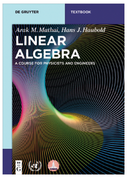
Arak M. Mathai and Hans J. Haubold

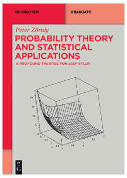
**Probability and Statistics**

De Gruyter Textbook

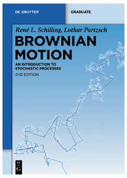
## Also of Interest


DE GRUYTER
TEXTBOOK
Arak M. Mathai, Hans J. Haubold
LINEAR ALGEBRA
A COURSE FOR PHYSICISTS AND ENGINEERS

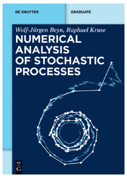
*Linear Algebra. A Course for Physicists and Engineers*
Arak M. Mathai, Hans J. Haubold, 2017
ISBN 978-3-11-056235-4, e-ISBN (PDF) 978-3-11-056250-7,
e-ISBN (EPUB) 978-3-11-056259-0


DE GRUYTER
GRADUATE
Peter Zörnig
PROBABILITY THEORY AND STATISTICAL APPLICATIONS
A PROFOUND TREATISE FOR SELF-STUDY

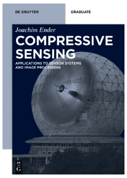
*Probability Theory and Statistical Applications.*
*A Profound Treatise for Self-Study*
Peter Zörnig, 2016
ISBN 978-3-11-036319-7, e-ISBN (PDF) 978-3-11-040271-1,
e-ISBN (EPUB) 978-3-11-040283-4


DE GRUYTER
GRADUATE
René L. Schilling, Lothar Partzsch
BROWNIAN MOTION
AN INTRODUCTION TO STOCHASTIC PROCESSES
2ND EDITION

*Brownian Motion. An Introduction to Stochastic Processes*
René L. Schilling, Lothar Partzsch, 2014
ISBN 978-3-11-030729-0, e-ISBN (PDF) 978-3-11-030730-6,
e-ISBN (EPUB) 978-3-11-037398-1


DE GRUYTER
GRADUATE
Wolf-Jürgen Beyn, Raphael Kruse
NUMERICAL ANALYSIS OF STOCHASTIC PROCESSES

*Numerical Analysis of Stochastic Processes*
Wolf-Jürgen Beyn, Raphael Kruse, 2018
ISBN 978-3-11-044337-0, e-ISBN (PDF) 978-3-11-044338-7,
e-ISBN (EPUB) 978-3-11-043555-9



*Compressive Sensing.*
*Applications to Sensor Systems and Image Processing*
Joachim Ender, 2018
ISBN 978-3-11-033531-6, e-ISBN (PDF) 978-3-11-033539-2,
e-ISBN (EPUB) 978-3-11-039027-8

Arak M. Mathai and Hans J. Haubold

# Probability and Statistics

A Course for Physicists and Engineers

DE GRUYTER

**Mathematics Subject Classification 2010**
60-00, 60-01, 60A05, 60D05, 60E05, 62-00, 62-01, 62P30, 62P35

**Authors**

Prof. Dr. Arak M. Mathai
McGill University
Department of Mathematics and Statistics
805 Sherbrooke St. West
Montreal, QC H3A 2K6
Canada
mathai@math.mcgill.ca

Prof. Dr. Hans J. Haubold
United Nations Office for Outer Space Affairs
Vienna International Centre
P.O. Box 500
1400 Vienna
Austria
hans.haubold@gmail.com

ISBN 978-3-11-056253-8
e-ISBN (PDF) 978-3-11-056254-5
e-ISBN (EPUB) 978-3-11-056260-6

**Library of Congress Cataloging-in-Publication Data**
A CIP catalog record for this book has been applied for at the Library of Congress.

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: VTeX, Lithuania
Printing and binding: CPI books GmbH, Leck
Cover image: Pasieka, Alfred / Science Photo Library
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

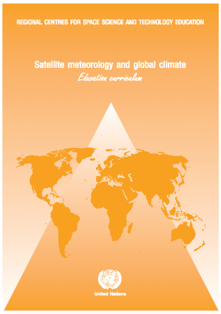
Due to great interest shown by students, a proposal was made to the Department of Science and Technology, Government of India, New Delhi, India (DST), for conducting a sequence of 10-day mathematics camps for undergraduates during holidays so that in one sequence of four camps the essential basic mathematics could be covered. DST approved this project with full financial support for 30 candidates in each camp. Principals of colleges in Kerala, India, were asked to select up to 5 motivated students from each college. The camp is run at CMSS from 08:30 am until 6:00 pm, continuously for 10 days of nearly 40 hours of lectures and 40 hours of problem-solving sessions. In one camp, Modules 1, 2, 3, on basic linear algebra for all disciplines, are covered. In a second camp, Modules 4 and 5, covering the topics of sequences, limits, continuity and differential calculus and the basic integral calculus, are covered. In a third camp, Module 6, covering the topics of sample space, probability, random variables, expected values, statistical distributions and their applications, are covered. In the fourth camp, Modules 7 and 9, covering the topics of sampling distributions, statistical inference, prediction and model building, design of experiments and some non-parametric tests, are covered.

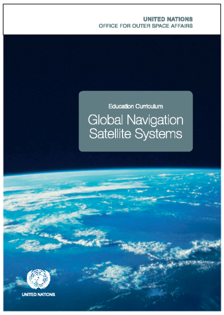
The basic sequence of four camps, covering basic mathematics, probability and statistics, were repeated over the years from 2007 to 2014. There were also a four to six weeks course every year at the research level, meant for MSc and PhD students, and young faculty, known as SERC (Science and Engineering Research Council of the Department of Science and Technology, Government of India, New Delhi) Schools. The notes from these schools were brought out by CMSS in book form every year. Selected notes from the SERC schools of 1995 to 2006 was brought out as *Special Functions for Applied Scientists* by Springer, New York in 2008, authored by A. M. Mathai and Hans J. Haubold. Selected notes from the SERC Schools of 2007 to 2012 appeared in 2017 as a research level book by Springer, New York. Selected notes from the SERC Schools of 2013 to 2015 will appear as a research level book on *Matrix Methods and Fractional Calculus* by World Scientific Publishing, all by the same authors.

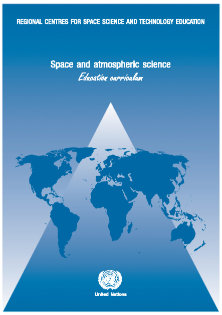
The present book on probability and statistics consists of 16 chapters. Chapters 1 to 9 cover the topics of random experiments, sample space, probability, how to assign probabilities to individual events, random variables, expected values, statistical distributions, collections of random variables and the central limit theorem. The statistics part consists of Chapters 10 to 16 covering the topic of sampling distributions, point estimation, interval estimation, tests of hypotheses, prediction, regression and model building problems, design of experiments and analysis of variance, some non-parametric tests, questions and answers.

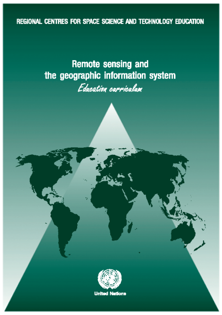
All concepts in probability and statistics are explained properly and illustrated with real-life examples. The material is developed slowly so that the book can be used as a self-study material also. Proper interpretations of the concepts of variance, co-

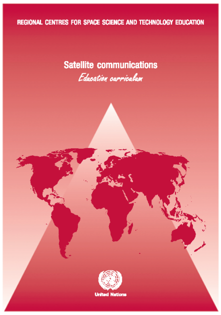
https://doi.org/10.1515/9783110562545-201


REGIONAL CENTRES FOR SPACE SCIENCE AND TECHNOLOGY EDUCATION
Satellite meteorology and global climate
Education curriculum
United Nations


UNITED NATIONS
OFFICE FOR OUTER SPACE AFFAIRS
Education Curriculum
Global Navigation Satellite Systems
UNITED NATIONS


REGIONAL CENTRES FOR SPACE SCIENCE AND TECHNOLOGY EDUCATION
Space and atmospheric science
Education curriculum
United Nations


REGIONAL CENTRES FOR SPACE SCIENCE AND TECHNOLOGY EDUCATION
Remote sensing and
the geographic information system
Education curriculum
United Nations



variance, correlation, multiple correlation, a statistical hypothesis, meaning of tests of a statistical hypothesis, difference between regression and correlation analysis, difference between prediction and model building, etc. are given. An introduction to matrix-variate distributions such as matrix-variate Gaussian, matrix-variate gamma and matrix-variate beta are also given.

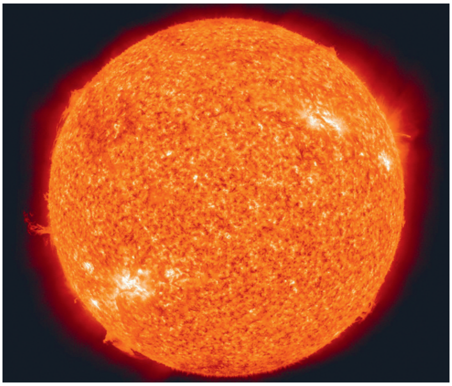
Since 2004, the material in this book was made available to UN-affiliated Regional Centres for Space Science and Technology Education, located in India, China, Morocco, Nigeria, Jordan, Brazil and Mexico (http://www.unoosa.org/oosa/en/ourwork/psa/regional-centres/index.html).

Since 1988, the material was taken into account for the development of education curricula in the fields of remote sensing and geographic information systems, satellite meteorology and global climate, satellite communications, space and atmospheric science and global navigation satellite systems (http://www.unoosa.org/oosa/en/ourwork/psa/regional-centres/study_curricula.html).

As such, the material was considered to be a prerequisite for applications, teaching and research in space science and technology. It was also a prerequisite for the nine-months post-graduate courses in the five disciplines of space science and technology, offered by the Regional Centres on an annual basis to participants from all 194 Member States of the United Nations.

Since 1991, whenever suitable at the research level, the material in this book was utilized in lectures in a series of annual workshops and follow-up projects of the so-called Basic Space Science Initiative of the United Nations (http://www.unoosa.org/oosa/en/ourwork/psa/bssi/index.html).

As such, the material was considered a prerequisite for teaching and research in astronomy and physics. Astronomy has a long history of exploiting observational data to estimate parameters and quantify uncertainty in physical models. Problems in astronomy propelled the development of many statistical techniques, from classical least squares estimation to contemporary methods such as diffusion entropy analysis. Late twentieth century advances in data collection, such as automation of telescopes and use of CCD cameras, resulted in a dramatic increase in data size and complexity, producing a surge in use and development of statistical methodology. Astronomers use these data sets for a diverse range of science goals, including modeling formation of galaxies, finding earth-like planets, estimating the metric expansion of space and classifying transients. This book reviews common data types and statistical methodology currently in use in space science, with the goal of making applications more accessible to methodological and applied statisticians. Naturally, the courses focused on the analysis and modeling of observational data (image data, spectral data, functional data, time series) emanating from a main sequence star, the Sun (figure below), utilizing solar particle, photon and neutrino radiation. Exercises went even that far to analyze solar neutrino data coming to the convincing conclusion that the solar neutrino flux is varying over time but not discovering a conclusive result of the physics that may drive such variation.



The Space and Atmospheric Sciences education curricula provides opportunities to teach basic space science. The development of the education curricula (illustrated above) started in 1988 at UN Headquarters in New York, the specific GNSS curriculum emanated only in 1999 after the UNISPACE III Conference, held at and hosted by the United Nations at Vienna.

CMSS
15 June 2017

A. M. Mathai and Hans J. Haubold
Peechi, Kerala, India

# Preface

Upon requests from colleges in Kerala, India, the Centre for Mathematical and Statistical Sciences (CMSS) had decided to conduct a series of remedial courses on selected topics. These topics were suggested by the college teachers themselves so that by participating in these courses they could be better prepared to teach the material in their classes. A series of such courses were conducted by CMSS from 1985 to 2002 at its Trivandrum Campus, Kerala, India. The notes written up for such courses and then class-tested at the University of Texas at El Paso, USA, formed Modules 1, 2, 3.

Modules Series of CMSS are meant for self-study. On selected basic topics in mathematical sciences, materials are developed with a lot of illustrative examples, starting from the fundamentals. Examples are taken from real-life situations so that the students can see the relevance of mathematics in solving real-life problems. The subject matter is developed very slowly so that there is time for absorbing the materials.

Modules 1, 2, 3 are on vectors, matrices, determinants and their applications. Module 4 is on limits, continuity, differentiability and differential calculus. Module 5 is on integrals and integration. Module 6 is on basic probability and random variables. This is Chapters 1 to 9 of the materials on basic probability and statistics. Module 7 is on statistics. Module 8 is on stochastic processes. Module 9 is about questions and answers, mainly the questions asked by the students over past years and their answers. Modules 1 to 9 are expected to cover the basic materials needed for a proper study of their own subjects for students from statistics, physics, engineering areas and other applied areas. This present material is a combined version of Modules 6, 7 and 9.

CMSS
15 June 2017

A. M. Mathai and Hans J. Haubold
Peechi, Kerala, India

https://doi.org/10.1515/9783110562545-202

# Acknowledgement

Dr B. D. Acharya (passed away), Dr H. K. N. Trivedi (now retired), Dr P. K. Malhotra and Dr Ashok K. Singh of the Mathematical Sciences Division of DST took the initiative in getting the funds released for the undergraduate training programs and SERC Schools of CMSS. CMSS would like to express its gratitude to these visionaries. A lot of people have contributed, directly or indirectly, in making CMSS Modules in their final printed forms. When Dr A. M. Mathai had written up the notes, he had set it in Plain TeX. Dr Shanoja R. Naik (Shanoja Pai) and Dr T. Princy of CMSS spent a lot of time in setting all the figures in the modules and translating to LaTeX and they deserve special thanks. All the PhD scholars at CMSS, Dr Nicy Sebastian, Dr Seema S. Nair, Dr Dhannya P. Joseph, Dr Dilip Kumar, Dr Naiju M. Thomas, Dr Anitha Thomas, Dr Sona Jose and Dr Ginu Varghese put in their time and efforts to make the Modules in the final format. Ms Sini Devassy did all the photo shop work for making the Modules ready for printing. Ms R. Girija looked after the office matters. The authors would like to thank all of them for their sincere and dedicated efforts.

https://doi.org/10.1515/9783110562545-203

# Contents

# List of Tables

# List of Symbols

| | |
|---|---|
| $\{\ldots\}$ | a set (Section 1.1, p. 2) |
| $\cup, \cap$ | union, intersection of sets (Section 1.1, p. 4) |
| $P(A)$ | probability of the event $A$ (Section 1.4, p. 18) |
| $P(n,r), {}_nP_r$ | number of permutations (Section 2.2, p. 33) |
| $(a)_k$ | Pochhammer symbol (Section 2.2, p. 36) |
| $\binom{n}{r}$ | number of combinations (Section 2.3, p. 38) |
| $\sum$ | sum notation (Section 2.4, p. 41) |
| $\prod$ | product notation (Section 2.4, p. 41) |
| $P(A\|B)$ | conditional probability (Section 2.5, p. 49) |
| $Pr\{x \leq a\}$ | probability of the event $\{x \leq x\}$ (Section 3.1, p. 61) |
| $F_x(x)$ | distribution function (Section 3.1, p. 65) |
| $E[\psi(x)]$ | expected value (Section 4.2, p. 83) |
| $\mu_r'$ | $r$-th moment about the origin (Section 4.3, p. 95) |
| $\mu_r$ | $r$-th central moment (Section 4.3, p. 96) |
| $\mu_{[r]}$ | factorial moment (Section 4.3, p. 96) |
| $M(t)$ | moment generating function (Section 4.3.1, p. 100) |
| $M_f(s)$ | Mellin transform of $f$ (Section 4.3.2, p. 102) |
| $\Gamma(\alpha)$ | gamma function (Section 6.1, p. 134) |
| $B(\alpha,\beta)$ | beta function (Section 6.4.1, p. 142) |
| $\sim$ | distributed as (Section 6.7, p. 149) |
| $N(\mu,\sigma^2)$ | normal or Gaussian population (Section 6.7, p. 149) |
| $\mathrm{d}x \wedge \mathrm{d}y$ | wedge product of differentials (Section 7.1, p. 171) |
| $\mathrm{Cov}(x,y)$ | covariance (Section 7.4, p. 187) |
| $M(t_1,\ldots,t_k)$ | moment generating function (Section 7.4.2, p. 191) |
| $\rho$ | correlation coefficient (Section 7.4.3, p. 198) |
| $E(x\|y)$ | conditional expectation (Section 7.5, p. 201) |
| $\mathrm{Cov}(X)$ | covariance matrix in the vector $X$ (Section 8.2, p. 223) |
| $J$ | Jacobian (Note 8.1, p. 235) |
| $\mathrm{d}X$ | wedge product of differentials in $X$ (Note 8.1, p. 235) |
| $\|A\|$ | determinant of the matrix $A$ (Note 8.1, p. 234) |
| $\Gamma_p(\alpha)$ | real matrix-variate gamma (Section 8.5, p. 243) |
| iid | independently and identically distributed (Section 10.1, p. 259) |
| mgf | moment generating function (Section 10.1, p. 259) |
| $N(0,1)$ | standard normal (Section 10.2, p. 264) |
| $\chi_\nu^2$ | chi-square variable (Section 10.2, p. 265) |
| $\mathrm{Var}(\cdot)$ | variance of $(\cdot)$ (Note 10.4, p. 268) |
| $\chi_\nu^2(\lambda)$ | non-central chi-square (Example 10.8, p. 275) |
| $t_\nu$ | Student-$t$ (Definition 10.6, p. 277) |
| $F_{m,n}$ | F-statistic (Definition 10.7, p. 280) |

https://doi.org/10.1515/9783110562545-204

| | |
|---|---|
| $X'$ | transpose of $X$ (Section 10.5, p. 285) |
| $\Sigma > 0$ | positive definite sigma (Note 10.11, p. 288) |
| $x_{n:1}$ | smallest order statistic (Section 10.6, p. 293) |
| $x_{n:n}$ | largest order statistic (Section 10.6, p. 293) |
| $x_{n:r}$ | $r$-th order statistic (Section 10.6, p. 294) |
| $I_n(\theta)$ | information in the sample (Note 11.9, p. 334) |
| $H_0$ | null hypothesis (Section 13.2, p. 383) |
| $H_1$ | alternate hypothesis (Section 13.2, p. 383) |
| MPT | most powerful test (Section 13.2, p. 387) |
| UMPT | uniformly most powerful test (Section 13.2, p. 388) |
| $D_n$ | Kolmogorov–Smirnov statistic (Section 13.9, p. 421) |
| $\rho_{1.(2\ldots k)}$ | multiple correlation coefficient (Section 14.5, p. 459) |
| $\eta_{1.(2\ldots k)}$ | multiple correlation ratio (Section 14.5, p. 467) |
| ANOVA | analysis of variance (Section 15.2, p. 499) |

# 1 Random phenomena

## 1.1 Introduction

In day-to-day life, we come across several situations, some of which are deterministic in nature but most of them are non-deterministic in nature. We will give some general ideas of deterministic and non-deterministic situations first, and then we will introduce a formal definition to define random quantities.

You got up in the morning and went for a morning walk for one hour. You had this routine and you had decided to go for a walk. The event that you went for a walk did not just happen but it was predetermined. You got ready and went to college as predetermined. If a normal human being consumes potassium cyanide or catches hold of a high voltage live electric wire, then the result is predetermined; the person dies or it is a sure event. Such situations or events are predetermined events. But there are several events which cannot be determined beforehand.

### 1.1.1 Waiting time

For going to college, you had to wait for a bus, which was scheduled to arrive at 8 a.m. The bus usually does not leave until after 8 a.m. if it arrives to the stop earlier because the bus has to cater to the regular passengers from that stop. But the actual amount of time that you had to wait on a particular day for that bus was not under your control, which could not be predetermined because due to traffic congestion on the way or due to many other reasons the bus could be late. The waiting time on that day might be 5 minutes or even 20 minutes. The waiting time then is a non-deterministic or random quantity.

Suppose that you went to a hospital for a routine medical check-up. Suppose that the check-up consists of taking your weight and measuring height, checking your blood pressure, taking blood samples, waiting for the doctor's physical examination, waiting for the result of the blood test, waiting for a chest x-ray, etc. Thus the total waiting time $T$ consists of several waiting times for the individual items which are all non-deterministic or random in nature or these are variables with some sort of randomness associated with them, where $T$ is of the form $T = t_1 + t_2 + \cdots + t_k$ if $t_j$ is the waiting time for the $j$-th item, such as a blood test, and if there are $k$ such items. If $t_j, j = 1, 2, \dots, k$ are all random quantities, then $T$ itself is a random quantity or these variables (durations of waiting) are not of predetermined durations. If the check-up consisted of only three items, then the total waiting time would be of the form $T = t_1 + t_2 + t_3$, where $t_j, j = 1, 2, 3$ are the individual waiting times.

Suppose you take your car for service. Suppose that the service consists of an oil and filter change, checking and correcting fluid levels, etc. Here also, each component

https://doi.org/10.1515/9783110562545-001

of the waiting time is of non-deterministic durations. The waiting time for the first child birth for a fertile woman from the time of co-habitation until child birth is a random quantity or of non-deterministic duration. The waiting time for a farmer for the first rainfall after the beginning of the dry season is a random quantity since it cannot be predetermined for sure.

### 1.1.2 Random events

Consider the event of occurrence of the first flood in the local river during a rainy season, flood is taken as the water level being above a threshold level. It is a random event because it cannot be predetermined. The event of a lightning strike or thunderbolt at a particular locality is a random event, and may be a rare event, but random in nature. The event of a snake bite in a particular village is a random event. If you are tying to swim across a river and if you have done this several times and if you are sure that you can successfully complete the swim, then your next attempt to swim across is a sure event and it is not a random event because there is no element of uncertainty about the success. If you tried this before and if you had a few occasions of failures and a few occasions of successes, then you will not be sure about your next attempt to swim across. Then the event of successful completion of the next attempt has become a random event. If you throw a stone into a pond of water and if your aim is to see whether the stone sinks in water, then it is deterministic in nature because you know about the physical laws and you know that the stone will sink in water. The outcome or the event is not random. Suppose that your aim is to see at which location on the pond where the stone will hit the water surface, then the location cannot be predetermined and the outcome or the event is a random event. If you throw a cricket ball upward, then the event that it comes down to Earth is not a random event because you know for sure that the ball has to come down due to the pull of gravity, but how high the ball will go, before starting to come down, is not predetermined, and hence it is a random quantity.

With the above general ideas in mind, we will give a systematic definition for a random event and then define the chance of occurrence of a random event or the probability of occurrence of a random event. To this end, we will start with the definition of a “random experiment”, then we will proceed to define events of various types.

**Note 1.1.** It is assumed that the students are familiar with the notations for a set and operations such as union, intersection and complementation on sets. Those who are familiar with these items may skip this note and proceed to the next section. For the sake of those who are not familiar or forgotten, a brief description is given here.

**(i) A set.** It is a well-defined collection of objects. The objects could be anything. For example, if the objects are the numbers $2, -1, 0$ in a set $A$, then it is written as $A = \{2, -1, 0\}$, that is, the objects are put within curly brackets. The order in which the objects are written is not important. We could have written the same set in different forms such as

$$A = \{2, -1, 0\} = \{0, -1, 2\} = \{-1, 0, 2\} \tag{N1.1}$$

and so on, in fact the 6 different ways in which the same set $A$ of 3 objects can be written here. The objects are called "elements" of the set and written as

$$2 \in A, \quad -1 \in A, \quad 0 \in A, \quad 5 \notin A, \quad 10 \notin A \tag{N1.2}$$

and they are read as 2 in $A$ or 2 is an element of $A$, $-1$ in $A$, 0 in $A$, whereas 5 not in $A$ or 5 is not an element of $A$, 10 not in $A$, etc. The symbol $\in$ indicates "element of". If $B$ is another set, say,

$$B = \{m, \theta, 8, *\}, \quad m \in B,\ \theta \in B,\ 8 \in B,\ * \in B \tag{N1.3}$$

then $B$ contains the elements, the Latin letter $m$, the Greek letter $\theta$ (theta), number 8 and the symbol $*$. Thus the objects or the elements of a set need not be numbers always. Some frequently used sets are the following:

$N = \{1, 2, 3, \dots\}$ = the set of natural numbers or positive integers;

$N_0 = \{0, 1, 2, \dots\}$ = the set of non-negative integers;

$R = \{x \mid -\infty < x < \infty\}$ = the set of all real numbers, where the vertical bar "|" indicates "such that" or "given that". It is read as "all values of $x$ such that minus infinity less than $x$ less than infinity";

$A_1 = \{x \mid -10 \le x \le 5\} = [-10, 5]$ = the closed interval from $-10$ to 5, closed means that both end points are included;

$A_2 = \{x \mid -10 < x \le 5\} = (-10, 5]$ = the semi-open (semi-closed) interval from $-10$ to 5, open on the left and closed on the right. Usually we use a square bracket "[" or "]" to show where it is closed, and open bracket "(" or ")" to show where it is open. In this example, the left side is open or the point $-10$ is not included and the right side is closed or the point 5 is included;

$A_3 = \{x \mid -10 \le x < 5\} = [-10, 5)$ = interval from $-10$ to 5, closed on the left and open on the right;

$A_4 = \{x \mid -10 < x < 5\} = (-10, 5)$ = open interval from $-10$ to 5;

$C = \{a + ib \mid -\infty < a < \infty, -\infty < b < \infty, i = \sqrt{-1}\}$ = the set of all complex numbers.

**(ii) Equality of sets.** Two sets are equal if they contain the same elements.

For example, if $A = \{2, 7\}$ and $B = \{x, 7\}$, then $A = B$ if and only if $x = 2$.

**(iii) Null set or vacuous set or empty set.** If a set has no elements, then it is called a null set, and it is denoted by $\phi$ (Greek letter phi) or by a big O.

For example, consider the set of all real solutions of the equation $x^2 = -1$. We know that this equation has no real solution but it has two imaginary solutions $i$ and $-i$ with

$i^2 = (-i)^2 = -1$. Thus the set of real solutions here is $\phi$ or it is an empty set. If the set contains a single element zero, that is, $\{0\}$, it is not an empty set. It has one element.

**(iv) Union of two sets.** The set of all elements which belong to a set $A_1$ or to the set $A_2$ is called the union of the sets $A_1$ and $A_2$ and it is written as $A_1 \cup A_2$. The notation $\cup$ stands for "union".

For example, let $A = \{2, -1, 8\}$ and $B = \{\theta, 8, 0, x\}$ then

$$A \cup B = \{2, -1, 8, \theta, 0, x\}.$$

Here, 8 is common to both $A$ and $B$, and hence it is an element of the union of $A$ and $B$ but it is not written twice. Thus all elements which are in $A$ or in $B$ (or in both) will be in $A \cup B$. Some examples are the following:

For the sets $N$ and $N_0$ above, $N \cup N_0 = N_0$;

$$A = \{x \mid -2 \le x \le 8\}, \quad B = \{x \mid 3 \le x \le 10\} \quad \Rightarrow \quad A \cup B = \{x \mid -2 \le x \le 10\}$$

where the symbol $\Rightarrow$ means "implies";

$$A = \{x \mid 1 \le x \le 3\}, \quad B = \{x \mid 5 \le x \le 9\}$$
$$\Rightarrow \quad A \cup B = \{x \mid 1 \le x \le 3 \text{ or } 5 \le x \le 9\}.$$

**(v) Intersection of two sets.** The set of all elements, which is common to two sets $A$ and $B$, is called the intersection of $A$ and $B$ and it is written as $A \cap B$, where the symbol $\cap$ stands for "intersection".

Thus, for the same sets $A = \{2, -1, 8\}$ and $B = \{\theta, 8, 0, *\}$ above, $A \cap B = \{8\}$ or the set containing only one element 8 because this is the only element common to both $A$ and $B$ here. Some examples are the following:

For the sets $N$ and $N_0$ above, $N \cap N_0 = N$ = the set of all natural numbers:

$$A = \{x \mid 0 \le x \le 7\}, \quad B = \{x \mid 5 \le x \le 8\} \quad \Rightarrow \quad A \cap B = \{x \mid 5 \le x \le 7\};$$
$$A = \{x \mid -1 < x < 2\}, \quad B = \{x \mid 4 < x < 12\} \quad \Rightarrow \quad A \cap B = \phi$$

which is an empty set because there is no element common to $A$ and $B$.

**(vi) Subset of a set.** A set $C$ is said to be a subset of the set $A$ if all elements of $C$ are also elements of $A$ and it is written as $C \subset A$ and read as $C$ is a subset of $A$ or $C$ is contained in $A$ or $A$ contains $C$.

From this definition, it is clear that for any set $B$, $B \subset B$. Let $A = \{2, -1, 8\}$ and let $C = \{8, -1\}$, then $C \subset A$. Let $D = \{8, -1, 5\}$, then $D$ is not a subset of $A$ because $D$ contains the element 5 which is not in $A$. Consider the empty set $\phi$, then from the definition it follows that

$$\phi \text{ is a subset of every set;} \quad \phi \subset A,\ \phi \subset C,\ \phi \subset D.$$

Consider the sets

$$A = \{x \mid 2 \le x \le 8\}, \quad B = \{x \mid 3 \le x \le 5\} \quad \Rightarrow \quad B \subset A.$$

Observe that the set of all real numbers is a subset of the set of all complex numbers. The set of all purely imaginary numbers, $\{a + ib$ with $a = 0\}$, is a subset of the set of all complex numbers. The set of all positive integers greater than 10 is a subset of the set of all positive integers.

**(vii) Complement of a set.** Let $D$ be a subset of a set $A$ then the complement of $D$ in $A$ or that part which completes $D$ in $A$ is the set of all elements in $A$ but not in $D$ when $D$ is a subset of $A$, and it is usually denoted by $D^c$ or $\bar{D}$. We will use the notation $D^c$.

Let $A = \{-3, 0, 6, 8, -11\}$ and let $D = \{0, 6, -11\}$ then $D^c = \{-3, 8\}$. Note that $D \subset A$ and $D \cup D^c = A$ and $D \cap D^c = \phi$. In general,

$$B \subset A \quad \Rightarrow \quad B^c \subset A, \quad B \cup B^c = A \quad \text{and} \quad B \cap B^c = \phi.$$

As another example,

$$A = \{x \mid 0 \le x \le 8\}, \quad B = \{x \mid 0 \le x \le 3\} \quad \Rightarrow \quad B^c = \{x \mid 3 < x \le 8\}$$

which is the complement of $B$ in $A$. Consider another subset here, $D = \{x \mid 3 < x < 5\}$, then what is the complement of $D$ in $A$? Observe that $D$ is a subset of $A$. Hence $D^c$ consists of all points in $A$ which are not in $D$. That is,

$$D^c = \{x \mid 0 \le x \le 3, 5 \le x \le 8\}$$

or the union of the pieces $0 \le x \le 3$ and $5 \le x \le 8$.

The above are some basic details about sets and basic operations on sets. We will be concerned about sets which are called “events”, which will be discussed next.

## Exercises 1.1

Classify the following experiments/events as random or non-random, giving justifications for your statements.

**1.1.1.** A fruits and vegetable vendor at Palai is weighing a pumpkin, selected by a customer, to see:
(a) the exact weight;
(b) whether the weight is more than 5 kg;
(c) whether the weight is between 3 and 4 kg.

**1.1.2.** For the same vendor in Exercise 1.1.1, a pumpkin is picked from the pile blindfolded and weighed to see (a), (b), (c) in Exercise 1.1.1.

**1.1.3.** A one kilometer stretch of a road in Kerala is preselected and a reporter is checking to see:
(a) the total number of potholes on this stretch;
(b) the total number of water-logged potholes on this stretch;
(c) the maximum width of the potholes on this stretch;
(d) the maximum depth of the potholes in this stretch.

**1.1.4.** For the same experiment in Exercise 1.1.3, a one kilometer stretch is selected by some random device and not predetermined.

**1.1.5.** A child is trying to jump across a puddle of water to see whether she can successfully do it.

**1.1.6.** Abhirami is writing a multiple choice examination consisting of 10 questions where each question is supplied with four possible answers of which one is the correct answer to the question. She wrote the answers for all the 10 questions, where:
(a) she knew all the correct answers;
(b) she knew five of the correct answers;
(c) she did not know any of the correct answers.

**1.1.7.** A secretary in an office is doing bulk-mailing. The experiment is to check and see how many pieces of mail are sent:
(a) without putting stamps;
(b) with an insufficient value of stamps;
(c) with wrong address labels.

**1.1.8.** A secretary is typing up a manuscript. The event of interest is to see on average that the number of mistakes per page is
(a) zero;
(b) less than 2;
(c) greater than 10.

**1.1.9.** A dandelion seed is flying around in the air with its fluffy attachment. The event of interest is to see:
(a) how high the seed will fly;
(b) when the seed will enter over your land;
(c) where on the your land the seed will settle down.

**1.1.10.** Sunlight hitting at a particular spot in our courtyard is suddenly interrupted by a cloud passing through. The event of interest is to see:
(a) at what time the interruption occurred;
(b) at which location on the light beam the interruption occurred.

**1.1.11.** You are looking through the binoculars to see the palm-lined beautiful shore of the other side of Vembanad Lake. Your view is interrupted by a bird flying across your line of view. The event of interest is:
(a) at which point the view is interrupted (distance from the binocular to the starting position of the bird);
(b) at what time the interruption occurred;
(c) the duration of interruption.

**1.1.12.** A student leader in Kerala is pelting stones at a passenger bus. The event of interest is to see:
(a) how many of his fellow students join him in pelting stones;
(b) extent of the money value of destruction of public property;
(c) how many passengers are hit by stones;
(d) how many passengers lose their eyes by the stones hitting the eyes.

## 1.2 A random experiment

Randomness is associated with the possible outcomes in a random experiment, not in the conduct of the experiment itself. Suppose that you consulted a soothsayer or "kaniyan" and conducted an experiment as per the auspicious time predicted by the kaniyan; the experiment does not become a random experiment. Suppose that you threw a coin. If the head (one of the sides of the coin) comes up, then you will conduct the experiment, otherwise not. Still the experiment is not a random experiment. Randomness is associated with the possible outcomes in your experiment.

**Definition 1.1** (Random experiment). A random experiment is such that the possible outcomes of interest, or the items that you are looking for, are not deterministic in nature or cannot be predetermined.

(i) Suppose that the experiment is that you jump off of a cliff to see whether you fly out in the north, east, west, south, up or down directions. In this experiment, we know the physical laws governing the outcome. There are not six possible directions in which you fly, but for sure there is only one direction, that is, going down. The outcome is predetermined and the experiment is not a random experiment.

(ii) If you put a 5-rupee coin into a bucket of water to see whether it sinks or not, there are not two possibilities that either it sinks or it does not sink. From the physical laws, we know that the coin sinks in water and it is a sure event. The outcome is predetermined.

(iii) If you throw a coin upward to see whether one side (call it heads) or the other side (call it tails) will turn up when the coin falls to the floor, assuming that the coin will not stay on its edge when it falls to the floor, then the outcome is not predetermined. You do not know for sure whether heads = $\{H\}$ will turn up or tails = $\{T\}$ will turn up. These are the only possible outcomes here, heads or tails, and these outcomes are not predetermined. This is a random experiment.

(iv) A child played with a pair of scissors and cut a string of length 20 cm. If the experiment is to see whether the child cut the string, then it is not a random experiment because the child has already cut the string. If the experiment is to see at which point on the string the cut is made, then it is a random experiment because the point is not determined beforehand.

The set of possible outcomes in throwing a coin once, or the outcome set, denoted by $S$, is then given by $S = \{H, T\} = \{T, H\}$.

**Definition 1.2** (A sample space or outcome set). The set of all possible outcomes in a random experiment, when no outcome there allows a subdivision in any sense, is called the sample space $S$ or the outcome set $S$ for that experiment.

The sample space for the random experiment of throwing a coin once is given by

$$S = \{H, T\} = \{T, H\}.$$

If the coin is tossed twice, then the possible outcomes are $H$ or $T$ in the first trial and $H$ or $T$ in the second trial. Then

$$S = \{(H,H),(H,T),(T,H),(T,T)\} \tag{1.1}$$

where, for example, the point $(H, T)$ indicates heads in the first trial and tails in the second trial. In this experiment of throwing a coin twice, suppose someone says that the sample space is

$$S_1 = \{\text{two heads, one head, zero head}\} = \{\text{zero tail, one tail, two tails}\},$$

is $S$ or $S_1$ the sample space here? Note that one tail or one head means that there are two possibilities of heads first and tails next or tails in the first trial and heads in the second trial. Thus the point "one head" or "one tail" allows a subdivision into two points $(H, T)$ and $(T, H)$ and in both of these points there is only exactly one head or exactly one tail. Hence $S_1$ allows one of its elements to be subdivided into two possible elements. Therefore, we will not take $S_1$ as the sample space here but $S$ in (1.1) is the sample space here.

**Example 1.1.** Construct the sample space of rolling a die once [A die is a cube with the faces marked with the natural numbers $1, 2, 3, 4, 5, 6$] and mark the subsets $A_1$ = the number is even, $A_2$ = the number is greater than or equal to 3.

**Solution 1.1.** The possibilities are that one of the numbers $1, 2, 3, 4, 5, 6$ will turn up. Hence

$$S = \{1,2,3,4,5,6\}$$

and the subsets are $A_1 = \{2,4,6\}$ and $A_2 = \{3,4,5,6\}$.

**Note 1.2.** In $S$, the numbers need not be taken in the natural order. They could have been written in any order, for example, $S = \{4,3,1,2,5,6\} = \{6,5,2,1,3,4\}$. Similarly, $A_1 = \{4,2,6\} = \{6,2,4\}$. In fact, $S$ can be represented in $6! = (1)(2)(3)(4)(5)(6) = 720$ ways. Similarly, $A_2$ can be written in $4! = 24$ ways and $A_1$ in $3! = 6$ ways, because in a set, the order in which the elements appear is unimportant.

**Definition 1.3** (An event or a random event). Any subset $A$ of the sample space $S$ of a random experiment is called an event or a random event. Hereafter, when we refer to an event, we mean a random event defined in a sample space or a subset of a sample space.

**Definition 1.4** (Elementary events). If a sample space consists of $n$ individual elements, such as the example of throwing a coin twice where there are 4 elements or 4 points in the sample space $S$, the singleton elements in $S$ are called the elementary events.

Let $A$ be an event in the sample space $S$, then $A \subset S$, that is, $A$ is a subset of $S$ or all elements in $A$ are also elements of $S$. If $A$ and $B$ are two subsets in $S$, that is,

$$A \subset S \quad \text{and} \quad B \subset S \quad \text{then } A \cup B \subset S$$

is called the event of occurrence of either $A$ or $B$ (or both) or the occurrence of at least one of $A$ and $B$. $A \cup B$ means the set of all elements in $A$ or in $B$ (or in both). Also $A \cap B$ is called the simultaneous occurrence of $A$ and $B$. Intersection of $A$ and $B$ means the set of all elements common to $A$ and $B$. Here, $\cup$ stands for "union" and $\cap$ stands for "intersection", $A \subset S$ stands for $A$ is contained in $S$ or $S$ contains $A$:

$$A \cup B = \text{occurrence of at least } A \text{ or } B.$$
$$A \cap B = \text{simultaneous occurrence of } A \text{ and } B.$$

Thus, symbolically, $\cup$ stands for "either, or" and $\cap$ stands for "and".

**Example 1.2.** In a random experiment of rolling a die twice, construct:
(1) the sample space $S$, and identify the events;
(2) $A$ = event of rolling 8 (sum of the face numbers is 8);
(3) $B$ = event of getting the sum greater than 10.

**Solution 1.2.** Here, in the first trial, one of the six numbers can come, and in the second trial also, one of the six numbers can come. Hence the sample space consists of all ordered pairs of numbers from 1 to 6. That is,

$$S = \{(1,1),(1,2),\dots,(1,6),(2,1),\dots,(6,6)\}.$$

There are 36 points in $S$. $A$ is the event of rolling 8. This can come by having 2 in the first trial and 6 in the second trial or 3 in the first trial and 5 in the second trial and so on. That is,

$$A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}.$$

The event of getting the sum greater than 10 means the sum is 11 or 12. Therefore,

$$B = \{(5,6),(6,5),(6,6)\}.$$

**Example 1.3.** In the experiment of throwing a coin twice, construct the events of getting:
(1) exactly two heads;
(2) at least one head;
(3) at least one tail, and interpret their unions and intersections.

**Solution 1.3.** Let $A$ = event of getting exactly two heads, $B$ = event of getting at least one head, and $C$ = event of getting at least one tail. Here, the sample space

$$S = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Then

$$\begin{aligned} &A = \{(H,H)\}, \quad B = \{(H,T),(T,H),(H,H)\}, \\ &C = \{(T,H),(H,T),(T,T)\}. \end{aligned}$$

At least one head means exactly one head or exactly two heads (one or more heads), and similar interpretation for $C$ also. [The phrase “at most” means that number or less, that is, at most 1 head means 1 head or zero head.] $A \cup B = B$ since $A$ is contained in $B$ here. Thus occurrence of $A$ or $B$ or both here means the occurrence of $B$ itself because occurrence of exactly 2 heads or at least one head implies the occurrence of at least one head. $A \cup C = \{(H,H),(H,T),(T,H),(T,T)\}$ = occurrence of exactly 2 heads or at least one tail (or both), which covers the whole sample space or which is sure to occur, and hence $S$ can be called the sure event. $A \cap B = \{(H,H)\}$ because this is the common element between $A$ and $B$. The simultaneous occurrence of exactly 2 heads and at least one head is the same as saying the occurrence of exactly 2 heads. $A \cap C =$ null set. There is no element common to $A$ and $C$. A null set is usually denoted by a big O or by the Greek letter $\phi$ (phi). Here, $A \cap C = \phi$. Also observe that it is impossible to have the simultaneous occurrence of exactly 2 heads and at least one tail because there is nothing common here. Thus the null set $\phi$ can be interpreted as the impossible event. Note that, by definition,

$$A \cup B = B \cup A, \quad A \cap B = B \cap A, \quad A \cup C = C \cup A, \quad A \cap C = C \cap A.$$

Now,

$$B \cup C = B \quad \text{or} \quad C(\text{or both}) = \{(H,T),(T,H),(H,H),(T,T)\} = S$$

which is sure to happen because the event of getting at least one head or at least one tail (or both) will cover the whole sample space:

$$B \cap C = \{(H,T),(T,H)\} =$$

event of getting exactly one head = event of getting exactly one tail, since the common part is the occurrence of exactly one head or exactly one tail, which then is the intersection of $B$ and $C$.

Also singleton elements are called *elementary events in a sample space*. Thus in Example 1.3, there are 4 elementary events in $S$. Also, the non-occurrence of an event $A$ is denoted by $\bar{A}$ or $A^c$. We will use the notation $A^c$ to denote the non-occurrence of $A$. In Example 1.3, if $A$ is the event of getting exactly 2 heads, then $A^c$ will be the event of the non-occurrence of $A$, which means the occurrence of exactly one head or exactly zero heads (or exactly 2 tails). Thus $A^c = \{(H,T),(T,H),(T,T)\}$ where $A = \{(H,H)\}$.

**Notation 1.1.** $A^c$ = non-occurrence of the event $A$ when $A$ and $A^c$ are in the same sample space $S$.

Note that if $A$ and $B$ are two events in the same sample space $S$, then if $A \cap B = \phi$, this means that they cannot occur simultaneously or the occurrence of $A$ excludes the occurrence of $B$ and vice versa. Then $A$ and $B$ will be called *mutually exclusive events*.

$A \cap B = \phi \Rightarrow A$ and $B$ are mutually exclusive or the occurrence of $A$ excludes the occurrence of $B$ and vice versa.

In the light of the above discussion, we have the following general results and interpretations for events in the same sample space:

$$S = \text{sample space} = \text{sure event} \tag{i}$$

$$\phi = \text{null set} = \text{impossible event} \tag{ii}$$

[Note that by assumption a null set is a subset of every set.]

$$A \cup B = \text{occurrence of at least } A \text{ or } B \text{ or both} \tag{iii}$$

$$A \cap B = \text{simultaneous occurrence of } A \text{ and } B \tag{iv}$$

$$A \cap B = \phi \text{ means } A \text{ and } B \text{ are mutually exclusive} \tag{v}$$

$$A \cup B = S \text{ means that } A \text{ and } B \text{ are totally exhaustive events} \tag{vi}$$

$$A^c = \text{complement of } A \text{ in } S = \text{non-occurrence of the event } A. \tag{vii}$$

Single elements in $S$ are called *elementary events* if $S$ consists of distinct finite number of elements.

**Note 1.3.** In the example of throwing a coin once, suppose that a physicist is capable of computing the position of the coin and the amount of pressure applied when it was thrown, all the forces acting on the coin while it is in the air, etc., then the physicist may be able to tell exactly whether that throw will result in a head or tail for sure. In that case, the outcome is predetermined. Hence one can argue that an experiment becomes random only due to our lack of knowledge about the various factors affecting the outcome. Also note that we do not have to really throw a coin for the description of our random experiment to hold. We are only looking at the possible outcomes "if" a coin is thrown. After it is thrown, we already know the outcome. In the past, a farmer used to watch the nature of the cloud formation, the coolness in the wind, the direction of the wind, etc. to predict whether a rain is going to come on that day. His prediction might be wrong 70% of the times. Nowadays, a meteorologist can predict, at least in the temperate zones, the arrival of rain including the exact time and the amount of rainfall, very accurately at least one to two days beforehand. The meteorologist may be wrong in less than 1% of the time. Thus, as we know more and more about the factors affecting an event, we are able to predict its occurrence more and more accurately, and eventually possibly exactly. In the light of the above details, is there anything called a random experiment?

Before concluding this section, let us examine one more point. The impossible event $\phi$ is often misinterpreted. Suppose that a monkey is given a computer to play with. Assume that it does not know any typing but only playing with the keyboard with the English alphabet. Consider the event that the monkey's final creation is one of the speeches of President Bush word-by-word. This is not a logically impossible event and we will not denote this event by $\phi$. We use $\phi$ for logically impossible events. The event that the monkey created one of the speeches is almost surely impossible. Such events are called *almost surely impossible events*. Consider the event of the best student in this class passing the next test. We are almost sure that she will pass but only due to some unpredicted mishap she may not pass. This is *almost surely a sure event* but not logically a sure event, and hence this cannot be denoted by our symbol $S$ for a sure event.

## Exercises 1.2

**1.2.1.** Write down the outcome set or the sample space in the following experiments:
(a) A coin is tossed 2 times (assume that only head = $H$ or tail = $T$ can turn up);
(b) Two coins together are tossed once;
(c) Two coins together are tossed two times.

**1.2.2.** Write down the outcome set or sample space in the following experiment:

(a) A die is rolled two times;
(b) Two dies are rolled together once;
(c) Two dies together are rolled two times.

**1.2.3.** Count the number of elementary events or sample points in the sample spaces of the following experiments, if possible:
(a) A coin is tossed 20 times;
(b) Two coins together are tossed 15 times;
(c) A die is rolled 3 times;
(d) Two dies together are rolled 5 times.

**1.2.4.** Write down the sample space in the following experiments and count the number of sample points, if possible:
(a) A random cut is made on a string of a length of 20 cm (one end is marked zero and the other end 20 and let $x$ be the distance from zero to the point of cut);
(b) Two random cuts are made on a string of a length of 20 cm and let $x$ and $y$ be the distances from zero to the points of cut, respectively.

**1.2.5.** There are 4 identical tags numbered $1, 2, 3, 4$. These are put in a box and well shuffled and tags are taken blind-folded. Write down the sample spaces in the following situations, and count the number of sample points in each case, if possible:
(a) One tag is taken from the box;
(b) One tag is taken. Its number is noted and then returned to the box, again shuffled and a second tag is taken (this is called sampling with replacement);
(c) In (b) above, the first tag is kept aside, not returned to the box, then a second tag is taken after shuffling (this is called sampling without replacement);
(d) Two tags are taken together in one draw.

**1.2.6.** Write down the sample space in the following cases when cards are taken, blind-folded, from a well-shuffled deck of 52 playing cards, and count the number of sample points in each case, if possible:
(a) One card is taken;
(b) Two cards are taken at random with replacement;
(c) Two cards are taken at random without replacement;
(d) Two cards together are taken in one draw.

**1.2.7.** Construct the sample spaces in the following experiments:
(a) Checking the life-time $x$ of one electric bulb;
(b) Checking the life-times $x$ and $y$ of two electric bulbs.

**1.2.8.** In Exercise 1.2.1 (b), write down the following events, that is, write down the corresponding subsets of the sample space:

(a) $A$ = the event of getting at most one tail;
(b) $B$ = the event of getting exactly 3 tails;
(c) $C$ = the event of getting at most one head;
(d) Interpret the events (i): $A \cap B$, (ii): $A \cup B$, (iii): $A \cap C$, (iv): $A \cup C$, (v): $A^c$, (vi): $(A \cup C)^c$, (vii): $(A \cap C)^c$.

**1.2.9.** In Exercise 1.2.2 (b), write down the following events:
(a) $A$ = the number in the first trial is bigger than or equal to the number in the second trial;
(b) $B$ = the sum of the numbers in the two trials is (i): bigger than 12, (ii): between 8 and 10 (both inclusive), (iii): less than 2.

**1.2.10.** In Exercise 1.2.4 (a), write down the following events and give graphical representations:
(a) $A$ = event that the smaller portion is less than 5 cm;
(b) $B$ = event that the smaller portion is between 5 and 10 cm;
(c) $C$ = the event that the smaller portion is less than $\frac{1}{2}$ of the larger portion.

**1.2.11.** In Exercise 1.2.4 (b), write down the following events and give graphical representations:
(a) $x < y$;
(b) $x + y < 10$;
(c) $x^2 \leq y^2$;
(d) $xy \leq 5$.

**1.2.12.** In Exercise 1.2.5 (b), write down the following events:
(a) $A$ = event that the first number is less than or equal to the second number;
(b) $B$ = event that the first number is less than or equal to $\frac{1}{2}$ of the second number.

**1.2.13.** In Exercise 1.2.5 (c), write down the same events $A$ and $B$ in Exercise 1.2.12.

## 1.3 Venn diagrams

We will borrow the graphical representation of sets as Venn diagrams from set theory. In a Venn diagrammatic representation, a set is represented by a closed curve, usually a rectangle or circle or ellipse, and subsets are represented by closed curves within the set or by points within the set or by regions within the set, see Figures 1.1, 1.2, 1.3. Examples are the following: $A \subset S$, $B \subset S$, $C \subset S$, $D \subset S$, $E \subset S$ all are events in the same sample space $S$. In the Venn diagram in Figure 1.2, $A$ and $B$ intersect, $A$ and $C$ intersect, $B$ and $C$ intersect, $A, B, C$ all intersect and $D$ and $E$ do not intersect. $A \cap B$ is the shaded region, also $A \cap B \cap C$ is the shaded region. $D \cap E = \phi$ or they are mutually exclusive. By

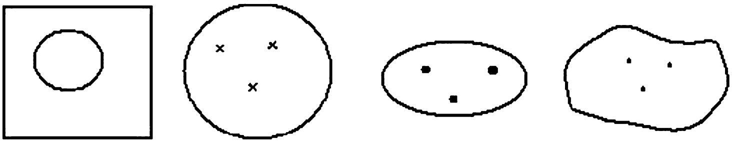


Figure 1.1: Venn diagrams for sample space.

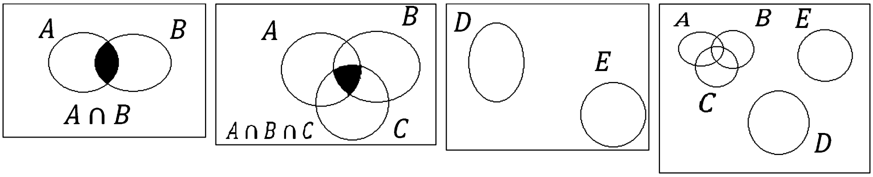


Figure 1.2: Representations of events.

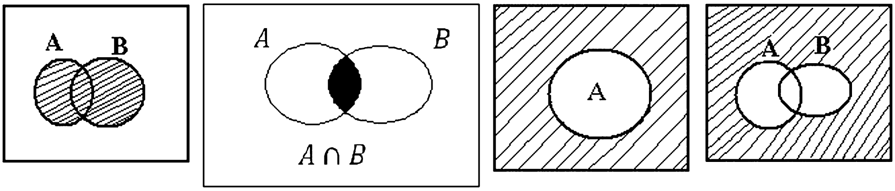


Figure 1.3: Union, intersection, complement of events.

looking at the Venn diagram, we can see the following general properties. $A \cup B$ can be split into three mutually exclusive regions $A \cap B^c$, $A \cap B$, $A^c \cap B$. That is,

$$\begin{aligned} A \cup B &= (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B) \\ &= A \cup (A^c \cap B) = B \cup (B^c \cap A) \end{aligned} \tag{1.2}$$

where

$$\begin{aligned} &(A \cap B^c) \cap (A \cap B) = \phi, \quad (A \cap B^c) \cap (A^c \cap B) = \phi, \\ &(A \cap B) \cap (A^c \cap B) = \phi, \end{aligned} \tag{1.3}$$

$$\begin{aligned} A \cap (A^c \cap B) &= \phi, \\ B \cap (B^c \cap A) &= \phi \end{aligned} \tag{1.4}$$

or they are all mutually exclusive events. Also note that for any event $A$,

$$A \cup \phi = A, \quad A \cap \phi = \phi \tag{1.5}$$

which also means that the sure event $S$ and the impossible event $\phi$ are mutually exclusive events as well as totally exhaustive events.

Consider the set of events $A_1,\dots,A_k$ in the same sample space $S$, that is, $A_j \subset S$, $j = 1,2,\dots,k$ where $k$ could be infinity also or there may be countably infinite number of events. Let these events be mutually exclusive and totally exhaustive. That is,

$$\begin{aligned}&A_1 \cap A_2 = \phi, \quad A_1 \cap A_3 = \phi, \quad \dots, \quad A_1 \cap A_k = \phi\\ &A_2 \cap A_3 = \phi, \quad \dots, \quad A_2 \cap A_k = \phi, \quad \dots, \quad A_{k-1} \cap A_k = \phi \quad \text{and}\\ &\qquad S = A_1 \cup A_2 \cup \dots \cup A_k.\end{aligned}$$

This can also be written as follows:

$$A_i \cap A_j = \phi, \quad \text{for all } i \neq j,\ i,j = 1,2,\dots,k, \quad A_1 \cup A_2 \cup \dots \cup A_k = S.$$

Then we say that the sample space $S$ is partitioned into $k$ mutually exclusive and totally exhaustive events. We may represent this as the following Venn diagram in Figure 1.4.

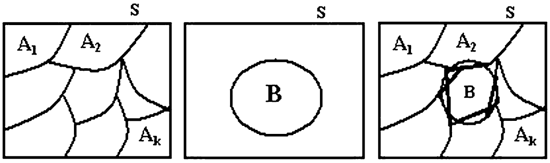


**Figure 1.4:** Partitioning of a sample space.

If $B$ is any other event in $S$ where $S$ is partitioned into mutually exclusive events $A_1,A_2,\dots,A_k$, then note from the Venn diagram that $B$ can be written as the union of mutually exclusive portions $B \cap A_1, B \cap A_2, \dots, B \cap A_k$, some of the portions may be null, that is,

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_k)$$

with

$$(B \cap A_1) \cap (B \cap A_2) = \phi, \quad \dots, \quad (B \cap A_{k-1}) \cap (B \cap A_k) = \phi. \tag{1.6}$$

## Exercises 1.3

**1.3.1.** Consider the sample space $S$ when a coin is tossed twice:

$$S = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Indicate the correct statements from the following list of statements:

(a) $(H,H) \in S$;

(b) $\{(H,H)\} \in S$;
(c) $\{(H,H)\} \subset S$;
(d) $\{(H,T),(T,H)\} \subset S$;
(e) $\{(H,T),(T,H)\} \in S$.

**1.3.2.** Let $S = \{x \mid 0 \le x \le 10\}$. List the correct statements from the following list of statements:
(a) $2 \in S$;
(b) $\{2\} \in S$;
(c) $A = \{x \mid 2 \le x < 5\} \in S$;
(d) $A = \{x \mid 2 \le x < 5\} \subset S$.

**1.3.3.** For the same sample space $S$ in Exercise 1.3.1, let $A = \{(H,H)\}$, $B = \{(H,H),(H,T),(T,H)\}$, $C = \{(T,T)\}$. Draw one Venn diagram each to illustrate the following: (1) $A$ and $A^c$; (2) $A \cup B$, $A \cap B$, $A \cup B \cup C$, $A \cap B \cap C$.

**1.3.4.** By using the definition of sets, prove that the following statements hold with reference to a sample space $S$ where $A, B, C$ are events in $S$:
(a) $A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$;
(b) $A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$.

**1.3.5.** For the same sample space $S$ and events $A, B, C$ in Exercise 1.3.3, verify the following results with the help of Venn diagrams and interpret each of the events:
(a) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
(b) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
(c) $(B \cup C)^c = B^c \cap C^c$;
(d) $(A \cap B)^c = A^c \cup B^c$.

**1.3.6.** By constructing an example of your own, verify the fact that for three events in the same sample space $S$:
(a) $A \cup B = A \cup C$ need not imply that $B = C$;
(b) $A \cap B = A \cap C$ need not imply that $B = C$.

**1.3.7.** For events $A, B, C$ in the same sample space $S$, prove the following results:
(a) $(A \cup B)^c = A^c \cap B^c$;
(b) $(A \cap B)^c = A^c \cup B^c$;
(c) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
(d) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

**1.3.8.** For events $A_1, A_2, \ldots$ in the same sample space $S$, prove the following results:
(a) $(A_1 \cup A_2 \cup \cdots)^c = A_1^c \cap A_2^c \cap \cdots$;
(b) $(A_1 \cap A_2 \cap \cdots)^c = A_1^c \cup A_2^c \cup \cdots$.

## 1.4 Probability or chance of occurrence of a random event

Now, we shall try to assign a numerical value for the chance of occurrence of a random event in a well-defined sample space $S$. Let $P(A)$ denote the probability or chance of occurrence of the event $A \subset S$.

**Notation 1.2.** $P(A) =$ the probability of the event $A$.

We are trying to give a meaning to the following types of statements that you hear every day:

(1) There is a 95% chance of rain today $\Rightarrow P(A) = 0.95$ where the symbol $\Rightarrow$ stands for "implies", and $A$ is the event of having rain.

(2) The chance of winning a certain lottery is one in a million $\Rightarrow P(A) = \frac{1}{10\,000\,000}$.

(3) The chance of a flood on campus today is nearly zero $\Rightarrow P(A) \approx 0$.

(4) The chance that Miss Cute will win the beauty contest is more than Miss Wise wins $\Rightarrow P(A) > P(B)$ where $A$ is the event that Miss Cute wins and $B$ is the event that Miss Wise wins.

We can define this probability by using some postulates or axioms. Postulates or axioms are logically consistent and non-overlapping types of basic assumptions that we make to define something. Usually such postulates are taken by taking into consideration plausible properties that we would like to have for the item to be defined. There is no question of proving or disproving these basic assumptions. The following three postulates will be used to define $P(A)$ the probability or chance of occurrence of the event $A$.

### 1.4.1 Postulates for probability

(i) $0 \leq P(A) \leq 1$ or the probability of an event is a number between 0 and 1, both inclusive;

(ii) $P(S) = 1$ or the probability of the sure event is 1;

(iii) $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$ whenever $A_1, A_2, \ldots$ are mutually exclusive [The events may be finite or countably infinite in number].

Thus $P(\cdot)$ coming out of the above three axioms will be called the probability of the event $(\cdot)$. Let us see what will be that quantity in the simple example of throwing a coin once.

**Example 1.4.** Compute the probability of getting a head when a coin is tossed once.

**Solution 1.4.** We have already computed the sample space for this experiment:

$$S = \{H, T\}.$$

For convenience, let $A$ be the event of getting a head, then $A = \{H\}$ and let $B$ be the event of getting a tail, that is, $B = \{T\}$. When we get a head, we cannot get a tail at the same time, and hence $A$ and $B$ are mutually exclusive events or $A \cap B = \phi$. In a trial, one of the events $A$ or $B$ must occur, that is, either a head or a tail will turn up because we have ruled out the possibility that the coin will fall on its edge. Thus $A \cup B = S$. Thus we have

$$A \cap B = \phi \quad \text{and} \quad A \cup B = S.$$

From postulate (ii),

$$1 = P(S) = P(A \cup B).$$

Now from postulate (iii),

$$P(A \cup B) = P(A) + P(B)$$

since $A \cap B = \phi$. Therefore,

$$1 = P(A) + P(B) \quad \Rightarrow \quad P(A) = 1 - P(B).$$

We can only come up to this line and cannot proceed further. The above statement does not imply that $P(A) = \frac{1}{2}$. There are infinitely many values $P(A)$ can take so that the equation $1 = P(A) + P(B)$ is satisfied. In other words, by using the postulates we cannot compute $P(A)$ even for this simple example of throwing a coin once.

**Example 1.5.** Compute the probability of getting a sum bigger than 10 when a die is rolled twice.

**Solution 1.5.** Here, the sample space consists of 36 elementary events as seen before. Let $A$ be the event of getting a sum bigger than 10, which means a sum 11 or 12. Then $A = \{(5,6),(6,5),(6,6)\}$ and let $B$ be the event of getting a sum less than or equal to 10 or the complement of $A$ or $B = A^c$. But for any event $A$, $A \cup A^c = S$ and $A \cap A^c = \phi$ or these two are mutually exclusive and totally exhaustive events. Therefore, by postulates (ii) and (iii) we can come up to the stage:

$$1 = P(A) + P(B).$$

We know that the event $A$ has 3 of the 36 elementary events in the sample space $S$. But we cannot jump into the conclusion that therefore $P(A) = \frac{3}{36}$ because from the definition of probability, as given by the postulates, does not depend upon the number of sample points or elementary events favorable to the event. If someone makes such a conclusion and writes the probability as $\frac{3}{36}$ in the above case, it will be wrong, which may be seen from the following considerations:

(i) A housewife is drying clothes on the terrace of a high-rise apartment building. What is the probability that she will jump off the building? There are only two possibilities: either she jumps off or she does not jump off. Therefore, if you say that the probability is $\frac{1}{2}$ obviously, you are wrong. The chance is practically nil that she will jump off the building.

(ii) What is the probability that there will be a flash flood on this campus in the next 5 minutes? There are two possibilities: either there will be flash flood or there will not be a flash flood. If you conclude that the probability is therefore $\frac{1}{2}$ you are definitely wrong because we know that the chance of a flash flood here in the next 5 minutes is practically nil.

(iii) At this place, tomorrow can be a sunny day, or a cloudy day or a mixed sunny and cloudy day, or a rainy day. What is the probability that tomorrow will be rainy day? If you say that the probability is $\frac{1}{4}$ since we have identified four possibilities, you can be obviously wrong because these four possibilities need not have the same probabilities. Since today is sunny and since it is not a rainy season, most probably tomorrow will also be a sunny day.

(iv) A child cuts a string of 40 cm in length into two pieces while playing with a pair of scissors. Let one end of the string be marked as 0 and the other end as 40. What is the probability that the point of the cut is in the sector from 0 to 8 cm? Here, the number of sample points is infinite, not even countable. Hence by the misuse of the idea that the probability may be the number of sample points favorable to the event to the total number of sample points, we cannot come up with an answer, even though wrong, as in the previous examples. The total number as well as the number of points favorable to the event cannot be counted. Then how do we calculate this probability?

(v) A person is throwing a dart at a square board of 100 cm in length and width. What is the probability that the dart will hit in a particular $10\,\text{cm} \times 10\,\text{cm}$ region on the board? Here, also we cannot count the number of sample points even if to misuse the numbers to come up with an answer. Then how do we compute this probability?

(vi) A floor is paved with square tiles of length $m$ units. A circular coin of diameter $d$ units is thrown upward, where $d < m$. What is the probability that the coin will fall clean within a tile, not cutting its edges or corners? This is the famous Buffon's "clean tile problem" from where the theory of probability has its beginning. How do we answer these types of questions (definitely not by counting the sample points)?

From the above examples, it is clear that the probability of an event does not depend upon the number of elementary events in the sample space and the number of elementary events favorable to the event. It depends upon many factors. Also we have seen that our definition of probability, through the three axioms, does not help us to evaluate the probability of a given event. In other words, the theory is useless, when it comes to the problem of computing the probability of a given event or the theory is not applicable to a practical situation, unless we introduce more assumptions or extraneous considerations.

Before we introduce some rules, we can establish some general results by using the axioms for a probability.

**Result 1.1.** *Probability of an impossible event is zero, that is,* $P(\phi) = 0$.

**Proof 1.1.** Consider the sure event $S$ and the impossible event $\phi$. From the definitions

$$S \cup \phi = S \quad \text{and} \quad S \cap \phi = \phi.$$

Hence from postulates (ii) and (iii):

$$1 = P(S) = P(S \cup \phi) = P(S) + P(\phi) = 1 + P(\phi)$$

since $S \cap \phi = \phi$. But probability is a real number. Therefore,

$$1 = 1 + P(\phi) \quad \Rightarrow \quad P(\phi) = 0.$$

**Result 1.2.** *Probability of non-occurrence = 1-probability of occurrence or* $P(A^c) = 1 - P(A)$.

**Proof 1.2.** We note that $A$ and $A^c$ are mutually exclusive and totally exhaustive events, and hence from axioms (ii) and (iii) we have

$$S = A \cup A^c \quad \text{and} \quad A \cap A^c = \phi \quad \Rightarrow \quad 1 = P(A) + P(A^c) \quad \Rightarrow \quad P(A^c) = 1 - P(A).$$

For example, if 0.8 is the probability that Abhirami will pass the next class test then 0.2 is the probability that she may not pass the next class test. If 0.6 is the probability that Abhirami may be the top scorer in the next class test, then 0.4 is the probability that she may not be the top scorer in the next class test.

**Result 1.3.** *For any two events $A$ and $B$ in the same sample space $S$,*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

**Proof 1.3.** From equation (1.2) or from the corresponding Venn diagram, we have seen that $A \cup B$ can be written as the union of three mutually exclusive events $A \cap B^c$, $A \cap B$ and $A^c \cap B$. Hence from axiom (iii),

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \tag{a}$$

$$P(A) = P(A \cap B^c) + P(A \cap B)$$

since

$$(A \cap B^c) \cap (A \cap B) = \phi \quad \Rightarrow \quad P(A \cap B^c) = P(A) - P(A \cap B). \tag{b}$$

Similarly,

$$P(B) = P(B \cap A^c) + P(A \cap B) \quad \Rightarrow \quad P(B \cap A^c) = P(B) - P(A \cap B). \tag{c}$$

Substituting (b) and (c) in (a), we have

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

This completes the proof. This can also be guessed from the Venn diagrammatic representation, taking probability as some sort of measure over the regions. The same measure over the region $A$ plus over the region $B$ will count twice over the region $A \cap B$, and hence once it should be subtracted. But this is not a proof but the results can be guessed from the Venn diagram. The above results can be extended for a set of three events or to a set of $k$ events, $k = 2, 3, \ldots$.

**Result 1.4.** *Let $A, B$ and $C$ be three events in the same sample space $S$. Then*

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$

**Proof 1.4.** The proof follows parallel to the steps in the proof for the Result 1.3, and hence it is left to the students. Also as an exercise, write down the formula for the probability of the union of $k$ events $A_1, \ldots, A_k$.

## 1.5 How to assign probabilities to individual events?

Now, we introduce a number of rules so that with the help of the axiomatic definition and the following rules one may be able to compute the probabilities in a number of situations.

**Rule 1.1** (Symmetry in the outcomes). *If the sample space consists of a finite number of distinct elements and if the physical characteristics of the experiments are such that, with respect to all factors which may affect the possible outcomes, there is no way of preferring one outcome to the other then the rule says to assign equal probabilities to the elementary events.*

In order to apply this rule, one has to have a sample space consisting of a finite number of elements, situations such as tossing a coin once or a number of times, rolling a die a number of times, predicting successful completion of a job when there are only a fixed number of alternatives, predicting rainfall, etc. The rule does not apply to situations such as the cutting a string where the sample space consists of a continuum of points, throwing a dart at a target where the sample space consists of regions, Buffon's clean tile problem where the sample space consists of a room paved with tiles or a finite planar region, and so on. The implication of Rule 1.1 is the following.

**Rule 1.1a.** *When there is symmetry in the outcome of a random experiment and when there are k elementary events in the sample space S, k being finite, and if m of the sample points (elementary events) are favorable to an event A, then the probability of the event A will be taken as*

$$P(A) = \frac{\textit{number of sample points favorable to A}}{\textit{total number of sample points}} = \frac{m}{k}. \tag{1.7}$$

Let us take the example of tossing a coin twice. The sample space is $S = \{(H,T),(T,H),(T,T)\}$. If the physical characteristics of the coin are such that there is no way of preferring one side to the other (in such a case we call the coin *an unbiased coin* or not loaded towards one side), the throwing of the coin is such that there is no advantage for one side over the other or, in short, with respect to all factors which may affect the outcomes, there is no advantage for one side over the other, then in this case we assign equal probabilities of $\frac{1}{4}$ to the individual elements in this sample space. That is, we assign, for the event of getting the sequence $H$ first and $T$ next a probability of $\frac{1}{4}$.

In some books, you may find the description saying when the "events are equally likely" they have equal probabilities. The statement is circumlocutory in the sense of using "probability" to define probability. Symmetry has nothing to do with the chances for the individual outcomes. We have the axioms defining probabilities and we have seen that the axioms are not sufficient to compute the probabilities in specific situations, and hence it is meaningless to say "equally likely events" when trying to compute the probabilities of events. Symmetry is concerned about the physical characteristics of the experiments and the factors affecting the outcomes and not about the chances of occurrence of the events.

The phrase used to describe symmetry are "unbiased coin" in the case of coins, "balanced die" in the case of rolling a die, and in other cases, we say "when there is symmetry in the experiment or symmetry in the outcomes".

Thus, in the example of tossing a coin if we ask:

What is the probability of getting a head when an unbiased coin is tossed once, then the answer is $\frac{1}{2}$. This value is assigned by us by taking into account of symmetry in the experiment, and not coming from the axioms or deduced from somewhere.

What is the probability of getting exactly one head when an unbiased coin is tossed twice?

*Answer*: Let $A$ be the event of getting exactly one head. Let $A_1$ be the event of getting the sequence $(H,T)$ and $A_2$ be the event of getting the sequence $(T,H)$. Then

$$A = A_1 \cup A_2 \quad \text{and} \quad A_1 \cap A_2 = \phi.$$

Therefore,

$$P(A) = P(A_1) + P(A_2)$$

by the third axiom. But we have assigned probabilities $\frac{1}{4}$ to individual outcomes because of the additional assumption of symmetry, and hence $P(A_1) = \frac{1}{4}$ and $P(A_2) = \frac{1}{4}$. Therefore,

$$P(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

**Example 1.6.** An unbiased coin is tossed (a) three times and (b) four times. What are the probabilities of getting the sequences (i) $HHT$, (ii) $THT$ in (a) and the sequences (iii) $HHTT$, (iv) $HHHH$ or $HTTT$ in (b)?

**Solution 1.6.** In (a), the sample space consists of all possible sequences of $H$ and $T$ and there are 8 such elementary events. They are available by looking at the problem of filling three positions by using $H$, and $T$. The first position can be filled in two ways, either $H$ or $T$. For each such choice, the second position can be filled in two ways. For each such choice, for the first and second positions the third can be filled in two ways so that the number of possible outcomes is $2 \times 2 \times 2 = 8$. They are the following:

$$HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT.$$

Since we assumed symmetry, all these 8 points are assigned probabilities $\frac{1}{8}$ each. Hence the answers to (i) and (ii) are

$$P\{(HHT)\} = \frac{1}{8} \quad \text{and} \quad P\{(THT)\} = \frac{1}{8}.$$

When the coin is tossed four times, the sample space consists of $2 \times 2 \times 2 \times 2 = 16$ elementary events. Due to symmetry, we assign probabilities $\frac{1}{16}$ to each of these points. Hence

$$P\{(HHHH)\} = \frac{1}{16}.$$

In (iv), the event of getting the sequences $HHHH$ or $HTTT$ means the union of two mutually exclusive events and by the third axiom, the probability is the sum of the probabilities. We have assigned probabilities $\frac{1}{16}$ each, and hence

$$P\{(HHHH \text{ or } HTTT)\} = P\{(HHHH)\} + P\{(HTTT)\} = \frac{1}{16} + \frac{1}{16} = \frac{1}{8}.$$

**Example 1.7.** A balanced die is rolled two times. What is the probability of (i) rolling 9 and (ii) getting a sum greater than or equal to 10?

**Solution 1.7.** When we say “balanced”, it means that we are assuming symmetry in the experiment and we are assigning equal probabilities to all elementary events. Here, there are 36 elementary events and each point will get probability of $\frac{1}{36}$ each. (i) Rolling 9 means the sum of the face numbers is 9. The possible points are $(3,6),(4,5),(5,4),(6,3)$. These are mutually exclusive because, for example, when the sequence $(3,6)$ comes at the same time another sequence cannot come. Let $A$ be the event of rolling 9, and let $A_1$ to $A_4$ denote the events of getting the sequences $(3,6),\dots,(6,3)$, respectively. Then

$$A = A_1 \cup A_2 \cup A_3 \cup A_4$$

and

$$\begin{aligned} A_1 \cap A_2 &= \phi, \quad A_1 \cap A_3 = \phi, \quad A_1 \cap A_4 = \phi, \\ A_2 \cap A_3 &= \phi, \quad A_2 \cap A_4 = \phi, \quad A_3 \cap A_4 = \phi. \end{aligned}$$

That is, they are all mutually exclusive. Hence by the third axiom in the definition of probability

$$P(A) = P(A_1) + P(A_2) + P(A_3) + P(A_4).$$

But we have assigned equal probabilities to elementary events. Hence

$$P(A) = \frac{1}{36} + \cdots + \frac{1}{36} = \frac{4}{36} = \frac{1}{9}.$$

Similarly, let $B$ be the event of getting the sum greater than or equal to 10, which means 10 or 11 or 12. The points favorable to this event are $(4,6),(5,5),(6,4),(5,6)$, $(6,5),(6,6)$ or 6 points are favorable to the event $B$ and since symmetry is assumed

$$P(B) = \frac{6}{36} = \frac{1}{6}.$$

**Rule 1.2.** *Assign probability 0 for almost surely impossible events and probability 1 for almost surely sure events.*

By assigning probability 0, we are not saying that the corresponding event is logically impossible. If an event is logically impossible, then its probability is zero as a consequence of the axioms defining probability. When we assign 1 to almost surely a sure event, we are not saying that the event is a sure event. For a logically sure event, the probability is 1 by the second axiom defining probability. But an assigned probability 1 does not mean that the event is a sure event.

**Rule 1.3.** *If the sample space consists of a continuum of points giving a line segment (or segments) of finite length (or lengths), such as a piece of string of a length of* 50 *cm, and if the experiment is to take a point from this line segment (or segments), such as* *a cut on this string, and if there is no preference of any sort in selecting this point, then assign probabilities proportional to the lengths, taking the total length as unity.*

When a point is selected from a line segment of finite length by using the rule of assigning probabilities proportional to the lengths, then we use the phrase: *a point is selected at random* from the line segment or we have a “random point” from this line segment or if a string is cut by using the above rule we say that we have a *random cut* of the string or we say that the point of cut is uniformly distributed over the line segment. These are all standard phrases used in this situation. Then, if an event $A$ is that the random point lies on a segment of length $m$ units out of a total length of $n$ units, $n \geq m$, then the rule says:

$$P(A) = \frac{m}{n}. \tag{1.8}$$

**Example 1.8.** A random cut is made on a string of 30 cm in length. Marking one end of the string as zero and the other end as 30, what is the probability that (i) the cut is between 10 and 11.7, (ii) the cut is between 10 and 10.001, (iii) the cut is at 10, and (iv) the smaller piece is less than or equal to 10 cm?

**Solution 1.8.** Since we use the phrase “random cut”, we are assigning probabilities proportional to the lengths. Let $x$ be the distance from the end marked 0 to the point of cut. Let $A$ be the event that $A = \{x \mid 10 \leq x \leq 11.7\}$, [this notation means: all values of $x$ such that $x$ is between 10 and 11.7, both the end points are included], $B$ be the event that $B = \{x \mid 10 \leq x \leq 10.001\}$, let $C$ be the event that $C = \{x \mid x = 10\}$ and let $D$ be the event that the smaller piece is less than or equal to 10 cm. The length of the interval in $A$ is $11.7 - 10.0 = 1.7$. Since we are assigning probabilities proportional to the lengths, we have

$$P(A) = \frac{11.7 - 10.0}{30} = \frac{1.7}{30} = \frac{17}{300}$$
$$P(B) = \frac{10.001 - 10.000}{30} = \frac{0.001}{30} = \frac{1}{30\,000}$$

and

$$P(C) = \frac{10 - 10}{30} = \frac{0}{30} = 0.$$

Since we are assigning probabilities proportional to the lengths and since a point does not have any length by definition, then according to this rule the probability that the cut is at a specific point, in a continuum of points, is zero. By assigning this value zero to this probability, we are not saying that it is impossible to cut the string at that point. As per our rule of assigning probabilities proportional to lengths, then since a point does not have length, the point will be assigned probability zero as per this rule. For the event $D$, the smaller piece is of length less than or equal to 10 cm in the following two situations:

$$D_1 = \{x \mid 0 \le x \le 10\} \quad \text{and} \quad D_2 = \{x \mid 20 \le x \le 30\}.$$

Therefore,

$$D = D_1 \cup D_2 \quad \text{where } D_1 \cap D_2 = \phi.$$

Hence

$$\begin{aligned} P(D) = P(D_1) + P(D_2) &= \frac{10-0}{30} + \frac{30-20}{30} \\ &= \frac{10}{30} + \frac{10}{30} = \frac{20}{30} = \frac{2}{3}. \end{aligned}$$

Note that this variable $x$ can be said to be *uniformly distributed over the line segment* $[0, 30]$ in this example.

**Note 1.4.** The above rule cannot be applied if the string is of infinite length such as a beam of light or laser beam or sound wave, etc. How do we compute probabilities in such situations of strings of infinite length?

**Rule 1.4.** *When a point is selected at random from a planar region of finite area, assign probabilities proportional to the area and when a point is selected at random from a higher dimensional space of finite hyper-volume, then assign probabilities proportional to the volume. According to this rule, if the total area is $\alpha$ and out of this, if $\mu(\alpha)$ of the area is favorable to an event $A$, then the probability of $A$ is assumed as*

$$P(A) = \frac{\mu(\alpha)}{\alpha} \tag{1.9}$$

*where $\alpha$ is the Greek letter alpha and $\mu$ is the Greek letter mu. Similarly, if $v$ is the total volume (or hyper-volume) of the space under consideration and if the fraction $\mu(v)$ of $v$ is favorable to an event $A$ then, as per the above rule, the probability of $A$ is taken as*

$$P(A) = \frac{\mu(v)}{v}. \tag{1.10}$$

Several items here need explanations: When a point is taken at random from a planar region of finite area $\alpha$, such as the point of hit of an arrow when the arrow is shot onto a wall of a length 10 meters and a width of 2 meters (area $= \alpha = 10 \times 2 = 20$ sq meters), here "at random" means that there is no preference of any sort for the point to be found anywhere on the planar region. Then we assign probabilities $\frac{\alpha_1}{\alpha}$ to every possible subregion of area $\alpha_1$ with a similar interpretation for higher dimensional situations. The standard terminology for length, area, volume, etc. is the following: length (one dimensional), area (two-dimensional), volume (3-dimensional), hyper-volume

(4 or higher dimensional). For simplicity, we say "volume" for 3 or higher dimensional cases, instead of saying "hyper-volume".

**Example 1.9.** A person trying dart throwing for the first time throws a dart at random to a circular board of a radius of 2 meters. Assuming that the dart hits the board, what is the probability that (1) it hits within the central region of radius 1 meter; (2) it hits along a horizontal line passing through the center and (3) it hits exactly at the center of the board as shown in Figure 1.5?

**Solution 1.9.** Assuming that the point of the hit is a random point on the board, we may assign probabilities proportional to the area. The total area of the board is the area of a circle with a radius of 2 meters:

$$\text{Total area} = \pi r^2 = \pi(2)^2 = 4\pi\,\mathrm{m}^2$$

where the standard notation $\mathrm{m}^2$ means square meters. (1) The area of the central region of the radius of one meter $= \pi(1)^2 = \pi\,\mathrm{m}^2$. Hence the required probability, denoted by $P(A)$, is

$$P(A) = \frac{\pi\,\mathrm{m}^2}{4\pi\,\mathrm{m}^2} = \frac{1}{4}.$$

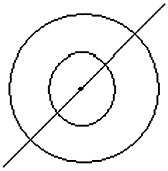


**Figure 1.5:** Circular board and circular, line, point targets.

For answering (2), we have to look at the area along a line passing through the center. But, by definition, a line has no area, and hence the area here is zero. Thus the required probability is $\frac{0}{4\pi} = 0$. In (3) also, a point has no area by definition, and hence the probability is zero.

**Note 1.5.** Note that the numerator and denominator here are in terms of square meters, but probability is a pure number and has no unit of measurement or does not depend on any unit of measurement.

**Note 1.6.** Also when assigning probabilities proportional to the area, remember that lines and points have no areas, and a point has no length or area but a line has length but no area. Similarly, when assigning probabilities proportional to the volume, remember that a planar region has no volume but it has area, and a line has no volume or area but has length, and a point has no length, area or volume.

### 1.5.1 Buffon's "clean tile problem"

**Example 1.10.** Solve Buffon's clean tile problem. That is, a circular coin of diameter $d$ is thrown upward. When it falls on the floor paved with identical square tiles of length $m$ with $d < m$, what is the probability that the coin will fall clean, which means that the coin will not cut any of the edges and corners of the tiles?

**Solution 1.10.** In Figure 1.6, a typical square tile is marked. Since the coin is tossed upward, we assume that the center of the coin could be anywhere on the tile if the coin has fallen on that tile. In other words, we are assuming that the center of the coin is a random point on the square tile or uniformly distributed over that square tile. In Figure 1.6, an inner square is drawn $\frac{d}{2}$ distance away from the boundaries of the outer square. If the center of the coin is anywhere on the boundaries of the inner square or in the region between the walls of the two squares, then the coin can touch or cut the walls of the outer square.

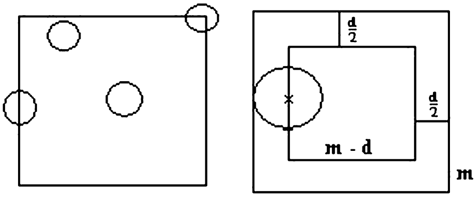


**Figure 1.6:** Square tile and circular coin.

If the center of the coin is strictly within the inner square, then the coin will fall clean. Therefore, the probability of the event, $A$ = the event that the coin falls clean, is given by

$$P(A) = \frac{\text{area of the inner square}}{\text{area of the outer square}} = \frac{(m - \frac{d}{2} - \frac{d}{2})^2}{m^2} = \frac{(m-d)^2}{m^2}. \tag{1.11}$$

This problem is generalized by looking at a floor paved with rectangular tiles of length $m$ units, width $n$ units and a circular coin of diameter $d$ units where $d < m$, $d < n$. This problem can be done in a similar way by looking at the center of the coin and assuming that the center is uniformly distributed over the rectangle. The floor can be paved with any symmetrical object such as a rhombus or general polygon, and a circular coin is tossed. The problem is to compute the probability that the coin will fall clean.

A three-dimensional generalization of the problem is to consider a prism with a square, rectangular, parallelogram or general polygonal base and a ball or sphere of radius $r$ is randomly placed inside the prism. Some illustrations are given in Figure 1.7. What is the probability that the ball will not touch any of the sides, base or top of the prism? When we move from a one-dimensional case to two or higher dimensions, then more axioms such as "invariance" is needed to define probability measures.

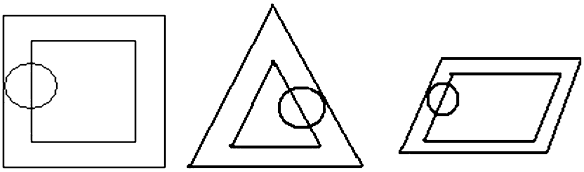


Figure 1.7: Tiles of various shapes.

Another basic problem that Buffon had looked into is called Buffon's needle problem. A floor is paved with parallel lines, $m$ units apart. A headless needle (or a line segment) of length $d$ is tossed up. What is the probability that the needle will touch or cut any of the parallel lines when the needle falls to the floor? There are several situations of interest here. One is the case of a short needle where the length of the needle, $d$, is less than $m$. Another case is when $d < 2m$ and $d > m$. Another case is a long needle which can cut a number of parallel lines. Remember that however long the needle may be there is a possibility that the needle need not cut any of the lines, for example, the needle can fall parallel to the lines.

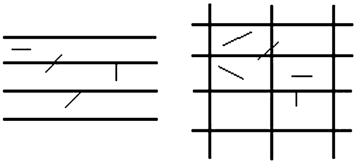


Figure 1.8: Buffon's needle problem.

Another needle problem is when the floor has horizontal and vertical lines making rectangular grids of length $m$ units and width $n$ units and a needle of length $d$ is tossed as shown in Figure 1.8. A generalization of this problem is the case when the needle can be of any shape, and need not be straight.

For dealing with Buffon's needle problem, we need the concepts of random variables and independence of random variables. Hence we will not do examples here.

When we combine geometry with probability, many interesting paradoxes can arise. The subject dealing with the combination of geometry and probability is called *Stochastic Geometry*. Students who are interested in this area can look into the book [4] and other papers of A. M. Mathai.

## Exercises 1.5

**1.5.1.** An unbiased coin is tossed until a head is obtained. What is the probability that the experiment is finished in (i) 4 or less number of trials, (ii) in 20 or less number of trials?

**1.5.2.** A balanced die is rolled 3 times. What is the probability of getting:

(a) sum greater than 14;

(b) all the face numbers are the same;
(c) at least two of the face numbers are the same;
(d) getting the sequences 666 or 121 or 112?

**1.5.3.** An unbiased coin is flipped 3 times. What is the probability of getting (1) exactly one head or (2) at least one head?

**1.5.4.** A box contains 6 identical chips numbered $1, 2, 3, 4, 5, 6$. Two chips are taken one-by-one at random (blind-folded after shuffling well) with replacement. What is the probability that (1) the first number is bigger than the second number? (2) the first number is less than $\frac{1}{2}$ of the second number?

**1.5.5.** What are the probabilities in Exercise 1.5.4 if the sampling is done without replacement?

**1.5.6.** In Exercise 1.5.4, what is the probability that (1) the number in the second trial is bigger than the number in the first trial and (2) the number in the second trial is bigger than that in the first trial, given that the first trial resulted in the number 1?

**1.5.7.** A box contains 7 identical marbles except for the color, 4 are red and 3 are green. Two marbles are picked at random one by one without replacement. What is the probability of getting:
(a) the sequence *RG* (red green);
(b) exactly one red and one green;
(c) *RR* (red red);
(d) exactly 2 red marbles?

**1.5.8.** In Exercise 1.5.7 suppose a subset of 2 marbles is taken at random or blindfolded by putting the hand in the box and taking 2 together. Answer (a), (b), (c), (d).

**1.5.9.** Two identical pieces of string of 20 cm are there. One end of each is marked zero and the other end 20. One string is cut at random. Let $x$ be the distance from zero to the point of cut. The second string is cut at random. Let $y$ be the distance from zero to the point of cut. Find the probability that:
(i) $x < y$, (ii) $x \leq y$, (iii) $x + y \leq 10$, (iv) $x + y \geq 30$,
(v) $10 \leq x \leq 15$, (vi) $5 \leq y \leq 20$, (vii) $5 \leq x \leq 10$ and $10 \leq y \leq 20$,
(viii) $x^2 + y^2 \leq 10$, (ix) $x^2 + y^2 = 10$.

**1.5.10.** A floor is paved with identical square tiles of side 10 cm. A circular coin with a diameter of 2 cm is tossed up. What is the probability that:
(a) the coin will fall clean;
(b) the coin will not fall clean;
(c) the coin will cut exactly one of the edges of the tiles?

**1.5.11.** In Exercise 1.5.10, if the coin is flipped twice, what is the probability that:

(a) on both occasions the coin will fall clean;
(b) in exactly one occasion it falls clean;
(c) on the first occasion it falls clean and on the second occasion it does not fall clean?

**1.5.12.** In Exercise 1.5.10, suppose that the sides of the tiles are $m$ units each and the diameter of the coin is $d$ units. What should the connection be between $m$ and $d$ so that the game is fair, which means the probability of the coin falling clean is the same as the probability it does not fall clean (in such a case, in a game of chance, both people betting on each of the two events of falling clean and not falling clean will have the same chance of winning at each trial).

**1.5.13.** Suppose that the floor is paved with identical rectangular tiles with lengths of 10 cm and a width of 5 cm and a coin with a diameter of 4 cm is tossed. What is the probability that the coin will fall clean?

**1.5.14.** Suppose that a floor is paved with identical rhombuses of side $m$ units and a circular coin of diameter $d$ is tossed. What is the probability that the coin will fall clean if $d$ is small such that it can fall clean?

**1.5.15.** In Exercise 1.5.13, if the floor is paved with identical equilateral triangles, then what will be the corresponding probability?

**1.5.16.** Answer the questions (a) and (b) in Exercise 1.5.10 if the tiles are (1) equilateral triangles with sides of 20 cm each, (2) parallelograms with sides of equal length of 20 cm and (3) hexagons with sides of 20 cm each.

# 2 Probability

## 2.1 Introduction

In Chapter 1, we have introduced the basic notion of probability. In the present chapter, we will explore more properties of probability, the idea of conditional probability, basic notions of independence of events, pair-wise independence, mutual independence, Bayes' theorem, etc. For the computations of probabilities in given situations, we will need some ideas of permutations and combinations. Students may be familiar with these aspects but for the sake of those who are not familiar, or forgotten, a brief description is given here as Sections 2.2 and 2.3. In Section 2.4, a note on sigma and pi notations is given. Those who already know these materials may skip these sections and go directly to Section 2.5.

## 2.2 Permutations

To permute means to rearrange and the number of permutations means the number of such rearrangements. We shall look into the problem of filling up some positions with some objects. For example, let there be $r$ seats in a row and $n$ individuals to be seated on these $r$ seats. In how many different ways can we select individuals from this set of $n$ individuals to occupy these $r$ seats. For example, suppose that there are $r = 2$ seats and $n = 5$ individuals. The first seat can be given to one of the 5 individuals, and hence there are five choices of filling up the first seat. When the first seat is already filled, there are 4 individuals left and one seat is left. Hence the second seat can be filled with one of the four remaining individuals or in 4 different ways. For each of the five choices for the first seat, there are four choices for the second seat. Hence the total number of choices for filling up these two seats is $5 \times 4 = 20$ ways. If $A, B, C, D, E$ denote the five individuals and if the first seat is given to $A$ then the sequences possible for the two seats are $AB$, $AC$, $AD$, $AE$. If $B$ is given the first seat again, there are four such choices, and so on. We can state this as the total number of permutations of five, taken two at a time, and it is 20. We can also state this as the total number of ordered sets of two from a set of five or the total number of sequences of two distinct items taken, from a set of five items.

Thus, if there are $n$ individuals and $r$ seats to be filled, $r \leq n$, then the total number of choices for filling up these $r$ seats with $n$ individuals is $n(n-1)(n-2)\cdots(n-(r-1))$.

**Notation 2.1.** $P(n,r) = {}_nP_r$ = Total number of permutations of $n$, taken $r$ at a time.

**Definition 2.1** (Permutations). The total number of permutations of $n$ distinct objects, taken $r$ at a time or the total number of ordered sets of $r$ distinct items from

https://doi.org/10.1515/9783110562545-002

the set of $n$ distinct items or the total number of sequences of $r$ items from a set of $n$ distinct items is given by

$$P(n,r) = n(n-1)(n-2)\cdots\big(n-(r-1)\big) = n(n-1)\cdots(n-r+1). \tag{2.1}$$

For example, the total number of permutations of 5 items, taken 3 at a time is $5\times 4\times 3 = 60$. The total number of permutations of 5, taken 5 at a time or all is $5\times 4\times 3\times 2\times 1 = 120$:

$$P(1,1) = 1,\quad P(2,1) = 2,\quad P(n,1) = n,\quad P(10,2) = (10)(9) = 90,$$
$$P(4,4) = (4)(3)(2)(1) = 24 = 4!,\quad P(-3,2) = \text{no meaning},$$
$$P\left(2,\frac{1}{2}\right) = \text{no meaning}.$$

**Notation 2.2.** $n!$ = factorial $n$ or $n$ factorial.

**Definition 2.2.**

$$n! = (1)(2)\cdots(n),\quad 0! = 1 \text{ (convention)}.$$

That is,

$$2! = (1)(2) = 2,\quad 3! = (1)(2)(3) = 6,\quad 4! = (1)(2)(3)(4) = 24,$$
$$(-2)! = \text{not defined},\quad \left(\frac{1}{2}\right)! = \text{not defined},$$

and so on. We can also write the number of permutations in terms of factorials:

$$P(n,r) = n(n-1)\cdots(n-r+1) = \frac{n(n-1)\cdots(n-r+1)(n-r)\cdots(2)(1)}{(n-r)(n-r-1)\cdots(2)(1)}$$

by multiplying and dividing by $(n-r)(n-r-1)\cdots(2)(1)$. That is,

$$P(n,r) = \frac{n!}{(n-r)!}. \tag{2.2}$$

If we want this formula, in terms of factorials, to hold for all $n$ then let us see what happens if we compute $P(n,n)$. Writing in terms of factorials, by substituting $r = n$ on the right side of the above equation (2.2), we have

$$P(n,n) = \frac{n!}{0!}.$$

But, from the original definition,

$$P(n,n) = n(n-1)\cdots(2)(1) = n!$$

Hence we need the convention $0! = 1$ if we want to use the representation of $P(n,r)$ in terms of factorials for all $r$ and $n$. [Mathematical conventions are convenient assumptions which will not contradict or interfere with any of the mathematical derivation or computation. Also note that all computations can be carried out without this convention also. If we do not want to use the convention, then at equation (2.2) write $r = 1, \dots, n-1$ and $P(n,n) = n!$.]

As a simple example, we can consider words in a text. Words of a distinct alphabet are an ordered set or ordered sequence of a distinct alphabet. If the alphabet is rearranged or permuted, then we obtain different words. House numbers in a city, postal codes in addresses, etc. are all ordered sets of numbers, and if the numbers are permuted we get other house numbers, other postal codes, etc.

**Example 2.1.** How many different 3-letter words can be made by using all of the alphabet in the word (1) "can", (2) how many different 4-letter words can be made by using all of the alphabet of the word "good", (3) how many 11-letter words can be made by using all of the alphabet in the word "Mississippi"?

**Solution 2.1.** (1) The different words are the following:

can, cna, anc, acn, nac, nca.

There are 6 = 3! such words. In (2), we have the letter "o" repeated 2 times. If the o's were different such as $o_1, o_2$, then the total number of words possible is $4! = 24$. But $o_1o_2$ or $o_2o_1$ will give the same $oo$. Note that $o_1, o_2$ can be permuted in 2! ways and all these permutations will produce the same word. Hence the total number of distinct words possible is

$$\frac{4!}{2!} = 12.$$

In (3), the letter "s" is repeated four times, "i" is repeated 4 times and "p" is repeated 2 times. Hence the total number of distinct words possible is

$$\frac{11!}{4!4!2!} = 34\,650.$$

**Example 2.2.** How many different number plates can be made containing only three digits if (1) repetition of numbers is allowed, (2) no repetition is allowed.

**Solution 2.2.** A number plate with 3 digits means filling up 3 positions with one of the numbers $0, 1, \dots, 9$. The first position can be filled in 10 different ways with one of the 10 numbers $0, 1, \dots, 9$. When a repetition is allowed, the second and third positions can also be filled in 10 different ways. Thus the total number of number plates possible is

$$10 \times 10 \times 10 = 10^3 = 1\,000 \quad \text{when repetition is allowed.}$$

When repetition is not allowed, then the first position can be filled in 10 ways, the second only in 9 ways and the third position in only 8 ways. Thus the total number of number plates possible is

$$10 \times 9 \times 8 = 720 \quad \text{when repetition is not allowed.}$$

We can also write the number of permutations by using the Pochhammer symbol, which is widely used in mathematical analysis.

**Notation 2.3.** $(\alpha)_k$: Pochhammer symbol, where $\alpha$ is the Greek letter alpha.

**Definition 2.3.**

$$(\alpha)_k = \alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k-1), \quad \alpha \neq 0,\ (\alpha)_0 = 1. \tag{2.3}$$

For example,

$$\begin{aligned}
&(1)_n = (1)(2)\cdots(1+n-1) = n!; \quad (2)_3 = (2)(3)(4) = 24;\\
&(-2)_3 = (-2)(-2+1)(-2+2) = 0; \quad \left(\frac{1}{2}\right)_2 = \left(\frac{1}{2}\right)\left(\frac{1}{2}+1\right) = \left(\frac{1}{2}\right)\left(\frac{3}{2}\right) = \frac{3}{4};\\
&(0)_2 = \text{not defined}; \quad (3)_0 = 1; \quad (5)_{-2} = \text{not defined}.
\end{aligned}$$

Note that the various factors in the Pochhammer symbol are in ascending order in the form $a(a+1)(a+2)\cdots$. Suppose we have factors in descending order such as $b(b-1)(b-2)\cdots$ then can we write this also in a Pochhammer symbol. The answer is in the affirmative. Consider the following:

$$\begin{aligned}
b(b-1)\cdots(b-k+1) &= (-1)^k(-b)(-b+1)\cdots(-b+k-1)\\
&= (-1)^k(-b)_k.
\end{aligned} \tag{2.4}$$

With the help of (2.4), we can write the number of permutations in terms of a Pochhammer symbol. The total number of permutations of $n$, taken $r$ at a time, is given by $P(n,r)$ where

$$\begin{aligned}
P(n,r) &= n(n-1)(n-2)\cdots(n-r+1) = (-1)^r(-n)(-n+1)\cdots(-n+r-1)\\
&= (-1)^r(-n)_r.
\end{aligned} \tag{2.5}$$

## Exercises 2.2

**2.2.1.** Evaluate the following numbers of permutations, if possible: (1) $P(4,2)$; (2) $P(3,4)$; (3) $P(-5,2)$; (4) $P(\frac{1}{2},2)$; (5) $P(\frac{3}{2},\frac{1}{2})$.

**2.2.2.** If there are 20 students in a class, then their birthdays could be any one of the 365 days $1, 2, \ldots, 365$. If no two birthdays are the same or if all students have distinct birthdays, then how many possibilities are there?

**2.2.3.** How many 3-letter words can be made by using the alphabets of the word, (1) mind; (2) big, with (a) no letter is repeated, (b) the letter i is present.

**2.2.4.** How many 3-digital number plates can be made (1) with no restriction; (2) no numbers should be repeated; (3) one number 5 must be present; (4) the plate should start with a number 5.

**2.2.5.** In how many ways 10 persons can be seated (a) on the straight line of 4 chairs; (b) on a circular table with 4 chairs?

**2.2.6.** Evaluate the following Pochhammer symbols: (1) $(-5)_2$; (2) $(-5)_5$; (3) $(-\frac{1}{2})_3$; (4) $(\frac{1}{3})_4$.

**2.2.7.** Convert the following number of permutations into Pochhammer notation: (1) $P(5,3)$; (2) $P(10,2)$; (3) $P(5,0)$; (4) $P(5,5)$.

**2.2.8.** From a box containing 3 red and 5 green identical marbles, three marbles are picked at random (i) with replacement; (ii) without replacement. How many sample points are there in the sample space?

**2.2.9.** In Exercise 2.2.8, if we are interested in the event of getting exactly 2 red and one green marble, then how many sample points are there favorable to this event?

**2.2.10.** A coin is tossed 3 times. Write down all possible sequences of head $H$ and tails $T$.

## 2.3 Combinations

In permutations, we were interested in the rearrangement or in sequences or in ordered sets or ordered subsets from the given set of objects. Suppose that we are not interested in the order but only in the subsets. For example, if we have 3 letters $a, b, c$ and if we are looking at the ordered subsets of two letters from these three letters then the ordered sequences are

$$ab, \quad ac, \quad ba, \quad bc, \quad ca, \quad cb$$

or there are $3 \times 2 = 6$ such ordered sets. Suppose that we are only concerned with the subsets of two letters from this set of three letters then the subsets are

$$\{a,b\}, \quad \{a,c\}, \quad \{b,c\}$$

because whether the sequence $ab$ or $ba$ appears it is the same subset of the letters $a$ and $b$.

How many subsets of $r$ elements are possible from a set of $n$ distinct elements? If a subset of $r$ elements is there, then we can order them in $r!$ ways to get all ordered sequences.

Hence the total number of subsets of $r$ elements from a set of $n$ elements = total number of permutations of $n$ taken $r$ at a time, divided by $r!$:

$$= \frac{P(n,r)}{r!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}. \tag{2.6}$$

This is known as the number of combinations of $n$ taken $r$ at a time. The standard notations used are $\binom{n}{r}$, ${}_nC_r$, $C(n,r)$. We will use the notation $\binom{n}{r}$.

**Notation 2.4.** $\binom{n}{r}$ = the number of combinations of $n$, taken $r$ at a time = the number of subsets of $r$ distinct elements from the set of $n$ distinct elements.

**Definition 2.4.** The number of combinations of $n$, taken $r$ at a time or the number of possible subsets of $r$ distinct elements from a set of $n$ distinct elements, is given by

$$\binom{n}{r} = \frac{P(n,r)}{r!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}. \tag{2.7}$$

$$= \frac{(-1)^r(-n)_r}{r!} \quad \text{(in terms of Pochhammer symbol).} \tag{2.8}$$

From this definition itself, the following properties are evident by substituting for $r$:

$$\binom{n}{n} = \binom{n}{0} = 1; \quad \binom{n}{1} = \binom{n}{n-1} = n;$$

$$\binom{n}{2} = \binom{n}{n-2} = \frac{n(n-1)}{2!}, \quad \binom{-3}{2} = \text{not defined}; \quad \binom{\frac{1}{2}}{\frac{1}{4}} = \text{not defined}.$$

From the representation in terms of factorials, we have the following results for all $r$:

$$\binom{n}{r} = \binom{n}{n-r}, \quad r = 0,1,2,\dots,n \quad \Rightarrow \tag{2.9}$$

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1} \tag{2.10}$$

$$\binom{n}{0} = \binom{n}{n}, \quad \binom{n}{1} = \binom{n}{n-1}, \quad \binom{n}{2} = \binom{n}{n-2}, \quad \text{and so on.}$$

For example,

$$\binom{100}{98} = \binom{100}{100-98} = \binom{100}{2} = \frac{(100)(99)}{2!} = 4\,950,$$

$$\binom{210}{210} = \binom{210}{210-210} = \binom{210}{0} = 1,$$

$$\binom{10}{7} = \binom{10}{3} = \frac{(10)(9)(8)}{3!} = \frac{(10)(9)(8)}{(3)(2)(1)} = 120.$$

Note that for the definitions of the numbers of permutations and combinations to hold both $n$ and $r$ must be non-negative integers, $0, 1, 2, \ldots$. When evaluating the number of permutations or the number of combinations in a given situation, do not use the representations in terms of factorials, use the basic definitions, that is,

$$P(n,r) = n(n-1)\cdots(n-r+1) \quad \text{and} \quad \binom{n}{r} = \frac{P(n,r)}{r!} = \frac{n(n-1)\cdots(n-r+1)}{r!}.$$

The representations in terms of factorials are useful for theoretical developments. When factorials are evaluated for large numbers, the computer is not going to give you the correct value. It will print out a number, which is the maximum number that the computer can handle, and does not need to be equal to the value of that factorial. Hence if that large factorial is divided by another big factorial, and not the same as the numerator factorial, the computer will give the value as 1.

**Example 2.3.** A box contains 7 identical marbles, except for color, of which 4 are red and 3 are green. Two marbles are selected at random (a) one by one with replacement; (b) one by one without replacement; (c) two marbles together. (i) Compute the numbers of sample points in these cases; (ii) compute the probabilities of getting the sequence $(RG) = (R = \text{red},\ G = \text{green})$ in (a) and (b); (iii) compute the probabilities of getting exactly one red and one green marbles in (a), (b) and (c).

**Solution 2.3.** (a) It is like filling two positions with 7 objects. The first position can be filled in 7 ways, and since the first marble is put back, the second position can also be filled in 7 ways, and hence the total number of sample points is $7^2 = 49$ and the sample space consists of all such 49 pairs of marbles.

(b) Here, the sampling is done without replacement and hence the first position can be filled in 7 ways and the second in 6 ways because the first marble is not put back. Hence the total number of sample points here is $7 \times 6 = 42$ and the sample space consists of all such 42 pairs of marbles.

(c) Here, we are looking at all possible subsets of 2 items from a set of 7 items, and hence the sample space consists of all such subsets of 2 items and the total number of sample points is

$$\binom{7}{2} = \frac{(7)(6)}{2!} = \frac{42}{2} = 21.$$

In order to compute the probabilities, we will assume symmetry in the outcomes because of the phrase "at random". Hence in (a) all the elementary events get the probabilities of $\frac{1}{49}$ each, in (b) $\frac{1}{42}$ each and in (c) $\frac{1}{21}$ each. Now we need to compute only how many sample points are favorable to the events.

(ii) If the first marble has to be red, then that can only come from the set of red marbles, and hence there are 4 choices to fill the first position and similarly there are

3 choices to fill the second position and the number of sample points favorable to the events in (a) and (b) is $4 \times 3 = 12$. Hence the required probabilities in (a) and (b) are the following:

$$\frac{12}{49} \text{ for (a)} \quad \text{and} \quad \frac{12}{42} \text{ for (b)}.$$

For answering (iii), one has to look into all possible sequences of getting exactly one red and one green in (a) and (b) and all subsets containing exactly one red and one green in (c). Exactly one red and one green can come from the two sequences $RG$ and $GR$ and the number of sample points favorable to the event in (a) and (b) is $4 \times 3 = 12$ plus $3 \times 4 = 12$, equal to 24. Hence the required probabilities in (a) and (b) are the following:

$$\frac{24}{49} \text{ for (a)} \quad \text{and} \quad \frac{24}{42} = \frac{4}{7} \text{ for (b)}.$$

In (c), the total number of sample points favorable to the event of getting exactly one red and one green marble is the following: One red can come only from the set of red marbles and this can be done in $\binom{4}{1} = 4$ ways and similarly the one green can come in $\binom{3}{1} = 3$ ways. Thus the total number of sample points favorable to the event is 12. Hence the required probability of getting exactly one red and one green marble is

$$\frac{12}{21} = \frac{4}{7}.$$

Observe that sampling without replacement and taking a subset of two produced the same result. This, in fact, is a general property.

## Exercises 2.3

**2.3.1.** From a deck of 52 playing cards (13 diamonds, 13 spades, 13 clubs, 13 hearts) a hand of 8 cards is to be taken. (a) How many possibilities are there in making this hand of 8? (b) How many possibilities are there in making a hand of 8 consisting of 5 spades and 3 clubs?

**2.3.2.** A committee of 5 people is to be formed consisting of 3 women and 2 men. There are 10 men and 5 women available for selection. In how many ways can this committee be formed?

**2.3.3.** The 6/36 lottery is where there are 36 specific numbers and 6 numbers will be selected at random one-by-one without replacement or a subset of 6 numbers from the set of 36 numbers is taken. How many points are there in the sample space?

**2.3.4.** The 7/49 lottery consists of 49 specific numbers and a subset of 7 is taken, either together or one-by-one without replacement. (a) How many sample points are there in this experiment? (b) If someone wishes to buy one lottery ticket to play 6/36 or 7/49, should she buy a ticket from 6/36 or 7/49 and why?

**2.3.5.** Show that

$$(1)\colon \sum_{r=0}^{4}\binom{4}{r}=16;\quad (2)\colon \sum_{r=0}^{5}\binom{5}{r}=32;\quad (3)\colon \sum_{r=0}^{n}\binom{n}{r}=2^n.$$

**2.3.6.** Show that

$$\sum_{r=0}^{2}\binom{3}{r}\binom{2}{2-r}=10;\quad \sum_{r=0}^{2}\binom{4}{r}\binom{3}{2-r}=21;$$
$$\sum_{s=0}^{r}\binom{m}{s}\binom{n}{r-s}=\binom{m+n}{r}.$$

**2.3.7.** Suppose there are $r$ indistinguishable balls and $n$ boxes. Balls are put into the boxes without any restriction. A box may receive none, one or more balls. In how many ways $r$ indistinguishable balls can be distributed into $n$ boxes and show that it is $\binom{n+r-1}{r}$.

**2.3.8.** Compute the combinations in Exercise 2.3.7 for (i) $r = 2$, $n = 3$; (ii) $r = 4$, $n = 3$; (iii) Verify the results in (i) and (ii) by the actual count.

**2.3.9.** Evaluate the following sum:

$$\binom{n}{0}+\binom{n}{1}+\cdots+\binom{n}{n}.$$

**2.3.10.** Evaluate the following sum:

$$\binom{n}{0}-\binom{n}{1}+\binom{n}{2}-\cdots+(-1)^n\binom{n}{n}.$$

## 2.4 Sum ∑ and product ∏ notation

**Notation 2.5.** $\sum$: notation for a sum.

**Definition 2.5.** $\sum_{j=1}^{n} a_j = a_1 + a_2 + \cdots + a_n$.

The standard notation used for a sum is $\sum$ (similar to Greek capital letter sigma). For example, if $x$ is any element of the set $\{2, -1, 0, 5\}$, then

$$\sum x = \text{sum of all elements in the set} = (2) + (-1) + (0) + (5) = 6.$$

If $a_1 = 50\,\text{kg}$, $a_2 = 45\,\text{kg}$, $a_3 = 40\,\text{kg}$, $a_4 = 55\,\text{kg}$, $a_5 = 40\,\text{kg}$ denote the weights in kilograms of five sacks of potato, then the total weight of all the five sacks will be the sum, which can be written as

$$\sum_{j=1}^{5} a_j = a_1 + a_2 + a_3 + a_4 + a_5 = 50 + 45 + 40 + 55 + 40 = 230\,\text{kg}.$$

Here, the notation $\sum a_j$ means write the first number, which is $a_1$ or $a_j$ for $j = 1$, add to it the number for $j = 2$, and so on until the last number. Thus

$$\sum_{j=1}^{n} b_j = b_1 + b_2 + \cdots + b_n = \sum_{i=1}^{n} b_i = \sum_{k=1}^{n} b_k,$$

the subscript can be denoted by any symbol $i, j, k$, etc. because in the notation for the sum or called *the sigma notation* the subscript is replaced by $1, 2, \ldots$ and the successive numbers are added up. If $c_1 = \text{Rs}\,100$, $c_2 = \text{Rs}\,250$, $c_3 = \text{Rs}\,150$ are the costs of three items bought by a shopper then the total cost is

$$\sum_{j=1}^{3} c_j = c_1 + c_2 + c_3 = 100 + 250 + 150 = \text{Rs}\,500.$$

If the shopper bought 4 items, all were of the same price Rs 50, then as per our notation

$$\sum_{j=1}^{4} 50 = 50 + 50 + 50 + 50 = 4 \times 50 = \text{Rs}\,200.$$

Thus one property is obvious. If $c$ is a constant, then

$$\sum_{j=1}^{n} c = n \times c = nc. \tag{2.11}$$

Suppose that the first day a person spent $a_1 = \text{Rs}\,20$ for breakfast and $b_1 = \text{Rs}\,35$ for lunch. In the second day, he spent $a_2 = \text{Rs}\,25$ for breakfast and $b_2 = \text{Rs}\,30$ for lunch. Then the total amount spent for the two days is given by

$$\begin{aligned}\sum_{i=1}^{2}(a_i + b_i) &= (a_1 + b_1) + (a_2 + b_2) = \sum_{i=1}^{2} a_i + \sum_{i=1}^{2} b_i \\ &= (20 + 25) + (35 + 30) = \text{Rs}\,110.\end{aligned}$$

Hence another general property is obvious

$$\sum_{j=1}^{n}(a_j + b_j) = \sum_{j=1}^{n} a_j + \sum_{j=1}^{n} b_j. \tag{2.12}$$

Another property is the following:

$$\sum_{j=1}^{k} ca_j = c\sum_{j=1}^{k} a_j = c(a_1 + \cdots + a_k); \quad \sum_{j=1}^{k}(ca_j + db_j) = c\sum_{j=1}^{k} a_j + d\sum_{j=1}^{k} b_j \tag{2.13}$$

where $c$ and $d$ are constants, free of $j$. The average of a set of numbers $x_1, \ldots, x_n$, denoted by $\bar{x}$, can be written as

$$\bar{x} = \frac{(x_1 + \cdots + x_n)}{n} = \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right). \tag{2.14}$$

For example, if the numbers are $2, -3, 5$ then as per our notation

$$x_1 = 2, \quad x_2 = -3, \quad x_3 = 5, \quad n = 3 \quad \text{and} \quad \bar{x} = \frac{(2) + (-3) + (5)}{3} = \frac{4}{3}.$$

Let us see what happens if we consider $\sum_{j=1}^{3}(x_j - \bar{x})$ here:

$$\begin{aligned}\sum_{j=1}^{3}(x_j - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) \\ &= \left(2 - \frac{4}{3}\right) + \left(-3 - \frac{4}{3}\right) + \left(5 - \frac{4}{3}\right)\end{aligned}$$

by adding up all terms by putting $j = 1, j = 2, \ldots$ and the sum is

$$= [(2) + (-3) + (5)] - 3\left(\frac{4}{3}\right) = 4 - 4 = 0.$$

This, in fact, is a general property. Whatever be the numbers $x_1, x_2, \ldots, x_n$:

$$\sum_{j=1}^{n}(x_j - \bar{x}) = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \bar{x} = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} x_j = 0 \tag{2.15}$$

since $\bar{x}$ is free of $j$ it acts as a constant and

$$\sum_{j=1}^{n} \bar{x} = n\bar{x} = n\frac{(\sum_{j=1}^{n} x_j)}{n} = \sum_{j=1}^{n} x_j.$$

Whatever be the numbers $x_1, x_2, \ldots, x_n$,

$$\sum_{j=1}^{n} x_j^2 = x_1^2 + x_2^2 + \cdots + x_n^2; \tag{2.16}$$

$$\begin{aligned}\left(\sum_{j=1}^{n} x_j\right)^2 &= (x_1 + \cdots + x_n)^2 = x_1^2 + \cdots + x_n^2 + 2x_1x_2 + \cdots + 2x_1x_n \\ &\quad + 2x_2x_3 + \cdots + 2x_2x_n + \cdots + 2x_{n-1}x_n \\ &= \sum_{j=1}^{n} x_j^2 + 2\sum_{i<j} x_ix_j = \sum_{j=1}^{n} x_j^2 + 2\sum_{i>j} x_ix_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_ix_j. \end{aligned} \tag{2.17}$$

For example, the sum $\sum_{i<j} x_ix_j$ means to take the sum of product of all terms where the first subscript is less than the second subscript or $i < j$. It is a double sum involving $i$ and $j$ but subject to the condition $i < j$. That is, for example,

$$(x_1 + x_2 + x_3)^2 = x_1^2 + x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3$$

which is the same as saying

$$(x_1 + x_2 + x_3)^2 = \sum_{j=1}^{3} x_j^2 + 2\sum_{i<j} x_ix_j$$

$$\begin{aligned}
&= x_1^2 + x_2^2 + x_3^2 + 2x_3x_1 + 2x_2x_1 + 2x_3x_2 \\
&= \sum_{j=1}^{3} x_j^2 + 2\sum_{i>j} x_ix_j \\
&= x_1^2 + x_2^2 + x_3^2 + 2x_2x_1 + 2x_3x_1 + 2x_3x_2 \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} x_ix_j \\
&= x_1x_1 + x_2x_2 + x_3x_3 + x_1x_2 + x_2x_1 \\
&\quad + x_1x_3 + x_3x_1 + x_2x_3 + x_3x_2
\end{aligned}$$

which is the same as saying the double sum without any restriction on $i$ and $j$, that is, $\sum_{i=1}^{3}\sum_{j=1}^{3}(x_ix_j)$. Some of the general properties of the sigma notation are the following: For any set of numbers $a_1, a_2, \ldots, b_1, b_2, \ldots$

$$\sum_{i=1}^{n}(a_ib_i) = a_1b_1 + a_2b_2 + \cdots + a_nb_n; \tag{2.18}$$

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n}(a_ib_j) &= \sum_{i=1}^{m} a_i\Big[\sum_{j=1}^{n} b_j\Big] \\
&= \sum_{i=1}^{m} a_i[b_1 + \cdots + b_n] \\
&= [a_1 + \cdots + a_m][b_1 + \cdots + b_n] = [b_1 + \cdots + b_n][a_1 + \cdots + a_m] \\
&= \sum_{j=1}^{n}\sum_{i=1}^{m}(b_ja_i)
\end{aligned} \tag{2.19}$$

or, in other words, we could have opened up the sum with respect to $i$ first or $j$ first the result would have remained the same.

If we have two or more subscripts, the sigma notation will be the same type. For example, let $w_{i,j}$ (which is also written as $w_{ij}$ without the comma between the subscripts if there is no possibility of confusion) be the weight of the $i$-th individual in the $j$-th age group. Suppose we have numbered some people from 1 to 40, say, $i = 1, \ldots, 40$ and categorized into 5 categories according to their ages such as age less than or equal to 20 in group 1, greater than 20 but less than or equal to 30 in group 2, greater than 30 but less than or equal to 40 in group 3, greater than 40 but less than or equal to 50 in group 4, greater than 50 in group 5. Then $j = 1, 2, 3, 4, 5$. We have $w_{10,5}$ the weight of the 10th person in the 5th age group, and if her weight is 55 kg then $w_{10,5} = 55$. Then the total weight of all the individuals is given by

$$\begin{aligned}
\sum_{i=1}^{40}\sum_{j=1}^{5} w_{i,j} &= \sum_{i=1}^{40}\Big[\sum_{j=1}^{5} w_{i,j}\Big] \\
&= \sum_{j=1}^{5}\Big[\sum_{i=1}^{40} w_{i,j}\Big]
\end{aligned}$$

that is, we can sum up *i* first or *j* first and it is also

$$
\begin{aligned}
&= w_{1,1} + w_{1,2} + \cdots + w_{1,5} + w_{2,1} + \cdots + w_{2,5} + \cdots + w_{40,5} \\
&= w_{1,1} + w_{2,1} + \cdots + w_{40,1} + w_{1,2} + \cdots + w_{40,2} + \cdots + w_{40,5}.
\end{aligned}
$$

Thus we have the following general rule:

$$
\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} &= \sum_{i=1}^{m}\left[\sum_{j=1}^{n} x_{ij}\right] \\
&= \sum_{j=1}^{n}\left[\sum_{i=1}^{m} x_{ij}\right].
\end{aligned}
\tag{2.20}
$$

### 2.4.1 The product notation or pi notation

Just like the notation for a sum, calling it the sigma notation, we have a notation for a product, calling it the pi notation.

**Notation 2.6.** $\prod$: the product notation.

**Definition 2.6.**

$$\prod_{j=1}^{n} a_j = a_1 \times a_2 \times \cdots \times a_n = a_1 a_2 \cdots a_n.$$

For example, if $a_1 = 5$, $a_2 = -1$, $a_3 = 2$ then

$$\prod_{j=1}^{3} a_j = a_1 a_2 a_3 = (5)(-1)(2) = -10.$$

If $a_1 = a_2 = a_3 = 5$, then $\prod_{j=1}^{3} a_j = (5)(5)(5) = 5^3 = 125$. Thus, in general we have the following result:

$$\prod_{j=1}^{n} c = c^n; \quad \prod_{j=1}^{n} c \neq c \prod_{j=1}^{n} 1 \tag{2.21}$$

whenever *c* is a constant.

$$\prod_{j=1}^{n}(a_j + b_j) = (a_1 + b_1)(a_2 + b_2)\cdots(a_n + b_n) \neq \prod_{j=1}^{n} a_j + \prod_{j=1}^{n} b_j. \tag{2.22}$$

$$\left(\prod_{i=1}^{n} a_i\right)\left(\prod_{j=1}^{m} b_j\right) = \left(\prod_{j=1}^{m} b_j\right)\left(\prod_{i=1}^{n} a_i\right) = a_1 \cdots a_n b_1 \cdots b_m. \tag{2.23}$$

But if the brackets are not there, then let us see what happens.

$$\prod_{i=1}^{2} a_i \prod_{j=1}^{3} b_j = \left(a_1 \prod_{j=1}^{3} b_j\right)\left(a_2 \prod_{j=1}^{3} b_j\right) = a_1 a_2 \left(\prod_{j=1}^{3} b_j\right)^2$$

opening up $i$ first, and it is also $= a_1 a_2 (b_1 b_2 b_3)^2$. Let us see what happens if we open up $j$ first.

$$\begin{aligned}\prod_{i=1}^{2} a_i \prod_{j=1}^{3} b_j &= \left(\prod_{i=1}^{2} a_i b_1\right)\left(\prod_{i=1}^{2} a_i b_2\right)\left(\prod_{i=1}^{2} a_i b_3\right)\\ &= b_1 b_2 b_3 \left(\prod_{i=1}^{2} a_i\right)^3 = b_1 b_2 b_3 (a_1 a_2)^3.\end{aligned}$$

Hence it is clear that if product notations are written without brackets then a product need not determine a unique quantity, or the notation becomes meaningless. Hence remember to put proper brackets at appropriate places, otherwise the notation can become meaningless:

$$\prod_{i=1}^{2}(a_i - c) = (a_1 - c)(a_2 - c) = a_1 a_2 - c(a_1 + a_2) + c^2 \neq \prod_{i=1}^{2} a_i - \prod_{i=1}^{2} c = a_1 a_2 - c^2.$$

$$\prod_{j=1}^{n}(x - a_j) = (x - a_1)(x - a_2)\cdots(x - a_n). \tag{2.24}$$

$$\prod_{j=1}^{n} c a_j = (ca_1)(ca_2)\cdots(ca_n)] = c^n a_1 \cdots a_n \neq c \prod_{j=1}^{n} a_j.$$

## Exercises 2.4

**2.4.1.** If $x \in \{3, 5, 0\}$ and $y \in \{-2, -5, 0, 6\}$, then compute (i) $\sum x$; (ii) $\sum y$; (iii) $\sum(x + y)$.

**2.4.2.** If $x_1 = 2$, $x_2 = -3$, $x_3 = 0$, $x_4 = 5$, then compute

(i) $\bar{x}$;

(ii) $\sum_{i=1}^{4}(x_i - \bar{x})$;

(iii) $\sum_{i=1}^{4} x_i^2$;

(iv) $\sum_{i=1}^{4}(x_i - \bar{x})^2$;

(v) $\sum_{i=1}^{4} |x_i|$ (absolute value means the magnitude without the sign, when $x_i$ is real, that is $|6| = 6$, $|-6| = 6$, $|0| = 0$, $|-\frac{1}{2}| = \frac{1}{2}$, $|-3^2| = 3^2 = 9$);

(vi) $\sum_{i=1}^{4} |x_i - \bar{x}|$.

**2.4.3.** For general numbers $x_1, \dots, x_n$, derive the following general results:

$$\sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - n(\bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2;$$

$$\left[\sum_{j=1}^{n}(x_j - \bar{x})^2\right]^{\frac{1}{2}} \neq \sum_{j=1}^{n}(x_j - \bar{x}); \tag{2.25}$$

$$\left|\sum_{j=1}^{n}(x_j - \bar{x})\right| \neq \sum_{j=1}^{n}|x_j - \bar{x}|.$$

[Give counter examples wherever something is to be disproved.]

**2.4.4.** Evaluate the following:
(i) $\prod_{i=1}^{5}(10)$;
(ii) $\prod_{i=1}^{3}(a_i - 3)$ where $a_1 = 5$, $a_2 = 0$, $a_3 = -2$;
(iii) $\prod_{i=1}^{3}(a_i - b_i)$ where $(a_1, b_1) = (2, 3)$, $(a_2, b_2) = (5, 1)$, $(a_3, b_3) = (1, -2)$, and show that it is not equal to $\prod_{i=1}^{3} a_i - \prod_{i=1}^{3} b_i$.

**2.4.5.** Write the following by using a double product notation:

$$(a_1 - a_2)(a_1 - a_3)\cdots(a_1 - a_n)(a_2 - a_3)\cdots(a_2 - a_n)\cdots(a_{n-1} - a_n).$$

**2.4.6.** Open up the following and write as a sum:
(i) $(x - a_1)(x - a_2)(x - a_3)$;
(ii) $(x - a_1)(x - a_2)\cdots(x - a_n)$.

**2.4.7.** Open up the following:
(i) $(\prod_{i=1}^{2} a_i)(\sum_{j=1}^{3} b_j)$;
(ii) $(\prod_{i=1}^{2} a_i)\sum_{j=1}^{3} b_j$;
(iii) $\prod_{i=1}^{2} a_i(\sum_{j=1}^{3} b_j)$;
(iv) $\prod_{i=1}^{2} a_i \sum_{j=1}^{3} b_j$.

**2.4.8.** Evaluate the following:
(i) $\prod_{i=1}^{2}[\prod_{j=1}^{2}(a_i - b_j)]$;
(ii) $\prod_{j=1}^{2}[\prod_{i=1}^{2}(a_i - b_j)]$.

**2.4.9.** If $a_1 = 2$, $a_2 = -1$ then evaluate
(i) $(\prod_{i=1}^{2} a_i)^2$ and
(ii) $\prod_{i=1}^{2} a_i^2$.

**2.4.10.** For paired values $(x_1, y_1), \ldots, (x_n, y_n)$ show that

$$\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y}.$$

**Example 2.4.** In Example 2.3, suppose that the marbles are taken at random, one by one, without replacement. What is the probability that (a) the second marble taken is green, given that the first marble removed is a red marble? (b) the second marble is green?

**Solution 2.4.** Let $A$ be the event that the first marble removed is red and let $B$ be the event that the second marble is green. (a) If it is already known that the first marble removed is red then there are only 6 marbles left in the box, out of which 3 are green, and hence the required probability is $\frac{3}{6} = \frac{1}{2}$. That is,

$$P(B \text{ given that } A \text{ has occurred}) = \frac{3}{6} = \frac{1}{2}.$$

What is $A \cap B$ here? This is the event that the first marble is red and the second marble is green or getting the sequence $RG$. This probability is already evaluated in Example 2.3. That is,

$$P(A \cap B) = \frac{12}{42} = \frac{2}{7}.$$

What is the probability $P(A)$ that the first marble is red? This can be computed either looking at the first trial alone, where there are 7 marbles out of which 4 are red, and hence the probability is $P(A) = \frac{4}{7}$. We can also look at it after the completion of the experiment of taking two marbles one by one without replacement. Then the first marble is red if we have the sequence $RG$ or $RR$. The total number of points favorable to this event is $RG$ giving $4 \times 3 = 12$ plus $RR$ giving $4 \times 3 = 12$, and hence the probability

$$P(A) = \frac{12 + 12}{42} = \frac{4}{7}.$$

One interesting property may be noted from the above. That is,

$$P(A \cap B) = \frac{12}{42} = P(A)P(B \text{ given that } A \text{ has occurred}) = \frac{4}{7} \times \frac{3}{6}.$$

This is a general property that we shall discuss next, after checking (b).

(b) Here, we need the probability for the second marble to be green. This can happen in two ways, under the sequence $RG$ or $GG$. The sample points favorable to these two sequences is $4 \times 3 = 12$ plus $3 \times 2 = 6$ or 18. Hence

$$P(B) = \frac{18}{42} = \frac{3}{7}.$$

It is equivalent to taking one marble at random and the probability for that marble being green.

## 2.5 Conditional probabilities

We will examine probability of the type $B$ given $A$ or the probability of an event given that some other event has occurred. In some cases that information will change the probability, that is, the probability of a conditional statement and that of an unconditional statement may be different, as seen from Example 2.4. We will introduce a formal notation and definition for such conditional statements here.

**Notation 2.7.** $P(B|A)$ = probability of $B$ given $A$ = the conditional probability of $B$ given that $A$ has already occurred, where $A$ and $B$ are two events in the same sample space.

Here, the notation is a vertical bar after $B$ and it should not be written as $B/A$ or $\frac{B}{A}$ and these have no meaning when $A$ and $B$ are events. Conditional probability can be defined in terms of the probability for simultaneous occurrence and the marginal probability or the probability of the conditioned event.

**Definition 2.7.** The conditional probability of $B$ given $A$ is the probability of the simultaneous occurrence of $B$ and $A$ divided by the probability of $A$ when $P(A) \neq 0$. That is,

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, \quad P(A) \neq 0, \quad \Rightarrow \quad P(A \cap B) = P(A)P(B|A) \quad \Rightarrow$$
$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B), \quad P(A) \neq 0, \quad P(B) \neq 0. \tag{2.26}$$

Thus the probability of intersection can be written as the conditional probability times the marginal probability of the conditioned event. This rule can be extended to any number of events:

$$\begin{aligned} P(A \cap B \cap C) &= P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C), \\ &P(B \cap C) \neq 0, \quad P(C) \neq 0. \end{aligned} \tag{2.27}$$

Extending this result, we have

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_k) &= P(A_1|A_2 \cap A_3 \cap \cdots \cap A_k), \\ P(A_2 \cap \cdots \cap A_k) &\neq 0 \\ &= P(A_1|A_2 \cap \cdots \cap A_k)P(A_2|A_3 \cap \cdots \cap A_k) \\ &\quad \times \cdots P(A_{k-1}|A_k)P(A_k), \\ P(A_2 \cap \cdots \cap A_k) &\neq 0, \quad \ldots, \quad P(A_k) \neq 0. \end{aligned} \tag{2.28}$$

**Example 2.5.** A box contains 4 red and 3 green identical marbles. Marbles are taken at random one by one (a) without replacement; (b) with replacement. What is the probability of getting (i) the sequence $RRG$; (ii) the sequence $RGR$; (iii) exactly 2 red and one green marble.

**Solution 2.5.** (i)(a) Let the marbles be selected at random without replacement. Let $A$ be the event that the first marble is red, $B$ be the event that the second marble is red and $C$ be the event that the third marble is green. Then the sequence $RRG$ means $A \cap B \cap C$. By using the rule in (2.28), we have

$$\begin{aligned}P(A\cap B\cap C)=P(C|B\cap A)P(B|A)P(A)&=\frac{3}{5}\times\frac{3}{6}\times\frac{4}{7}\\ =P(A)P(B|A)P(C|A\cap B)&=\frac{4}{7}\times\frac{3}{6}\times\frac{3}{5}=\frac{6}{35}.\end{aligned}$$

For the first marble being red is $P(A)=\frac{4}{7}$ because there are 7 marbles out of which 4 are red and the marbles are picked at random. If one red marble is removed, then the probability of getting another red marble is $P(B|A)=\frac{3}{6}$ because there are only 6 marbles left out of which 3 are red. If two red marbles are removed, then there are only 5 marbles out of which 3 are green, and hence $P(C|A\cap B)=\frac{3}{5}$. By a similar argument, the probability in (ii)(a) is

$$P(\{RGR\})=\frac{4}{7}\times\frac{3}{6}\times\frac{3}{5}=\frac{6}{35}.$$

(i), (ii)(b) Let the marbles be taken with replacement. If marbles are returned each time, then the probability remains the same. Then probability of getting a red in any trial is $\frac{4}{7}$ and the probability of getting a green in any trial is $\frac{3}{7}$. The occurrence or non-occurrence of an event in the first trial does not affect the probability of occurrence of an event in the second trial, as so on. Again, by using the same formula (2.27), we have

$$P(\{RRG\})=\frac{4}{7}\times\frac{4}{7}\times\frac{3}{7}=\left(\frac{4}{7}\right)^2\left(\frac{3}{7}\right);\quad P(\{RGR\})=\frac{4}{7}\times\frac{3}{7}\times\frac{4}{7}=\left(\frac{4}{7}\right)^2\left(\frac{3}{7}\right).$$

(iii) Note that exactly 2 red and one green, out of three marbles taken can come in

$$\binom{3}{2}=\binom{3}{1}=3$$

ways. These are the sequences $RRG, RGR, GRR$. By using (2.27), we see that for each of these sequences the probability remains the same. Hence the probabilities for (iii)(a) and (iii)(b) are respectively,

$$3\times\frac{6}{35}=\frac{18}{35}\quad\text{and}\quad 3\times\left(\frac{4}{7}\right)^2\left(\frac{3}{7}\right).$$

**Example 2.6.** In Example 2.5, suppose that three marbles are taken together at random. What is the probability of getting exactly 2 red and one green marbles?

**Solution 2.6.** This is a matter of selecting subsets of size 3 or of 3 elements. The total number of sample points possible is

$$\binom{7}{3}=\frac{7\times6\times5}{1\times2\times3}=35.$$

The total number of sample points favorable to the event is

$$\binom{4}{2}\binom{3}{1}=\left(\frac{4\times3}{1\times2}\right)\times(3)=18$$

because the red marbles can come only from the set of red marbles and there are 4 red and a subset of 2 is taken, which can be done in $\binom{4}{2}$ ways and similarly the green marbles can be selected in $\binom{3}{1}$ ways. Note that for each selection of red, the green can be selected in $\binom{3}{1}$ ways and vice versa, and hence the total number of sample points favorable to the event is the product of the two combinations. Hence the required probability in (iii)(a) is

$$\frac{\binom{4}{2}\binom{3}{1}}{\binom{7}{3}} = \frac{18}{35}.$$

When sampling is done with replacement then the probability of getting a red marble at any trial is $\frac{4}{7}$ and the probability of getting a green marble at any trial is $\frac{3}{7}$. The total number of ways of getting 2 red or 1 green in 3 trials is $\binom{3}{1} = \binom{3}{2}$. Hence the answer for (iii)(b) is

$$\binom{3}{2}\left(\frac{4}{7}\right)^2\left(\frac{3}{7}\right) = 3\left(\frac{4}{7}\right)^2\left(\frac{3}{7}\right).$$

**Definition 2.8** (Statistical independence or product probability property (*PPP*))**.** If

$$P(A \cap B) = P(A)P(B) \tag{2.29}$$

then the events $A$ and $B$ are said to be independent events or said to satisfy the product probability property. If three events $A, B, C$ are such that

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C) \tag{2.30}$$

then $A, B, C$ are said to be *pairwise independent events*. In addition to (2.30) if further,

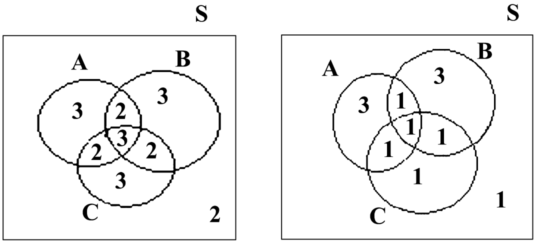
$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

then the events $A, B, C$ are said to be *mutually independent events*. Pairwise independence need not imply mutual independence. A set of events $A_1, A_2, \ldots, A_k$ are said to be mutually independent events if for all subsets of the set $\{A_1, \ldots, A_k\}$ the product probability property holds or the probability of the intersection is the product of the probabilities of individual events, that is,

$$P(A_{i_1} \cap \cdots \cap A_{i_r}) = P(A_{i_1}) \cdots P(A_{i_r}) \tag{2.31}$$

for all different subscripts $(i_1, \ldots, i_r)$, $r = 2, \ldots, k$. This means for every intersection of two, three, …, $k$ distinct events the probability of the intersection is the product of the individual probabilities.

From the following figure, it can be seen that pair-wise independence need not imply mutual independence.



(a) (b)

**Figure 2.1:** Pairwise and mutual independence.

In Figure 2.1 (a), a sample space with symmetry in the outcomes and with 20 sample points, three events $A, B, C$, is given. The numbers in the various regions indicate the numbers of points falling in various regions. Each of the 20 sample points has a probability of $\frac{1}{20}$ each. A total of 10 sample points fall in each of $A, B$ and $C$. Five points each fall in the intersections $A \cap B$, $A \cap C$, $B \cap C$, three sample points fall in $A \cap B \cap C$ and two sample points are in the complementary region of $A \cup B \cup C$. Note that

$$\begin{aligned}
P(A) &= \frac{10}{20} = \frac{1}{2} = P(B) = P(C);\\
P(A \cap B) &= \frac{5}{20} = \frac{1}{4} = P(A)P(B);\\
P(A \cap C) &= \frac{5}{20} = \frac{1}{4} = P(A)P(C);\\
P(B \cap C) &= \frac{5}{20} = \frac{1}{4} = P(B)P(C);\\
P(A \cap B \cap C) &= \frac{3}{20} \neq P(A)P(B)P(C) = \frac{1}{8}.
\end{aligned}$$

Hence $A, B, C$ are pair-wise independent but not mutually independent.

Some students may be thinking that $P(A \cap B \cap C) = P(A)P(B)P(C)$ is sufficient to guarantee mutual independence. This is not sufficient. In Figure 2.1 (b), let us assume symmetry and let the numbers of elementary events be as shown there in the sets $A, B, C$, 6 in $A$, 6 in $B$, 4 in $C$ and one outside, thus a total of 12 points. Then

$$\begin{aligned}
P(A) &= \frac{6}{12} = \frac{1}{2}; \quad P(B) = \frac{6}{12} = \frac{1}{2}; \quad P(C) = \frac{4}{12} = \frac{1}{3};\\
P(A \cap B \cap C) &= \frac{1}{12} = P(A)P(B)P(C);\\
P(A \cap B) &= \frac{2}{12} = \frac{1}{6} \neq P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.
\end{aligned}$$

Hence $P(A \cap B \cap C) = P(A)P(B)P(C)$ need not imply $P(A \cap B) = P(A)P(B)$.

**Note 2.1.** Independence of events should not be confused with mutually exclusive events. The phrase "independent" is one of the unfortunate terms in statistical literature. This can create a wrong impression in the minds of students as if the events have nothing to do with each other or they are mutually exclusive. When we say that the events $A$ and $B$ are independent they depend on each other a lot, the dependence is in the form of a product probability or the probability of intersection is the product of the individual probabilities, that is,

$$P(A \cap B) = P(A)P(B).$$

Hence the students may observe that $A$ and $B$ depend on each other through this product probability property (*PPP*), and hence this author has suggested to replace "independence of events" with events satisfying product probability property. The students must keep in mind that

*independence of events has nothing to do with mutually exclusiveness of events.*

Two events can be mutually exclusive and not independent or mutually exclusive and independent or not mutually exclusive and independent or not mutually exclusive and not independent.

Now the students may wonder from where this word "independent" originated. This has to do with conditional statements. We had defined conditional probability of $A$ given $B$ as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{for } P(B) \neq 0.$$

Now, if the product probability property holds then $P(A \cap B) = P(A)P(B)$. Then in this case

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A) \quad \text{when } P(B) \neq 0. \tag{2.32}$$

This means that conditional probability of $A$ given $B$ is the same as the marginal or unconditional probability of $A$ when the product probability property holds. In other words, the probability of $A$ is not affected by the occurrence or non-occurrence of $B$ and in this sense, independent of the occurrence of $B$. This is from where the word "independent" came in. But this word has created a lot of confusion when this concept is applied in practical situations. Hence it is much safer to say when the product probability property or *PPP* holds instead of saying when there is independence. We have given examples of sampling with replacement where *PPP* holds or where the events are independent.

### 2.5.1 The total probability law

Two important results on conditional probability are the total probability law and Bayes' theorem. Both deal with a partitioning of the sample space. Let a sample space be partitioned into mutually exclusive and totally exhaustive events $A_1, \dots, A_k$ and let $B$ be any other event in the same sample space as in Figure 2.2. From the Venn diagram one may note that $B$ is partitioned into mutually exclusive pieces $B \cap A_1, B \cap A_2, \dots$, $B \cap A_k$, some of which may be empty. That is,

$$\begin{aligned} S &= A_1 \cup A_2 \cup \cdots \cup A_k. \\ A_i \cap A_j &= \phi, \quad \text{for all } i \neq j = 1, \dots, k. \\ B &= (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k), \\ (B \cap A_i) \cap (B \cap A_j) &= \phi, \quad \text{for all } i \neq j = 1, \dots, k. \end{aligned}$$

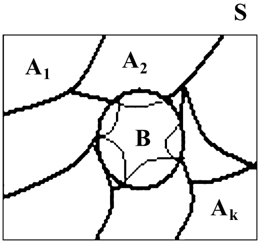


**Figure 2.2:** Total probability law.

Hence by the second and third axioms of probability we have

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_k) \tag{2.33}$$

which can be written, by using conditional probability, as

$$\begin{aligned} P(B) = {} & P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \\ & + \cdots + P(B|A_k)P(A_k). \end{aligned} \tag{2.34}$$

for $P(A_j) \neq 0$, $j = 1, \dots, k$. This equation (2.34) is known as the *total probability law* where the probability of the event $B$ is split into a sum of conditional probabilities of $B$ given $A_1, \dots, A_k$ and marginal probabilities of $A_1, \dots, A_k$. It is a probability law connecting conditional probabilities and marginal probabilities when the marginal probabilities are non-zeros. We can get many interesting applications of this probability law.

**Example 2.7.** Dr Joy is not a very good medical practitioner. If a patient goes to him, the chance that he will diagnose the patient's symptoms properly is 30%. Even if the diagnosis is correct his treatment is such that the chance of the patient dying is 60% and if the diagnosis is wrong the chance of the patient dying is 95%. What is the probability that a patient going to Dr Joy dies during treatment?

**Solution 2.7.** Let $A_1$ be the event of a correct diagnosis, and $A_2$ that of a wrong diagnosis. Then $A_1 \cap A_2 = \phi$, $A_1 \cup A_2 = S$ the sure event. Let $B$ be the event of a patient of Dr Joy dying. Then the following probabilities are given:

$$P(A_1) = 0.3, \quad P(A_2) = 0.7, \quad P(B|A_1) = 0.6, \quad P(B|A_2) = 0.95.$$

We are asked to compute the probability of $B$. By the total probability law,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) = (0.6)(0.3) + (0.95)(0.7) = 0.845$$

or the chance of the patient dying is 84.5%.

**Example 2.8.** Mr Narayanan is a civil engineer with Kerala government. He is asked to design an over bridge (sky way). The chance that his design is going to be faulty is 60% and the chance that his design will be correct is 40%. The chance of the over bridge collapsing if the design is faulty is 90%; otherwise, due to other causes, the chance of the over bridge collapsing is 20%. What is the chance that an over bridge built by Mr Narayanan will collapse?

**Solution 2.8.** Let $A_1$ be the event that the design is faulty and $A_2$ be the event that the design is not faulty. Then $A_1 \cap A_2 = \phi$ and $A_1 \cup A_2 = S$ a sure event. Let $B$ be the event of the over bridge collapsing. Then we are given the following:

$$P(A_1) = 0.6, \quad P(A_2) = 0.4, \quad P(B|A_1) = 0.9, \quad P(B|A_2) = 0.2.$$

We are asked to compute the probability of $B$. From the total probability law,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) = (0.9)(0.6) + (0.2)(0.4) = 0.62.$$

There is a 62% chance of the over bridge designed by Mr Narayanan collapsing.

### 2.5.2 Bayes' rule

Consider again the partitioning of the sample space into mutually exclusive and totally exhaustive events $A_1, \ldots, A_k$ and let $B$ be any event in the same sample space. In the total probability law, we computed the probability of $B$. Let us look into any one intersection of $B$ with the $A_j$'s, for example, consider $B \cap A_1$. From the definition of conditional probability, we can write

$$P(B \cap A_1) = P(A_1|B)P(B), \quad P(B) \neq 0.$$

Therefore,

$$\begin{aligned} P(A_1|B) &= \frac{P(B \cap A_1)}{P(B)}, \quad P(B) \neq 0 \\ &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)}, \qquad (2.35) \\ P(A_j) \neq 0, &\quad j = 1, \ldots, k. \end{aligned}$$

This equation (2.35) is known as *Bayes' rule, Bayes' law or Bayes' theorem*. It is named after a Christian priest, Rev. Bayes, who discovered this rule. The beauty of the result can be seen from many perspectives. It can be interpreted as a rule connecting prior and posterior probabilities in the sense that probabilities of the type $P(A_j|B)$ can be interpreted as posterior probabilities or the probability of the event $A_j$ computed after observing the event $B$ and $P(A_j)$ can be called prior probability of $A_j$ or the probability of $A_j$ computed before observing the event $B$. Bayes' rule also provides an inverse reasoning or establishes a connection between probabilities of the type $P(A_1|B)$ and $P(B|A_1)$, where one can be interpreted as the probability from cause to effect and the other from effect to cause.

If a patient died and if the relatives of the patient felt that the medical doctor attending to the patient was incompetent or the hospital was negligent, then they would like to have an estimate of the chance that the patient died due to the negligence or incompetence of the doctor, etc. What is the probability that the diagnosis was wrong given that the patient died? In the case of a bridge collapsing, the concerned general public may want to know the chance that the engineer's design was in fact faulty in the light of the bridge collapsing.

**Example 2.9.** In Example 2.7, what is the probability that Dr Joy's diagnosis was wrong in the light of a patient of Dr Joy dying?

**Solution 2.9.** Here, we are asked to compute the probability $P(A_2|B)$. But

$$\begin{aligned}P(A_2|B) &= \frac{P(B|A_2)P(A_2)}{P(B)} = \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)}\\ &= \frac{(0.95)(0.7)}{0.845} = \frac{0.665}{0.845} = \frac{133}{169} \approx 0.787.\end{aligned}$$

There is approximately a 78.7% chance that the doctor's diagnosis was wrong. There is a very good chance of a successful lawsuit against the doctor.

**Example 2.10.** In Example 2.8, what is the probability that the design of Mr Narayanan was faulty, in the light of an over bridge designed by him collapsing?

**Solution 2.10.** Here, we are asked to compute the probability $P(A_1|B)$. From Bayes' rule, we have

$$\begin{aligned}P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)}\\ &= \frac{(0.9)(0.6)}{(0.62)} = \frac{54}{62} \approx 0.87.\end{aligned}$$

There is approximately a 87% chance that Mr Narayanan's design was faulty. But there is no way of taking any action against Mr Narayanan due to job security in the Kerala system even if several people also died due to the collapse of the over bridge.

### 2.5.3 Entropy

Another concept associated with a partitioning of a sample space or a system

$$(A_1,p_1),\dots,(A_k,p_k),$$

where $A_1,\dots,A_k$ are mutually exclusive and totally exhaustive events (a partitioning of the sample space) and $p_1,\dots,p_k$ the associated probabilities, that is, $P(A_j)=p_j$, $j=1,\dots,k$ such that $p_j\geq 0$, $p_1+\cdots+p_k=1$, is the concept of *entropy* or *information* or *uncertainty*. This can be explained with a simple example for $k=2$.

Suppose that Mr Nimbus is contesting an election to be the chairman of the local township. Suppose that the only two possibilities are that either he wins or he does not win. Thus we have two events $A_1,A_2$ such that $A_1\cap A_2=\phi$, $A_1\cup A_2=S$ where $S$ is the sure event. Three local newspapers are predicting his chances of winning. The first newspaper gave a 50–50 chance of his winning, the second gave a 80–20 chance and the third gave a 60–40 chance. That is, if $A_1$ is the event of winning and $p=P(A_1)$ the true probability of winning, then the three estimates for this $p$ are $0.5, 0.8, 0.6$, respectively. We have three schemes here:

$$\text{Scheme 1: }\begin{pmatrix} A_1 & A_2\\ 0.5 & 0.5\end{pmatrix};$$

$$\text{Scheme 2: }\begin{pmatrix} A_1 & A_2\\ 0.8 & 0.2\end{pmatrix};$$

$$\text{Scheme 3: }\begin{pmatrix} A_1 & A_2\\ 0.6 & 0.4\end{pmatrix}.$$

In Scheme 1, there is quite a lot of uncertainty about the win, because it is a 50–50 situation with maximum uncertainty, a 50% chance of winning. In Scheme 2, the uncertainty is much less because it is a 80–20 situation. Whatever be that “uncertainty”, one can say this much that in Scheme 3 the uncertainty is in between the situations in Schemes 1 and 2. Lack of uncertainty is the “information” content in a scheme. If one can come up with a mathematical measure for this ‘information” content in a scheme, it has a lot of applications in practical situations such as sending a wireless message from one point and it is captured at another point. One would like to make sure that the message is fully captured in every respect or at least the information content is maximum. If a photo is transmitted, we would like to capture it with all of its full details.

Shannon in 1948 came up with a measure of “uncertainty” or “information” in a scheme. He developed it for communication networks. The measure is

$$S=-c\sum_{i=1}^{k}p_i\ln p_i$$

where $c$ is a constant and ln is the natural logarithm. He developed it by putting forward desirable properties as axioms, postulates or assumptions, and then deriving the expression mathematically. A whole discipline is developed and it is now known as “Information Theory” with a wide range of applications in almost all fields. The measure is simply called “entropy” in order to avoid possible misinterpretation if the term “information” or “uncertainty” is used. Such a procedure of developing a mathematical measure from basic postulates is called an axiomatic development. Axiomatic development of the basic concepts in Information Theory and Statistics may be seen from the book [14]. Shannon’s theory is extended also in various forms, some of which are given in the above mentioned book and some more applications may be seen from a recent paper by Mathai and Haubold [8].

## Exercises 2.5

**2.5.1.** An unbiased coin is tossed 10 times. What is the probability of getting (i) exactly 4 heads?; (ii) exactly 2 tails?; (iii) first head at the 10th trial?; (iv) the 3rd head at the 10th trial?

**2.5.2.** A class has 10 students. Each student has a birthday which can be one of the 365 days of the year, and no other information about the birthdays is available. (i): A student is selected at random. What is the probability that the student has the birthday on 1 February? (ii) What is the probability that their birthdays are all distinct, none coinciding with any other?

**2.5.3.** In a number lottery, each ticket has 3 digits. When the lottery is drawn, a specific sequence of 3 digits will win, the digits could be repeated also. A person has bought 4 tickets. What is the probability that one of his tickets is the winning ticket?

**2.5.4.** In Exercise 2.5.3, if repetition of the numbers is not allowed, then what is the answer?

**2.5.5.** From a well-shuffled deck of 52 playing cards, a hand of 8 is drawn at random. What is the probability that the hand contains 4 clubs, 2 spades, 1 heart and 1 diamond?

**2.5.6.** In a game, an unbiased coin is tossed successively. The game is finished when a head appears. What is the probability that (i) the game is over with less than or equal to 10 trials; (ii) the game is over at the 10th trial.

**2.5.7.** In the same game of tossing an unbiased coin successively, suppose that a person wins the game if a head appears. What is the probability of the person winning the game?

**2.5.8.** A balanced die is rolled twice. What is the probability of (i) rolling 6 (sum of the face numbers is 6)? (ii) getting an even number on both occasions? (iii) and even number comes in the first trial and odd number comes in the second trial?

**2.5.9.** In 6/36 lottery, there are 36 numbers and a given collection of 6 will win. A person has 3 such 6/36 tickets. What is the probability that one of these three is the winning ticket (assume that a person will not buy tickets with the same set of numbers on more than one ticket)?

**2.5.10.** In a 7/49 lottery, there are 49 numbers and a specific collection of 7 numbers wins. A person has 3 such tickets. (i) What is the probability that one of these is the winning ticket (assume that no two tickets will have the same set of numbers); (ii) Comparing with the probabilities in Exercises 2.5.9 and 2.5.10 (i), which lottery that a person should prefer 6/36 or 7/49?

**2.5.11.** A manufacturing unit of water heaters is known to produce 10% of defective items. A customer bought 3 water heaters from this manufacturer. What is the probability that (i) at least one of the three is defective; (ii) all are defective?

**2.5.12.** Vembanad Lake contains $n$ Karimeen (particular fish). A random sample of 50 Karimeen were caught and tagged and then released into the lake. After several months, a random sample of 100 Karimeen were caught. (i) What is the probability that this sample contains 5 tagged Karimeen? (ii) How will you estimate $n$, the total number of Karimeen in the Vembanad Lake based on this information that out of 100 caught 5 were found to be tagged?

**2.5.13.** A box contains 3 red and 5 green identical balls. Balls are taken at random, one by one, with replacement. What is the probability of getting (i) 3 red and 5 green in 8 trials; (ii) a red ball is obtained before a green ball is obtained.

**2.5.14.** In Exercise 2.5.13, if the balls are taken at random without replacement, what is the probability of getting (i) the sequence *RRGG* in four trials; (ii) *RGRR* in four trials? (ii) the third ball is green given that the first two were red? (iii) the third ball is green given that the first ball was red? (iv) the third ball is green and no other information is available.

**2.5.15.** Thekkady Wildlife Reserve is visited by people from Kerala, Tamilnadu, Karnataka and from other places. For any day, suppose that the proportions are 50%, 30%, 10%, 10%, respectively. Suppose that the probability that garbage will be thrown around at the reserve, on any day, by visitors from Kerala is 0.9, visitors from Tamilnadu is 0.9, visitors from Karnataka is 0.5 and for others it is 0.10. (i) What is the probability that the reserve will have garbage thrown around on any given day; (ii) On a particular day, it was found that the place had garbage thrown around, and what is the probability that it is done by Keralite visitors?

**2.5.16.** In a production process, two machines are producing the same item. Machine 1 is known to produce 5% defective (items which do not satisfy quality specifications) and Machine 2 is known to produce 2% defective. Sixty percent of the total production per day is by Machine 1 and 40% by Machine 2. An item from the day's production is taken at random and found to be defective. What is the probability that it was produced by Machine 1?

**2.5.17.** In a multiple choice examination, there are 10 questions and each question is supplied with 4 possible answers of which one is the correct answer to the question. A student, who does not know any of the correct answers, is answering the questions by picking answers at random. What is the probability that the student gets (i) exactly 8 correct answers; (ii) at least 8 correct answers; (iii) not more than three correct answers?

**2.5.18.** There are 4 envelopes addressed to 4 different people. There are 4 letters addressed to the same 4 people. A secretary puts the letters at random to the four envelopes and mails. All letters are delivered. What is the probability that none gets the letter addressed to him/her?

**2.5.19.** Construct two examples each of practical situations where you have two events $A$ and $B$ in the same sample space such that they are (i) mutually exclusive and independent; (ii) mutually exclusive and not independent; (iii) not mutually exclusive but independent; (iv) not mutually exclusive and not independent.

**2.5.20.** For the events $A_1, \dots, A_k$ in the same sample space $S$, show that
(i) $P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq p(A_1) + P(A_2) + \cdots + P(A_k)$;
(ii) $P(A_1 \cap A_2 \cap \cdots \cap A_k) \geq P(A_1) + \cdots + P(A_k) - (k-1)$.

**2.5.21.** For two events $A$ and $B$ in the same sample space $S$, show that if $A$ and $B$ are independent events (that is satisfying the product probability property) then (i) $A$ and $B^c$; (ii) $A^c$ and $B$; (iii) $A^c$ and $B^c$ are independent events, where $A^c$ and $B^c$ denote the complements of $A$ and $B$ in $S$.

# 3 Random variables

## 3.1 Introduction

Random variables constitute an extension of mathematical variables just like complex variables providing an extension to the real variable system. Random variables are mathematical variables with some probability measures attached to them. Before giving a formal definition to random variables, let us examine some random experiments and some variables associated with such random experiments. Let us take the simple experiment of an unbiased coin being tossed twice.

**Example 3.1.** Tossing an unbiased coin twice. The sample space is

$$S = \{(H,T),(T,H),(H,H),(T,T)\}.$$

There are four outcomes or four elementary events. Let $x$ be the number of heads in the elementary events or in the outcomes. Then $x$ can take the values $0,1,2$, and thus $x$ is a variable here. But we can attach a probability statement to the values taken by this variable $x$. The probability that $x$ takes the value zero is the probability of getting two tails and it is $\frac{1}{4}$. The probability that $x$ takes the value 1 is the probability of getting exactly one head, which is $\frac{1}{2}$. The probability that $x$ takes the value 2 is $\frac{1}{4}$. The probability that $x$ takes any another value, other than $0,1,2$, is zero because it is an impossible event in this random experiment. Thus the probability function, associated with this variable $x$, denoted by $f(x)$, can be written as follows:

$$f(x) = \begin{cases} 0.25, & \text{for } x = 0 \\ 0.50, & \text{for } x = 1 \\ 0.25, & \text{for } x = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

Here, $x$ takes individually distinct values with non-zero probabilities. That is, $x$ here takes the specific value zero with probability $\frac{1}{4}$, the value 1 with probability $\frac{1}{2}$ and the value 2 with probability $\frac{1}{4}$. Such random variables are called discrete random variables. We will give a formal definition after giving a definition for a random variable.

We can also compute the following probability in this case. What is the probability that $x \le a$ for all real values of $a$? Let us denote this probability by $F(a)$, that is,

$$F(a) = \Pr\{x \le a\} = \text{probability of the event } \{x \le a\}.$$

From Figure 3.1, it may be noted that when $a$ is anywhere from $-\infty$ to 0, not including zero, the probability is zero, and hence $F(a) = 0$. At $x = 0$, there is a probability $\frac{1}{4}$ and this remains the same for all values of $a$ from zero to 1 with zero included but 1 ex-

https://doi.org/10.1515/9783110562545-003

cluded, that is, $0 \le a < 1$. Remember that we are computing the sum of all probabilities up to and including point $x = a$, or we are computing the cumulative probabilities in the notation $\Pr\{x \le a\}$. There is a jump at $x = 1$ equal to $\frac{1}{2}$. Thus when $1 \le a < 2$, then all the probabilities cumulated up to $a$ is $0 + \frac{1}{4} + 0 + \frac{1}{2} + 0 = \frac{3}{4}$. When $a$ is anywhere $2 \le a < \infty$, all the probabilities cumulated up to $a$ will be $0 + \frac{1}{4} + 0 + \frac{1}{2} + 0 + \frac{1}{4} + 0 = 1$. Thus the cumulative probability function here, denoted by $F(a) = \Pr\{x \le a\}$, can be written as follows:

$$F(a) = \begin{cases} 0, & -\infty < a < 0 \\ 0.25, & 0 \le a < 1 \\ 0.75, & 1 \le a < 2 \\ 1, & 2 \le a < \infty. \end{cases}$$

Here, for this variable $x$, we can associate with $x$ a probability function $f(x)$ and a cumulative probability function $F(a) = \Pr\{x \le a\}$.

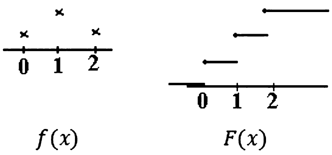


**Figure 3.1:** Left: Probability function $f(x)$; Right: Cumulative probability function $F(x)$.

**Notation 3.1.** $\Pr\{c \le x \le d\}$: probability of the event that $c \le x \le d$.

Now let us examine another variable defined over this same sample space. Let $y$ be the number of heads minus the number of tails in the outcomes. Then $y$ will take the value $-2$ for the sample point $(T, T)$ where the number of heads is zero and the number of tails is 2. The points $(H, T)$ and $(T, H)$ will give a value 0 for $y$ and $(H, H)$ gives a value 2 to $y$. If $f_y(y)$ denotes the probability function and $F_y(a) = \Pr\{y \le a\}$ the cumulative probability function, then we have the following, which may also be noted from Figure 3.2:

$$f_y(y) = \begin{cases} 0.25, & y = -2 \\ 0.5, & y = 0 \\ 0.25, & y = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

$$F_y(a) = \begin{cases} 0, & -\infty < a < -2 \\ 0.25, & -2 \le a < 0 \\ 0.75, & 0 \le a < 2 \\ 1, & 2 \le a < \infty. \end{cases}$$

Both $x$ and $y$ here are discrete variables in the sense of taking individually distinct values with non-zero probabilities. We may also note one more property that **on a given sample space any number of such variables can be defined**. The above ones, $x$ and $y$, are only two such variables.

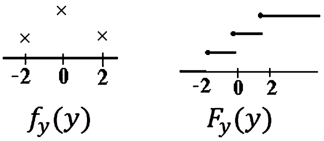


**Figure 3.2:** Left: Probability function of $y$; Right: Cumulative probability function of $y$.

Now, let us consider another example of a variable, which is not discrete. Let us examine the problem of a child playing with scissors and cutting a string of 10 cm into two pieces.

**Example 3.2** (Random cut of a string). Let one end of the string be denoted by 0 and the other end by 10 and let the distance from zero to the point of cut be $x$. Then, of course, $x$ is a variable because we did not know where exactly would be the cut on the string. What is the probability that the cut is anywhere in the interval $2 \le x \le 3.5$? In Chapter 1, we have seen that in a situation like this we assign probabilities proportional to the length of the intervals and then

$$\Pr\{2 \le x \le 3.5\} = \frac{3.5 - 2.0}{10} = \frac{1.5}{10} = 0.15.$$

What is the probability that the cut is between 2 and 2.001? This is given by

$$\Pr\{2 \le x \le 2.001\} = \frac{2.001 - 2.000}{10} = \frac{0.001}{10} = 0.0001.$$

What is the probability that the cut is exactly at 2?

$$\Pr\{x = 2\} = \frac{2 - 2}{10} = 0.$$

Here, $x$ is defined on a continuum of points and the probability that $x$ takes any specific value is zero because here the probabilities are assigned as relative lengths. A point has no length. Such variables, which are defined on continuum of points, will be called continuous random variables. We will give a formal definition after defining a random variable. A probability function which can be associated with this $x$, denoted by $f_x(x)$, will be of the following form:

$$f_x(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10 \\ 0, & \text{elsewhere.} \end{cases}$$

Let us see whether we can compute the cumulative probabilities here also. What is the probability that $x \le a$ for all real values of $a$? Let us denote this by $F_x(a)$. Then when

$-\infty < a < 0$, the cumulative probability is zero. When $0 \le a < 10$, it is $\frac{a}{10}$, probabilities being relative lengths, and when $10 \le a < \infty$ it is $\frac{10}{10} + 0 = 1$. Thus we have

$$F_x(a) = \begin{cases} 0, & -\infty < a < 0 \\ \frac{a}{10}, & 0 \le a < 10 \\ 1, & 10 \le a < \infty. \end{cases}$$

The probability function in the continuous case is usually called the density function. Some authors do not make a distinction; in both discrete and continuous cases, the probability functions are either called probability functions or density functions. We will use the term "probability function" in the discrete case and mixed cases and "density function" in the continuous case. The density and cumulative density, for the above example, are given in Figure 3.3.

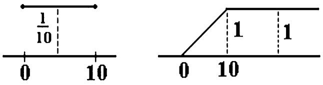


**Figure 3.3:** Left: Density function of $x$; Right: Cumulative density function of $x$.

Here, we may note some interesting properties. The cumulative probability function $F_x(a)$ could have been obtained from the density function by integration. That is,

$$F_x(a) = \int_{-\infty}^{a} f(t)\mathrm{d}t = 0 + \int_0^a \frac{1}{10}\mathrm{d}t = \left[\frac{t}{10}\right]_0^a = \frac{a}{10}.$$

Similarly, the density is available from the cumulative density function by differentiation since here the cumulative density function is differentiable. That is,

$$\left[\frac{\mathrm{d}}{\mathrm{d}a}F_x(a)\right]_{a=x} = \left[\frac{\mathrm{d}}{\mathrm{d}a}\frac{a}{10}\right]_{a=x} = \frac{1}{10} = f_x(x).$$

We have considered two discrete variables associated with the random experiment in Example 3.1 and one continuous random variable in Example 3.2. In all of the three cases, one could have computed the cumulative probabilities, or $\Pr\{x \le a\}$ was defined for all real $a$, $-\infty < a < \infty$. Such variables will be called random variables. Before giving a formal definition, a few more observations are in order. In the two discrete cases, we had the probability function, which were of the form:

$$f(x_*) = \Pr\{x = x_*\} \tag{3.1}$$

and the cumulative probability function was obtained by adding up the individual probabilities. That is,

$$F(a) = \Pr\{x \le a\} = \sum_{-\infty < x \le a} f(x). \tag{3.2}$$

In Example 3.2, we considered one continuous random variable $x$ where we had the density function

$$f_x(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10 \\ 0, & \text{elsewhere,} \end{cases}$$

and the cumulative density function

$$F_x(a) = \Pr\{x \le a\} = \begin{cases} 0, & -\infty < a < 0 \\ \frac{a}{10}, & 0 \le a < 10 \\ 1, & 10 \le a < \infty. \end{cases}$$

$$= \int_{-\infty}^{a} f_x(t)\mathrm{d}t. \tag{3.3}$$

**Definition 3.1** (Random variables)**.** Any variable $x$ defined on a sample space $S$ for which the cumulative probabilities $\Pr\{x \le a\}$ can be defined for all real values of $a$, $-\infty < a < \infty$, is called a real random variable $x$.

**Definition 3.2** (Discrete random variables)**.** Any random variable $x$ which takes individually distinct values with non-zero probabilities is called a discrete random variable and in this case the probability function, denoted by $f_x(x)$, is given by

$$f_x(x_*) = \Pr\{x = x_*\}$$

and obtained by taking successive differences in (3.2).

**Definition 3.3** (Continuous random variables)**.** Any random variable $x$, which is defined on a continuum of points, where the probability that $x$ takes a specific value $x_*$ is zero, is called a continuous random variable and the density function is available from the cumulative density by differentiation, when differentiable, or the cumulative density is available by integration of the density. That is,

$$f_x(x) = \left[\frac{\mathrm{d}}{\mathrm{d}a} F_x(a)\right]_{a=x} \tag{3.4}$$

or

$$F_x(a) = \int_{-\infty}^{a} f_x(t)\mathrm{d}t. \tag{3.5}$$

**Definition 3.4** (Distribution function)**.** The cumulative probability/density function of a random variable $x$ is also called the distribution function associated with that random variable $x$, and it is denoted by $F(x)$:

$$F(a) = [\Pr\{x \le a\}, -\infty < a < \infty]. \tag{3.6}$$

We can also define probability/density function and cumulative function, free of random experiments, by using a few axioms.

## 3.2 Axioms for probability/density function and distribution functions

**Definition 3.5** (Density/Probability function). Any function $f(x)$ satisfying the following two axioms is called the probability/density function of a real random variable $x$:

(i) $f(x) \geq 0$ for all real $x$, $-\infty < x < \infty$;

(ii) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ if $x$ is continuous; and $\sum_{-\infty<x<\infty} f(x) = 1$ if $x$ is discrete.

**Example 3.3.** Check whether the following can be probability functions for discrete random variables:

$$f_1(x) = \begin{cases} 2/3, & x = -2 \\ 1/3, & x = 5 \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_2(x) = \begin{cases} 3/4, & x = -3 \\ 2/4, & x = 0, \\ -1/4, & x = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_3(x) = \begin{cases} 3/5, & x = 0 \\ 3/5, & x = 1 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 3.3.** Consider $f_1(x)$. Here, $f_1(x)$ takes the non-zero values $\frac{2}{3}$ and $\frac{1}{3}$ at the points $x = -2$ and $x = 5$, respectively, and $x$ takes all other values with zero probabilities. Condition (i) is satisfied, $f(x) \geq 0$ for all values of $x$. Condition (ii) is also satisfied because $\frac{2}{3} + \frac{1}{3} + 0 = 1$. Hence $f_1(x)$ here can represent a probability function for a discrete random variable $x$. We could have also stated $f_1(x)$ as follows:

$$f_1(-2) = \frac{2}{3}; \quad f_1(5) = \frac{1}{3}; \quad f(x) = 0 \quad \text{elsewhere}$$

where, for example, $f_1(-2)$ means $f_1(x)$ at $x = -2$.

$f_2(x)$ is such that $\sum_x f_2(x) = 1$, and thus the second condition is satisfied. But $f_2(x)$ at $x = 2$ or $f_2(2) = -\frac{1}{4}$ which is negative, and hence condition (i) is violated. Hence $f_2(x)$ here cannot be the probability function of any random variable.

$f_3(x)$ is non-negative for all values of $x$ because $f_3(x)$ takes the values $0, \frac{3}{5}, \frac{3}{5}$ but

$$\sum_x f_3(x) = 0 + \frac{3}{5} + \frac{3}{5} = \frac{6}{5} > 1.$$

Here, condition (ii) is violated, and hence $f_3(x)$ cannot be the probability function of any random variable.

**Example 3.4.** Check whether the following can be density functions of some random variables:

$$f_1(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, b > a \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_2(x) = \begin{cases} cx^4, & 0 < x < 1 \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_3(x) = \begin{cases} \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, & 0 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_4(x) = \begin{cases} x, & 0 \le x < 1 \\ 2-x, & 1 \le x \le 2 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 3.4.** $f_1(x)$ is non-negative since it is either 0 or $\frac{1}{b-a}$ where $b-a > 0$. Hence condition (i) is satisfied. Now, check the second condition:

$$\int_{-\infty}^{\infty} f_1(x)\mathrm{d}x = 0 + \int_a^b \frac{1}{b-a}\mathrm{d}x = \left[\frac{x}{b-a}\right]_a^b = \frac{b-a}{b-a} = 1.$$

Hence the second condition is also satisfied. It is a density function of a continuous random variable. The graph is given in Figure 3.4.

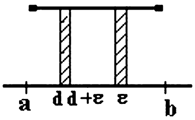


**Figure 3.4:** Uniform or rectangular density.

This density looks like a rectangle, and hence it is called a *rectangular density*. Since the probabilities are available as integrals or areas under the curve if we take any interval of length $\epsilon$ (epsilon) units, say from $d$ to $d+\epsilon$, then the probability that $x$ falls in the interval $d$ to $d+\epsilon$ or $d \le x \le d+\epsilon$ is given by the integral:

$$\int_d^{d+\epsilon} \frac{1}{b-a}\mathrm{d}x = \frac{\epsilon}{b-a}.$$

Since it is a rectangle, if we take an interval of length $\epsilon$ anywhere in the interval $a \le x \le b$, then the area will be the same as $\frac{\epsilon}{b-a}$ or we can say that the total area 1 is uniformly distributed over the interval $[a,b]$. In this sense, this density $f_1(x)$ is also called *uniform density*. Also we may observe here that these unknown quantities $a$ and $b$ could be any constants, free of $x$. As long as $b > a$, $f_1(x)$ is a density.

$f_2(x) \geq 0$ for all values of $x$ if $c > 0$ since either it is zero or $x^4$ in the interval $[0,1]$ which is positive. Thus condition (i) is satisfied if $c > 0$. Now, let us check condition (ii):

$$\int_{-\infty}^{\infty} f_2(x)\mathrm{d}x = 0 + \int_0^1 cx^4\mathrm{d}x$$
$$= \left[c\frac{x^5}{5}\right]_0^1 = \frac{c}{5}.$$

Hence condition (ii) is satisfied if $c = 5$. For $c = 5$, $f_2(x)$ is a density function.

$f_3(x)$ satisfies condition (i) when $\theta$ (theta) is positive because an exponential function can never be negative. Hence $f_3(x)$ takes zero or a positive value only. Now let us check the second condition:

$$\int_{-\infty}^{\infty} \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}\mathrm{d}x = 0 + \int_0^{\infty} \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}\mathrm{d}x = [-\mathrm{e}^{-\frac{x}{\theta}}]_0^{\infty} = 1.$$

Hence it is a density. Note that whatever be the value of $\theta$ as long as it is positive, $f_3(x)$ is a density, see Figure 3.5.

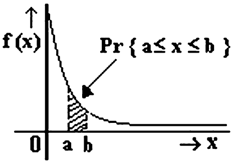


**Figure 3.5:** Exponential or negative exponential density.

Since this density is associated with an exponential function it is called an *exponential density*. Note that if $\theta$ is negative, then $\frac{1}{\theta} < 0$ even though the exponential function remains positive. Thus condition (i) will be violated. If $\theta$ is negative, then the exponent $-\frac{x}{\theta} > 0$ thereby the integral from 0 to $\infty$ will be $\infty$. Thus condition (ii) will also be violated. For $\theta \leq 0$ here, $f_3(x)$ cannot be a density. When integration is from 0 to $\infty$, the exponential function with a positive exponent cannot create a density we need not say “negative exponential density” and we simply say that it is an exponential density, and it is implied that the exponent is negative.

$f_4(x)$ is zero or $x$ in $[0,1)$ and $2 - x$ in $[1,2]$, and hence $f_4(x) \geq 0$ for all $x$ and condition (i) is satisfied. The total integral is available from the integrals over the several intervals:

$$\int_{-\infty}^{\infty} f_4(x)\mathrm{d}x = 0 + \int_0^1 x\mathrm{d}x + \int_1^2 (2-x)\mathrm{d}x + 0$$
$$= \left[\frac{x^2}{2}\right]_0^1 + \left[2x - \frac{x^2}{2}\right]_1^2 = \frac{1}{2} + \frac{1}{2} = 1.$$

Thus, condition (ii) is also satisfied and $f_4(x)$ here is a density.

The graph of this density looks like a triangle, and hence it is called a *triangular density* as shown in Figure 3.6.

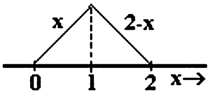


**Figure 3.6:** Triangular density.

**Definition 3.6** (Parameters). Arbitrary constants sitting in a density or probability function are called parameters.

In $f_1(x)$ of Example 3.4, there are two unknown quantities $a$ and $b$. Irrespective of the values of $a$ and $b$, as long as $b > a$ then we found that $f_1(x)$ was a density. Hence there are two parameters in that density. In $f_3(x)$ of Example 3.4, we had one unknown quantity $\theta$. As long as $\theta$ was positive, $f_3(x)$ remained as a density. Hence there is one parameter here in this density, and that is $\theta > 0$.

**Definition 3.7** (Normalizing constant). If a constant sitting in a function is such that for a specific value of this constant the function becomes a density or probability function then that constant is called the normalizing constant.

In $f_2(x)$ of Example 3.4, there was a constant $c$ but for $c = 5$, $f_2(x)$ became a density. This $c$ is the normalizing constant there.

**Definition 3.8** (Degenerate random variable). If the whole probability mass is concentrated at one point, then the random variable is called a degenerate random variable or a mathematical variable. Consider the following density/probability function:

$$f(x) = \begin{cases} 1, & x = b \\ 0, & \text{elsewhere.} \end{cases}$$

Here, at $x = b$ the whole probability mass 1 is there and everywhere else the function is zero. The random variable here is called a *degenerate random variable* or with probability 1 the variable $x$ takes the value $b$ or it is a mathematical variable. If there are two points such that at $x = c$ we have probability 0.9999 and at $x = d \neq c$ we have probability 0.0001, then it is not a degenerate random variable even though most of the probability is at one point $x = c$.

Thus, statistics or statistical science is a systematic study of random phenomena and random variables, extending the study of mathematical variables, and as such mathematical variables become special cases of random variables or as degenerate random variables. This author had coined the name “Statistical Science” when he launched the Statistical Science Association of Canada, which became the present Statistical Society of Canada. Thus in this author’s definition, statistical sciences has a wider coverage compared to mathematical sciences. But nowadays the term mathematical sciences is used to cover all aspects of mathematics and statistics.

**Example 3.5.** Compute the distribution function for the following probability functions:

$$f_1(x) = \begin{cases} 0.3, & x = -2 \\ 0.2, & x = 0, \\ 0.5, & x = 3 \\ 0, & \text{otherwise}; \end{cases}$$

$$f_2(x) = \begin{cases} c(\frac{1}{2})^x, & x = 0, 1, \\ 0, & \text{otherwise}. \end{cases}$$

**Solution 3.5.** The distribution function in the discrete case is

$$F(a) = \Pr\{x \le a\} = \sum_{-\infty < x \le a} f(x).$$

Hence for $f_1(x)$, it is zero for $-\infty < x < -2$, then there is a jump of 0.3 at $x = -2$, and so on. Therefore,

$$F(a) = \begin{cases} 0, & -\infty < a < -2 \\ 0.3, & -2 \le a < 0 \\ 0.5\,(= 0.3 + 0.2), & 0 \le a < 3 \\ 1, & 3 \le a < \infty. \end{cases}$$

It is a *step function*. In general, for a discrete case we get a step function as the distribution function.

For $f_2(x)$, the normalizing constant $c$ is to be determined to make it a probability function. If it is a probability function, then the total probability is

$$0 + \sum_{x=0}^{2} \left(c\frac{1}{2}\right)^x = 0 + c\left(1 + \frac{1}{2} + \frac{1}{4}\right) = c\frac{7}{4}.$$

Hence for $c = \frac{4}{7}$, $f_2(x)$ is a probability function and it is given by

$$f_2(x) = \begin{cases} 4/7, & x = 0 \\ 2/7, & x = 1 \\ 1/7, & x = 2 \\ 0, & \text{otherwise}. \end{cases}$$

Hence the distribution function is given by

$$F(x) = \begin{cases} 0, & -\infty < x < 0 \\ 4/7, & 0 \le x < 1 \\ 6/7, & 1 \le x < 2 \\ 1, & 2 \le x < \infty. \end{cases}$$

Again, note that it is a step function. The student may draw the graphs for the distribution function for these two cases.

**Example 3.6.** Evaluate the distribution function for the following densities:

$$f_1(x) = \begin{cases} \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, & 0 \le x < \infty \\ 0, & \text{otherwise;} \end{cases}$$

$$f_2(x) = \begin{cases} x, & 0 < x < 1 \\ 2 - x, & 1 \le x < 2 \\ 0, & \text{otherwise.} \end{cases}$$

**Solution 3.6.** The distribution function, by definition, in the continuous case is

$$F(t) = \int_{-\infty}^{t} f(x)\mathrm{d}x.$$

Hence in $f_1(x)$,

$$\begin{aligned} \int_{-\infty}^{t} f_1(x)\mathrm{d}x &= 0 + \int_0^t \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}\mathrm{d}x \\ &= [-\mathrm{e}^{-\frac{x}{\theta}}]_0^t = 1 - \mathrm{e}^{-\frac{t}{\theta}}, \quad 0 \le t < \infty, \end{aligned}$$

and zero from $-\infty < x < 0$. For $f_2(x)$, one has to integrate in different pieces. Evidently, $F(t) = 0$ for $-\infty < t < 0$. When $t$ is in the interval 0 to 1, the function is $x$ and its integral is $\frac{x^2}{2}$. Therefore,

$$\left[\frac{x^2}{2}\right]_0^t = \frac{t^2}{2}.$$

When $t$ is in the interval 1 to 2 the integral up to 1, available from $\frac{t^2}{2}$ at $t = 1$ which is $\frac{1}{2}$, plus the integral of the function $(2 - x)$ from 1 to $t$ is to be computed. That is,

$$\frac{1}{2} + \int_1^t (2 - x)\mathrm{d}x = \frac{1}{2} + \left[2x - \frac{x^2}{2}\right]_1^t = -1 + 2t - \frac{t^2}{2}.$$

When $t$ is above 2, the total integral is one. Hence we have

$$F(t) = \begin{cases} 0, & -\infty < t < 0 \\ \frac{t^2}{2}, & 0 \le t < 1 \\ -1 + 2t - \frac{t^2}{2}, & 1 \le t < 2 \\ 1, & t \ge 2. \end{cases}$$

The student is asked to draw the graphs of the distribution function in these two density functions.

### 3.2.1 Axioms for a distribution function

If we have a discrete or continuous random variable, the distribution function is $F(t) = \Pr\{x \le t\}$. Without reference to a random variable $x$, one can define $F(t)$ by using the following axioms:

(i) $F(-\infty) = 0$;
(ii) $F(\infty) = 1$;
(iii) $F(a) \le F(b)$ for all $a < b$;
(iv) $F(t)$ is right continuous.

Thus $F(t)$ is a monotonically non-decreasing (either it increases steadily or it remains steady for some time) function from zero to 1 when $t$ varies from $-\infty$ to $\infty$. The student may verify that conditions (i) to (iv) above are satisfied by all the distribution functions that we considered so far.

### 3.2.2 Mixed cases

Sometime we may have a random variable where part of the probability mass is distributed on some individually distinct points (discrete case) but the remaining probability is distributed over a continuum of points (continuous case). Such random variables are called *mixed cases*. We will list one example here, from where it will be clear how to handle such cases.

**Example 3.7.** Compute the distribution function for the following probability function for a mixed case:

$$f(x) = \begin{cases} \frac{1}{2}, & x = -2 \\ x, & 0 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

**Solution 3.7.** The definition for the distribution function remains the same whether the variable is discrete, continuous or mixed:

$$F(t) = \Pr\{x \le t\}.$$

For $-\infty < t < -2$, obviously $F(t) = 0$. There is a jump of $\frac{1}{2}$ at $t = -2$ and then it remains the same until 1. In the interval $[0, 1]$, the function is $x$ and its integral is

$$\int_0^t x\mathrm{d}x = \left[\frac{x^2}{2}\right]_0^t = \frac{t^2}{2}.$$

For $t$ greater than 1, the total probability 1 is attained. Therefore, we have

$$F(t)=\begin{cases}0, & -\infty<t<-2\\ \frac{1}{2}, & -2\le t<0\\ \frac{1}{2}+\frac{t^2}{2}, & 0\le t<1\\ 1, & t\ge 1.\end{cases}$$

The graph will look like that in Figure 3.7.

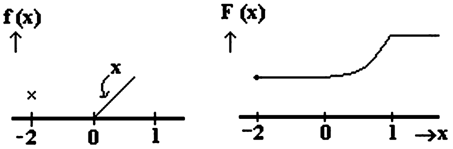


**Figure 3.7:** The distribution function for a mixed case.

Note that for $t$ up to 0 it is a step function then the remaining part is a continuous curve until 1 and then it remains steady at the final value 1.

**Example 3.8.** Compute the probabilities (i) $\Pr\{-2\le x\le 1\}$, (ii) $\Pr\{0\le x\le 1.7\}$ for the probability function

$$f(x)=\begin{cases}0.2, & x=-1,\\ 0.3, & x=0,\\ 0.3, & x=1.5,\\ 0.2, & x=2,\\ 0, & \text{otherwise}.\end{cases}$$

**Solution 3.8.** In the discrete case, the probabilities are added up from those at individual points. When $-2\le x\le 1$, the probabilities in this interval are 0, 0.2 at $x=-1$ and 0.3 at $x=0$. Therefore, the answer to (i) is $0+0.2+0.3=0.5$. When $0\le x\le 1.7$, the probabilities are 0, 0.3 at $x=0$ and 0.3 at $x=1.5$. Hence the answer to (ii) is $0+0.3+0.3=0.6$.

In the discrete case, the probability that $x$ falls in a certain interval is the sum of the probabilities from the corresponding distinct points with non-zero probabilities falling in that interval.

**Example 3.9.** Compute the following probabilities on the waiting time $t$, (i) $\Pr\{0\le t\le 2\}$, (ii) $\Pr\{3\le t\le 10\}$ if the waiting time has an exponential density with the parameter $\theta=5$.

**Solution 3.9.** The waiting time having an exponential density with parameter $\theta=5$ means that the density of $t$ is given by

$$f(t)=\begin{cases}\frac{1}{5}e^{-\frac{t}{5}}, & 0\le t<\infty\\ 0, & \text{elsewhere}.\end{cases}$$

Probabilities are the areas under the density curve between the corresponding ordinates or the integral of the density over the given interval. Hence for (i) the probability is given by

$$\int_0^2 \frac{1}{5}\mathrm{e}^{-\frac{t}{5}}\mathrm{d}t = [-\mathrm{e}^{-\frac{t}{5}}]_0^2 = 1 - \mathrm{e}^{-\frac{2}{5}}.$$

In a similar manner, the probability for (ii) is given by

$$\int_3^{10} f(t)\mathrm{d}t = [-\mathrm{e}^{-\frac{t}{5}}]_3^{10} = \mathrm{e}^{-\frac{3}{5}} - \mathrm{e}^{-\frac{10}{5}}.$$

The following shaded areas in Figure 3.8 are the probabilities.

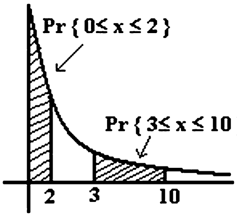


**Figure 3.8:** Probabilities in the exponential density.

In a continuous case, the probability of the variable $x$ falling in a certain interval $[a, b]$ is the area under the density curve over the interval $[a, b]$ or between the ordinates at $x = a$ and $x = b$.

## Exercises 3.2

**3.2.1.** Check whether the following are probability functions for some discrete random variables:

$$f_1(x) = \begin{cases} \frac{1}{2}, & x = -1 \\ \frac{1}{2}, & x = 1 \\ 0, & \text{elsewhere;} \end{cases} \qquad f_2(x) = \begin{cases} 2, & x = \frac{2}{3} \\ 1, & x = \frac{1}{3} \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_3(x) = \begin{cases} 1.2, & x = 0 \\ -0.2, & x = 1 \\ 0, & \text{elsewhere;} \end{cases} \qquad f_4(x) = \begin{cases} 0.8, & x = 1 \\ 0.3, & x = 2 \\ 0, & \text{otherwise.} \end{cases}$$

**3.2.2.** Check whether the following are density functions for some continuous random variables:

$$f_1(x) = \begin{cases} c(x^2 + 3x + 1), & 0 \le x \le 2 \\ 0, & \text{otherwise;} \end{cases}$$

$$f_2(x) = \begin{cases} \frac{c}{x^2}, & 1 \le x < \infty \\ 0, & \text{otherwise}; \end{cases}$$

$$f_3(x) = ce^{-\beta|x|}, \quad -\infty < x < \infty;$$

$$f_4(x) = \begin{cases} cx^2, & 0 < x < 2 \\ 6 - x, & 2 \le x \le 6 \\ 0, & \text{otherwise.} \end{cases}$$

**3.2.3.** An unbiased coin is tossed several times. If $x$ denotes the number of heads in the outcomes, construct the probability function of $x$ when the coin is tossed (i) once; (ii) two times; (iii) five times.

**3.2.4.** In a multiple choice examination, there are 8 questions and each question is supplied with 3 possible answers of which one is the correct answer to the question. A student, who does not know any of the correct answers, is answering the questions by picking the answers at random. Let $x$ be the number of correct answers. Construct the probability function of $x$.

**3.2.5.** In Exercise 3.2.4, let $x$ be the number of trials (answering the questions) at which the first correct answer is obtained, such as the third ($x = 3$) question answered is the first correct answer. Construct the probability function of $x$.

**3.2.6.** In Exercise 3.2.4, let the $x$-th trial resulted in the 3rd correct answer. Construct the probability function of $x$.

**3.2.7.** Compute the distribution function for each probability function in Exercise 3.2.1 and draw the corresponding graphs.

**3.2.8.** Compute the distribution function for each probability function in Exercise 3.2.2 and draw the corresponding graphs also.

**3.2.9.** Compute the distribution functions and draw the graphs in Exercises 3.2.3–3.2.6.

**3.2.10.** For the following mixed case, compute the distribution function:

$$f(x) = \begin{cases} \frac{1}{4}, & x = -5 \\ x, & 0 < x < 1, \\ \frac{1}{4}, & x = 5 \\ 0, & \text{otherwise.} \end{cases}$$

**3.2.11.** In Exercise 3.2.2, compute the following probabilities: (i) $\Pr\{1 \le x \le 1.5\}$ for $f_1(x)$; (ii) $\Pr\{2 \le x \le 5\}$ for $f_2(x)$; (iii) $\Pr\{-2 \le x \le 2\}$ for $f_3(x)$; (iv) $\Pr\{1.5 \le x \le 3\}$ for $f_4(x)$.

**3.2.12.** In Exercises 3.2.4 and 3.2.5, compute the probability for $2 \le x \le 5$, and in Exercise 3.2.6 compute the probability for $4 \le x \le 7$.

**Note 3.1.** For a full discussion of statistical densities and probability functions in common use, we need some standard series such as binomial series, logarithmic series, exponential series, etc. We will mention these briefly here. Those who are familiar with these may skip this section and go directly to the next chapter.

## 3.3 Some commonly used series

The following power series can be obtained by using the following procedure when the function is differentiable. Let $f(x)$ be differentiable countably infinite number of times and let it admit a power series expansion

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$$

then the coefficient

$$a_n = \frac{[\frac{\mathrm{d}^n}{\mathrm{d}x^n}f(x)|_{x=0}]}{n!}$$

or the series is

$$f(x) = f(0) + \frac{f^{(1)}(0)}{1!}x + \frac{f^{(2)}(0)}{2!}x^2 + \cdots \tag{3.7}$$

where $f^{(r)}(0)$ means to differentiate $f(x)$, $r$ times and then evaluate at $x = 0$. All of the following series are derived by using the same procedure.

### 3.3.1 Exponential series

$$\mathrm{e}^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^r}{r!} + \cdots \quad \text{for all } x. \tag{3.8}$$

$$\mathrm{e}^{-x} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \cdots + (-1)^r\frac{x^r}{r!} + \cdots \quad \text{for all } x. \tag{3.9}$$

### 3.3.2 Logarithmic series

Logarithm to the base e is called the natural logarithms and it is denoted by ln.

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \quad \text{for } |x| < 1. \tag{3.10}$$

For the convergence of the series, we need the condition $|x| < 1$:

$$\ln(1-x) = -\left[x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right], \quad |x| < 1. \tag{3.11}$$

### 3.3.3 Binomial series

The students are familiar with the binomial expansions for positive integer values, which can also be obtained by direct repeated multiplications, and the general result can be established by the method of induction:

$$
\begin{aligned}
&(1+x)^2 = 1 + 2x + x^2; \quad (a+b)^2 = a^2 + 2ab + b^2;\\
&(1+x)^3 = 1 + 3x + 3x^2 + x^3; \quad (a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3;\\
&(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \cdots + \binom{n}{n}x^n, \quad n = 1, 2, \ldots; \qquad (3.12)\\
&(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b + \cdots + \binom{n}{n}a^0 b^n,\\
&\qquad n = 1, 2, \ldots.
\end{aligned}
$$

What happens if the exponent is not a positive integer, if the exponent is something like $\frac{1}{2}, -20, -\frac{3}{2}$ or some general rational number $\alpha$ (alpha)? We can derive an expansion by using (3.7). Various forms of these are given below:

$$(1-x)^{-\alpha} = 1 + \frac{(\alpha)_1}{1!}x + \frac{(\alpha)_2}{2!}x^2 + \cdots + \frac{(\alpha)_r}{r!}x^r + \cdots, \quad |x| < 1. \qquad (3.13)$$

If $\alpha$ is not a negative integer, then we need the condition $|x| < 1$ for the convergence of the series. The Pochhammer symbol is

$$(\alpha)_r = \alpha(\alpha+1)\cdots(\alpha+r-1), \quad \alpha \neq 0, \ (\alpha)_0 = 1. \qquad (3.14)$$

Various forms of (3.13) can be obtained by replacing $x$ by $-x$ and $\alpha$ by $-\alpha$. For the sake of completeness, these will be listed here for ready reference:

$$(1+x)^{-\alpha} = [1-(-x)]^{-\alpha} = 1 - \frac{(\alpha)_1}{1!}x + \frac{(\alpha)_2}{2!}x^2 - \cdots, \quad |x| < 1. \qquad (3.15)$$

$$
\begin{aligned}
(1-x)^{\alpha} = (1-x)^{-(-\alpha)} &= 1 + \frac{(-\alpha)_1}{1!}x + \frac{(-\alpha)_2}{2!}x^2 + \cdots,\\
&\text{for } |x| < 1. \qquad (3.16)
\end{aligned}
$$

$$(1+x)^{\alpha} = [1-(-x)]^{-(-\alpha)} = 1 - \frac{(-\alpha)_1}{1!}x + \frac{(-\alpha)_2}{2!}x^2 - \cdots, \qquad (3.17)$$

for $|x| < 1$. In all cases, the condition $|x| < 1$ is needed for the convergence of the series except in the case when the exponent is a positive integer. When the exponent is $\alpha > 0$, then the coefficient of $\frac{x^r}{r!}$ is $(-\alpha)_r$. If $\alpha$ is a positive integer, then this Pochhammer symbol will be zero for some $r$ and the series will terminate into a polynomial, and hence the question of convergence does not arise. We have used the form $(1 \pm x)^{\pm\alpha}$. This is general enough because if we have a form

$$(a \pm b)^{\pm\alpha} = a^{\pm\alpha}\left(1 \pm \frac{b}{a}\right)^{\pm\alpha}$$

and thus we can convert to the form $(1 \pm x)$ by taking out $a$ or $b$ to make the resulting series convergent.

### 3.3.4 Trigonometric series

$$\begin{aligned}\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \\ \mathrm{e}^{ix} &= \cos x + i \sin x, \quad i = \sqrt{-1}.\end{aligned}$$

### 3.3.5 A note on logarithms

The mathematical statement

$$a^x = b$$

can be stated as the exponent $x$ is the logarithm of $b$ to the base $a$. For example, $2^3 = 8$ can be written as 3 (the exponent) is the logarithm of 8 to the base 2. The definition is restricted to $b$ being strictly a positive quantity when real or logarithm of negative quantities or zero is not defined in the real case. The standard notations used are the following:

$\log b \equiv \log_{10} b$ or common logarithm or logarithm to the base 10. When we say "$\log y$", it is a logarithm of $y$ to be base 10.

$\ln b \equiv \log_{\mathrm{e}} b$ or natural logarithm or logarithm to the base e. When we say "$\ln y$", it is a logarithm of $y$ to be base e.

For all other bases, other than 10 or e, write the base and write it as $\log_a b$. This note is given here because the students usually do not know the distinction between the notations "log" and "ln". For example,

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x}, \quad \frac{\mathrm{d}}{\mathrm{d}x} \log x = \frac{1}{x} \log_{10} \mathrm{e} \neq \frac{1}{x}.$$

**Note 3.2.** In Section 3.2.1, we have given an axiomatic definition of a distribution function and we defined a random variable with the help of the distribution function. Let us denote the distribution function associated with a random variable $x$ by $F(x)$. If $F(x)$ is differentiable at an arbitrary point $x$, then let us denote the derivative by $f(x)$. That is, $\frac{\mathrm{d}}{\mathrm{d}x} F(x) = f(x)$, which will also indicate that

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t.$$

In this situation, we call $F(x)$ an *absolutely continuous* distribution function. Absolute continuity is related to more general measures and integrals known as Lebesgue integrals. For the time being, if you come across the phrase "absolutely continuous distribution function", then assume that $F(x)$ is differentiable and its derivative is the density $f(x)$.

**Note 3.3.** Suppose that a density function $f(x)$ has non-zero part over the interval $[a,b]$ and zero outside this interval. When $x$ is continuous, then the probability that $x = a$, that is, $\Pr\{x = a\} = 0$ and $\Pr\{x = b\} = 0$. Then the students have the confusion whether $f(x)$ should be written as non-zero in $a \le x \le b$ or $a < x \le b$ or $a \le x < b$ or $a < x < b$. Should we include the boundary points $x = a$ and $x = b$ with the non-zero part of the density or with the zero part? For example, if we write an exponential density:

$$f(x) = \begin{cases} \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, & \theta > 0,\ 0 \le x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

should we write $0 < x < \infty$ or $0 \le x < \infty$. Note that if we are computing only probabilities then it will not make any difference. But if we are looking for a mode, then the function has a mode at $x = 0$ and if $x = 0$ is not included in the non-zero part of the density, then naturally we cannot evaluate the mode. For estimation of the parameters also, we may have similar problems. For example, if we consider a uniform density

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{elsewhere} \end{cases}$$

then what is known as maximum likelihood estimates [discussed in Module 7] for the parameters $a$ and $b$ do not exist if the end points are not included. That is, if the non-zero part of the density is written as $a < x < b$, then the maximum likelihood estimates for $a$ and $b$ do not exist. Hence when writing the non-zero part of the density include the end points of the interval where the function is non-zero.

**Note 3.4.** Note that when a random variables $x$ is continuous, then the following probability statements are equivalent:

$$\begin{aligned} \Pr\{a < x < b\} &= \Pr\{a \le x < b\} = \Pr\{a < x \le b\} = \Pr\{a \le x \le b\} \\ &= F(b) - F(a) \end{aligned}$$

where $F(x)$ is the distribution function. Also when $F(x)$ is absolutely continuous

$$F(b) - F(a) = \int_a^b f(t)\mathrm{d}t \quad \text{or} \quad \frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$$

where $f(x)$ is the density function.

## Exercises 3.3

**3.3.1.** By using a binomial expansion show that, for $n = 1, 2, \ldots$

$$2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}$$
$$0 = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} + \cdots \pm \binom{n}{n}$$

**3.3.2.** By using the identity,

$$(1+x)^m(1+x)^n \equiv (1+x)^{m+n}$$

and comparing the coefficient of $x^r$ on both sides show that

$$\sum_{s=0}^{r} \binom{m}{s}\binom{n}{r-s} = \binom{m+n}{r}, \quad m, n = 1, 2, \ldots.$$

**3.3.3.** By using the identity,

$$(1+x)^{n_1}(1+x)^{n_2}\cdots(1+x)^{n_k} \equiv (1+x)^{n_1+\cdots+n_k}$$

and comparing the coefficient of $x^r$ on both sides show that

$$\sum_{r_1}\cdots\sum_{r_k}\binom{n_1}{r_1}\binom{n_2}{r_2}\cdots\binom{n_k}{r_k} = \binom{n}{r}$$

where $r = r_1 + \cdots + r_k$, $n = n_1 + \cdots + n_k$, $n_j = 1, 2, \ldots, j = 1, \ldots, k$.

**3.3.4.** Show that

$$\sum_{m=1}^{n} m = \frac{n(n+1)}{2}; \quad \sum_{m=1}^{n} m^2 = \frac{n(n+1)(2n+1)}{6};$$
$$\sum_{m=1}^{n} m^3 = \left[\frac{n(n+1)}{2}\right]^2.$$

**3.3.5.** Compute the sums $\sum_{m=1}^{n} m^4$; $\sum_{m=1}^{n} m^5$; $\sum_{m=1}^{n} m^p$, $p = 6, 7, \ldots$.

**3.3.6.** Show that

$$a + ar + ar^2 + \cdots + ar^{n-1} = a\frac{(1-r^n)}{1-r}, \quad r \neq 1;$$
$$a + ar + ar^2 + \cdots = a\sum_{n=0}^{\infty} r^n = \frac{a}{1-r}, \quad \text{for } |r| < 1.$$

**3.3.7.** What is the infinite sum in Exercise 3.3.6 for (i) $r = 1$; (ii) $r = -1$; (iii) $r > 1$; (iv) $r < -1$.

**3.3.8.** Evaluate the sum $\sum_{x=k}^{\infty}\binom{x-1}{k-1}p^k q^{x-k}$, $q = 1 - p$, $0 < p < 1$.

**3.3.9.** Evaluate the sum $\sum_{x=0}^{n}\binom{n}{x}p^x \mathrm{e}^{tx} q^{n-x}$, $q = 1 - p$, $0 < p < 1$.

**3.3.10.** Compute the sum $\sum_{x=k}^{\infty}\binom{x-1}{k-1}p^k \mathrm{e}^{tx} q^{x-k}$, $q = 1 - p$, $0 < p < 1$.

# 4 Expected values

## 4.1 Introduction

Here, we will start with some commonly used probability functions and density functions and then we will define the concepts called expected values, moments, moment generating functions, etc. In Chapter 3, we have defined a probability function, a density function and distribution function or cumulative probability/density function. There is another term used in statistics and probability literature called "distributions" something like "exponential distribution", "normal distribution", etc. There is a possibility of confusion between a distribution and a distribution function. A distribution function is the cumulative probability/density function as defined in Chapter 3, whereas when we say that, for example, we have an exponential distribution or a variable $x$ is exponentially distributed, we mean that we have identified a random variable, a density function or a distribution function, and it is the random variable having exponential density. When we say we have a uniform distribution, it means that we have identified a random variable having a uniform or rectangular density or we have identified a random variable that is uniformly distributed. It is unfortunate that two technical terms, "distribution" and "distribution function", which are very similar, are used in statistical literature. Hopefully, the students will get accustomed to the technical terms fast and will not be confused.

We will introduce the concept of expected values first and then we will deal with commonly appearing probability/density functions or commonly appearing statistical distributions.

## 4.2 Expected values

**Notation 4.1.** $E(\cdot) = E[(\cdot)]$ expected value of $(\cdot)$.

**Definition 4.1.** Expected value of a function $\psi(x)$ of the random variable $x$: Let $x$ be a random variable and let $\psi(x)$, ($\psi$ is the Greek letter psi) be a function of $x$. Then the expected value of $\psi(x)$, denoted by $E[\psi(x)] = E\psi(x)$ is defined as

$$\begin{aligned} E[\psi(x)] = E(\psi(x)) &= \sum_{-\infty<x<\infty} \psi(x)f(x) \quad \text{when } x \text{ is discrete} \\ &= \int_{-\infty}^{\infty} \psi(x)f(x)\mathrm{d}x \quad \text{when } x \text{ is continuous} \end{aligned} \tag{4.1}$$

where $f(x)$ is the probability/density function.

https://doi.org/10.1515/9783110562545-004

When $f(x)$ is a density function, obtained by differentiating a distribution function $F(x)$, then instead of writing $f(x)\mathrm{d}x$ we may also write it as $\mathrm{d}F(x)$ in (4.1). Expected values need not exist for all functions $\psi(x)$ and for all random variables $x$.

**Example 4.1.** Consider the probability function:

$$f(x) = \begin{cases} 0.2, & x = -1 \\ 0.3, & x = 0 \\ 0.5, & x = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

Evaluate (i) $E[x]$; (ii) $E[2x^2 - 5x]$; (iii) 8.

**Solution 4.1.** In (i), the function $\psi(x) = x$ and by definition, we multiply $x$ with the probability function and sum up over all values of $x$. Here, the random variable takes only $-1, 0, 2$ with non-zero probabilities and all other values with zero probabilities. Hence when we multiply with the probability function everywhere it will be zeros except at the points $x = -1$, $x = 0$ and $x = 2$. Therefore, for (i),

$$E[x] = \sum_{-\infty<x<\infty} xf(x) = 0 + (-1)(0.2) + (0)(0.3) + (2)(0.5) = 0.8.$$

When $x$ takes the value $-1$, the corresponding probability is 0.2. Hence $xf(x) = (-1)(0.2) = -0.2$ at $x = -1$. Similarly, $xf(x) = (0)(0.3) = 0$ at $x = 0$ and $xf(x) = (2)(0.5) = 1.0$ at $x = 2$. Then we added up all these to get our final answer as 0.8.

(ii) Here, we need the expected value of $2x^2 - 5x$. That is,

$$E[2x^2 - 5x] = \sum_{-\infty<x<\infty} [2x^2 - 5x]f(x) = \sum_{-1,0,2} [2x^2 - 5x]f(x).$$

When $x = -1$, the probability $f(x) = 0.2$, and hence the corresponding value of

$$[2x^2 - 5x]f(x) = [2(-1)^2 - 5(-1)](0.2) = 2(-1)^2(0.2) - 5(-1)(0.2) = 1.4.$$

When $x = 0$, $2x^2 - 5x = 0$, and hence this term will be zero. When $x = 2$,

$$[2x^2 - 5x]f(x) = [2(2)^2 - 5(2)](0.5) = -1.$$

Hence the final sum $(1.4) + (0) + (-1) = 0.4$ is the answer to (ii). We may also note one interesting property that the above expected value can also be computed as

$$E[2x^2 - 5x] = 2E[x^2] - 5E[x].$$

In (iii), we do not have any variable $x$. Hence from the definition

$$E[8] = \sum_{-\infty<x<\infty} 8f(x) = 8 \sum_{-\infty<x<\infty} f(x) = 8$$

since the total probability or $\sum_x f(x) = 1$ by definition. This in fact is a general result.

**Result 4.1.** *Whatever be the random variable under consideration, the expected value of a constant is the constant itself. That is,*

$$E[c]=c \tag{4.2}$$

*whenever c is a constant.*

One more property is evident from the computations in Example 4.1. When a sum was involved in (ii), we could have obtained the final answer by summing up individually also. The following general result follows from the definition itself.

**Result 4.2.** *The expected value of a constant times a function is the constant times the expected value of the function, and the expected value of a sum is the sum of the expected values whenever the expected values exist. That is,*

$$E[c\psi(x)]=cE[\psi(x)] \tag{4.3}$$

$$\begin{aligned}E[a\psi_1(x)+b\psi_2(x)]&=E[a\psi_1(x)]+E[b\psi_2(x)]\\&=aE[\psi_1(x)]+bE[\psi_2(x)]\end{aligned} \tag{4.4}$$

*where a and b are constants and $\psi_1(x)$ and $\psi_2(x)$ are two functions of the same random variable $x$.*

Note that once the expected value is taken the resulting quantity is a constant and it does not depend on the random variable $x$ anymore.

**Result 4.3.** *For any random variable $x$, for which $E(x)$ exists,*

$$E[x-E(x)]=0. \tag{4.5}$$

The proof follows from the fact that $E(x)$ is a constant and the expected value of a constant is the constant itself. Taking expectation by using Result 4.2, we have

$$E[x-E(x)]=E[x]-E[E(x)]=E(x)-E(x)=0.$$

**Example 4.2.** If a discrete random variable takes the values $x_1,\dots,x_n$ with probabilities $p_1,\dots,p_n$, respectively, compute $E[x]$ for the general case as well as for the particular case when $p_1=p_2=\dots=p_n=\frac{1}{n}$.

**Solution 4.2.** We may note that a general discrete random variable is of the general type described in this example. The variable takes some values with non-zero probabilities and other values with zero probabilities. In Example 4.1, the variable took the values $-1,0,2$ with non-zero probabilities and other values with zero probabilities. If we draw a correspondence with Example 4.2, then $x_1=-1$, $x_2=0$, $x_3=2$, $n=3$. Let us consider the special case first. In the special case, all the probabilities are equal to $\frac{1}{n}$ each. Hence

$$\begin{aligned} E[x] &= \sum_{-\infty<x<\infty} xf(x) = 0 + x_1\left(\frac{1}{n}\right) + x_2\left(\frac{1}{n}\right) + \cdots + x_n\left(\frac{1}{n}\right) \\ &= \frac{x_1 + \cdots + x_n}{n} = \bar{x}. \end{aligned} \tag{4.6}$$

In this case, it is the average of the numbers $x_1, \dots, x_n$ or $E[x]$ in this case corresponds to the concept of average. In the general case,

$$E[x] = \sum_{-\infty<x<\infty} xf(x) = (x_1)(p_1) + \cdots + (x_n)(p_n) = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} \tag{4.7}$$

and note that $\sum_{i=1}^{n} p_i = 1$ by definition. The last expression in (4.7) is the expression for the *centre of gravity* of a physical system when $p_1, \dots, p_n$ are weights, or forces acting at the points $x_1, \dots, x_n$. Hence $E[x]$ can be interpreted in many ways.

> $E[x]$ is the *mean value of* $x$ or some sort of an average value of $x$ and it is the centre of gravity of a physical system.

**Example 4.3.** Evaluate the mean value of $x$ if $x$ has the following density:

$$f(x) = \begin{cases} \frac{1}{x^2}, & 1 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 4.3.** Since it is a continuous case, we integrate to find the expected values. Therefore,

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} xf(x)\mathrm{d}x \\ &= 0 + \int_{1}^{\infty} x\left(\frac{1}{x^2}\right)\mathrm{d}x \\ &= \int_{1}^{\infty} \frac{1}{x}\mathrm{d}x = [\ln x]_1^{\infty} = \infty. \end{aligned}$$

Here, the mean value is not a finite quantity. When an expected value is not a definite finite quantity, we say that the expected value does not exist.

> Expected value of $\psi(x)$ does not exist when:
> (i) $E[\psi(x)] = +\infty$
> (ii) $E[\psi(x)] = -\infty$
> (iii) $E[\psi(x)]$ oscillates between finite or infinite quantities, or the sum (in the discrete case) or the integral (in the continuous case) does not converge.

**Example 4.4.** If $x$ is uniformly distributed over the interval $[a, b]$, compute (i): the mean value of $x$, (ii): $E[x - E[(x)]^2$.

**Solution 4.4.** Since we are told that $x$ is uniformly distributed, the density of $x$ is a rectangular or uniform density. That is,

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b,\ b > a \\ 0, & \text{elsewhere.} \end{cases}$$

Hence the mean value of $x$, by definition,

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = 0 + \int_a^b x\frac{1}{b-a}\mathrm{d}x \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}. \end{aligned}$$

Thus, the mean value is the middle point of the interval here or the average of the end points as shown in Figure 4.1.

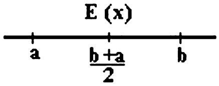


**Figure 4.1:** Expected value in the uniform density.

(ii) For computing this, either we can substitute for $E(x)$ and then compute directly, or simplify first and then compute. Let us simplify first by using Results 4.1 and 4.2:

$$\begin{aligned} E[x-E(x)]^2 &= E[x^2 - 2xE(x) + (E(x))^2] = E\left[x^2 - 2x\left(\frac{a+b}{2}\right) + \left(\frac{a+b}{2}\right)^2\right] \\ &= E(x^2) - 2\left(\frac{a+b}{2}\right)E(x) + \left(\frac{a+b}{2}\right)^2 \end{aligned}$$

since the expected value of a constant is the constant itself, and

$$= E(x^2) - 2\left(\frac{a+b}{2}\right)^2 + \left(\frac{a+b}{2}\right)^2 = E(x^2) - \left(\frac{a+b}{2}\right)^2.$$

This means that we have to only compute $E(x^2)$. But

$$E(x^2) = \int_a^b \frac{x^2}{b-a}\mathrm{d}x = \frac{(b^3-a^3)}{3(b-a)} = \frac{(a^2+ab+b^2)}{3}.$$

Hence

$$E[x-E(x)]^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

This quantity $E[x-E(x)]^2$ is a very important quantity and it is called the variance.

**Notation 4.2.** $\mathrm{Var}(x) = \sigma^2 = \sigma_x^2$ = variance of $x$, where $\sigma$ is the Greek letter sigma.

**Definition 4.2** (Variance of a random variable)**.** It is defined as the following expected value for any random variable $x$:

$$\mathrm{Var}(x) = \sigma_x^2 = E[x - E(x)]^2 = E(x^2) - [E(x)]^2. \tag{4.8}$$

Since $\mathrm{Var}(x) \geq 0$, we have from (4.8), $E(x^2) \geq [E(x)]^2 \Rightarrow [E(x^2)]^{\frac{1}{2}} \geq |E(x)|$. The last part of (4.8) is available by opening up and then simplifying by using Results 4.1 and 4.2 and by using the fact that $E(x)$ is a constant. It is already derived in the solution of Example 4.4.

**Notation 4.3.** $\sigma, \sigma_x, \sqrt{\mathrm{Var}(x)}$: Standard deviation of $x$.

**Definition 4.3** (Standard deviation)**.** The standard deviation of a real random variable $x$ is defined as the positive square root of the variance of $x$. It is abbreviated as $S.D$. That is,

$$S.D = \sigma_x = +\sqrt{\mathrm{Var}(x)}. \tag{4.9}$$

What are the uses of the variance and standard deviation? We have already seen that the mean value = $E(x)$ has the interpretation of average, central value, central tendency of $x$, centre of gravity of a physical system. Similarly, the variance corresponds to the *moment of inertia* in a physical system. Standard deviation is a mathematical distance of $x$ from the point $E(x)$ or from the centre of gravity of the system, and hence the standard deviation can be taken as a *measure of scatter or dispersion of* $x$ from the mean value = $E(x)$. A high value for the standard deviation means that the variable is more spread out and small value for standard deviation means that the variable is concentrated around the centre of gravity of the system or around the central value = $E(x)$.

### 4.2.1 Measures of central tendency

One measure of central tendency of the random variable $x$ is the mean value or expected value of $x$, $E(x)$. Other measures in common use are the median and the mode. The idea of a median value is that we are looking for a point $M$ such that the probability of the random variable $x$ falling below is equal to the probability of $x$ falling above $M$. In this sense, $M$ is a middle point. $M$ may not be unique and there may be several points qualifying to be medians for a given $x$. The following is a formal definition for the median.

**Notation 4.4.** $M$: median of $x$.

**Definition 4.4** (Median of a random variable $x$). Let $M$ be a point such that

$$\Pr\{x \le M\} \ge \frac{1}{2}$$

and

$$\Pr\{x \ge M\} \ge \frac{1}{2}.$$

All points $M$ satisfying these two conditions are called median points.

In some cases, we can have a unique point $M$ and in other cases we can have several values for $M$. We will examine some discrete and some continuous cases.

**Example 4.5.** Consider the following probability functions:

$$f_1(x) = \begin{cases} 0.4, & x = 1 \\ 0.2, & x = 2 \\ 0.4, & x = 7 \\ 0, & \text{elsewhere}; \end{cases}$$

$$f_2(x) = \begin{cases} 0.5, & x = -1 \\ 0.5, & x = 5 \end{cases}$$

Compute the medians in each case.

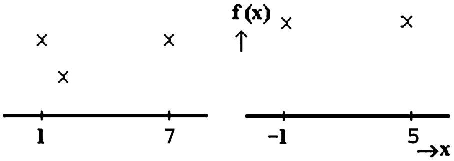


**Figure 4.2:** Left: Probability function $f_1(x)$; Right: Probability function $f_2(x)$.

**Solution 4.5.** In $f_1(x)$ of Figure 4.2, the point $x = 2$ satisfies both the conditions:

$$\Pr\{x \le 2\} = 0.6 > 0.5 \quad \text{and} \quad \Pr\{x \ge 2\} = 0.6 > 0.5$$

and hence $x = 2$ is the unique median point for this $x$.

In $f_2(x)$, any point from $-1$ to 5 will qualify or $-1 \le M \le 5$. If a unique value is preferred, then one can take the middle value of this interval, namely, $x = 2$ as the median.

**Example 4.6.** Compute the median point or points for the following densities:

$$f_1(x) = \begin{cases} \frac{1}{2}x, & 0 \le x \le 2 \\ 0, & \text{elsewhere}; \end{cases}$$

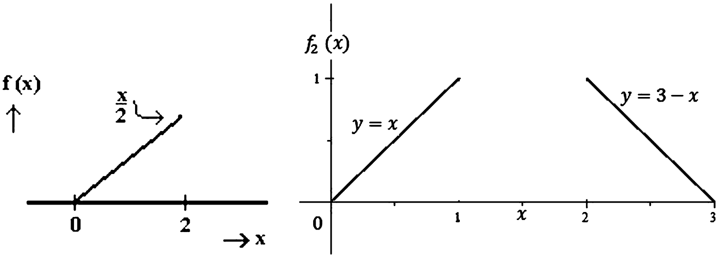


**Figure 4.3:** Left: Density $f_1(x)$; Right: Density $f_2(x)$.

$$f_2(x) = \begin{cases} x, & 0 \le x \le 1 \\ 3 - x, & 2 \le x \le 3 \\ 0, & \text{otherwise.} \end{cases}$$

**Solution 4.6.** For $f_1(x)$ of Figure 4.3, we can get a unique median point. Let us compute the probability

$$\Pr\{x \le M\} = \frac{1}{2}\int_0^M x\mathrm{d}x = \frac{M^2}{4}.$$

We can equate this to $\frac{1}{2}$ because in this case

$$\Pr\{x \le M\} = \frac{1}{2} = \Pr\{x \ge M\}.$$

Equating $\frac{M^2}{4}$ to $\frac{1}{2}$ we have $M = \sqrt{2}$.

In the case of $f_2(x)$, let us compute the probabilities in the two pieces where we have non-zero functions:

$$\int_0^1 x\mathrm{d}x = \frac{1}{2}$$

and

$$\int_2^3 (3 - x)\mathrm{d}x = \frac{1}{2}.$$

This means that any point from 1 to 2, $1 \le M \le 2$ will qualify to be a median point. If a single point is preferred, then 1.5 is the candidate.

Another measure used as a measure of central tendency of a random variable $x$ is the mode. This is grouped with measures of central tendency but it does not have much to do with measuring central tendency. The point(s) corresponding to local maximum (maxima) for the density curve (in the continuous case) is (are) taken as mode(s) in the continuous case, and the point(s) corresponding to the local maximum probability mass is (are) taken as mode(s) in the discrete case. In Example 4.5, we would have taken 1 and 7 as modes in $f_1(x)$, and −1 and 5 as modes in $f_2(x)$. In Example 4.6, we would have taken $x = 2$ as the mode in $f_1(x)$ and 1 and 2 as modes in $f_2(x)$.

### 4.2.2 Measures of scatter or dispersion

We have already introduced one measure of scatter, which is the standard deviation. This is a measure of the spread of $x$ from $E(x)$. If an arbitrary point $a$ is taken, then $\sqrt{E[x-a]^2}$ is a measure of scatter of $x$ from the point $a$, which is also called the *mean deviation of $x$ from $a$*. Other measures of scatter from $a$ are the following:

$$M_r(a) = E\left[|x-a|^r\right]^{\frac{1}{r}}, \quad r = 1,2,\ldots \tag{4.10}$$

$$M_1(a) = E|x-a| = \text{mean absolute deviation from } a \tag{4.11}$$

All these qualify to be mathematical distances of $x$ from the point $a$ and in this sense are measures of scatter of $x$ from the point $a$. Out of these, the mean deviation from $a$ and mean absolute deviation from $a$ are very important because they are very often used for statistical decision making. Hence we will list two basic properties here.

**Result 4.4.** *The mean deviation from $a$ is least when $a = E(x)$.*

**Proof.** Minimization of $\sqrt{E(x-a)^2}$ also implies the minimization of the square $E[x-a]^2$ and vice versa. Consider $E[x-a]^2$, add and subtract $E(x)$ inside and expand to obtain the following:

$$\begin{aligned} E[x-a]^2 &= E\left[x-E(x)+E(x)-a\right]^2 \\ &= E\left[x-E(x)\right]^2 - 2E\left[x-E(x)\right]\left[E(x)-a\right] + \left(E(x)-a\right)^2 \\ &= \mathrm{Var}(x) + \left[E(x)-a\right]^2 \end{aligned}$$

since $E(x)-a$ is a constant and $E[x-E(x)] = 0$. But the right-hand side has only one term depending on $a$ and both terms are non-negative. Hence the minimum is attained when the term containing $a$ is zero, that is, $[E(x)-a]^2 = 0 \Rightarrow a = E(x)$.

This result is very important in model building problems, in regression analysis, etc.

**Result 4.5.** *The mean absolute deviation is least when the deviations are taken from the median or $E|x-a|$ is least when $a$ = the median of $x$.*

The proof is slightly longer and it will be listed in the exercises.

### 4.2.3 Some properties of variance

(i) $\mathrm{Var}(c) = 0$ *when $c$ is a constant.*

The proof follows from the definition itself.

$$\mathrm{Var}(x) = E[x - E(x)]^2 = E[c - c]^2 = 0 \tag{4.12}$$

since $E(c) = c$ when $c$ is a constant.

(ii) *Let $y = x + a$ where $a$ is a constant. Then*

$$\mathrm{Var}(y) = \mathrm{Var}(x + a) = \mathrm{Var}(x)$$

*or the relocation of the variable $x$ does not affect the variance.*

When a constant is added to the variable, we say that the variable is relocated and $a$ here is the *location parameter*. Observe that $E(x + a) = E(x) + a$ and then

$$y - E(y) = (x + a) - (E(x) + a) = x - E(x)$$

thus $a$ is canceled. As an example let us consider the following problem. One pumpkin is randomly selected from a heap of pumpkins and weighed. Let $x_1$ be the weight observed. Later it came to the attention that the balance was defective and it always showed 100 grams more than the actual weight. If $x$ is the true weight of the pumpkin, then the observed weight $x_1 = x + 100$, weight being measured in grams. Will this faulty balance affect the mean value and variance of the true weight of a pumpkin selected at random? Let $y = x + c$ where $x$ is a random variable and $c$ is a constant. Then, denoting the mean value by $\mu$ (mu) and variance by $\sigma^2$ (sigma squared), we have

$$\mu = E(y) = E(x + c) = E(x) + c \tag{4.13}$$

The mean value is affected:

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(y) = E[y - E(y)]^2\\ &= E[(x + c) - (E(x) + c)]^2 = E[x - E(x)]^2 = \mathrm{Var}(x).\end{aligned} \tag{4.14}$$

Note that when the deviation $y - E(y)$ is taken the constant is eliminated, and hence the relocation of the variable will affect the mean value but it will not affect the variance.

The pumpkin was weighed in grams. Suppose we want to convert the weight into kilograms. Then

$$x \text{ grams} \quad \Rightarrow \quad z = \frac{1}{1000} x \text{ kilograms.}$$

Let us compare the variance of $x$ and variance of $z$. In general, let $z = bx$ where $b$ is a constant and let us denote the variances of $z$ and $x$ by $\sigma_z^2$ and $\sigma_x^2$, respectively, and the corresponding standard deviations by $\sigma_z$ and $\sigma_x$. Then by definition

$$\begin{aligned}\sigma_z^2 &= E[z - E(z)]^2 \quad \text{where } E(z) = E(bx) = bE(x)\\ &= E[bx - bE(x)]^2 = b^2E[x - E(x)]^2 = b^2\,\mathrm{Var}(x)\\ &= b^2\sigma_x^2.\end{aligned} \tag{4.15}$$

Thus the variance of $z$ is square of the scaling factor times the variance of $x$. Since standard deviation is the positive square root of the variance

$$\sigma_z = |b|\sigma_x. \tag{4.16}$$

Remember to take the absolute value. If the original $b$ is $-2$, then the multiplying factor for the standard deviation is $|-2| = 2$ and not $-2$. The term "scaling" came from multiplication by a positive constant but nowadays it is used to denote a constant multiple whether the constant is positive or negative, as long as it is not zero. If we look at a scaling and relocation, that is, let $t = cx + d$ where $c$ and $d$ are constants. Then

$$t = cx + d \quad \Rightarrow \quad \mathrm{Var}(t) = c^2\,\mathrm{Var}(x); \quad \sigma_t = |c|\sigma_x. \tag{4.17}$$

Some general properties may be observed. In general (for some special cases they may hold),

(iii) $\sigma_x^2 = E(x - E(x)]^2$ does not imply $\sigma_x = E[x - E(x)]$;

(iv) $\sigma_x^2 = E[x - E(x)]^2$ does not imply $\sigma_x = E[|x - E(x)|]$;

(v) $\sigma_x^2 = 0$ if and only if $x$ is a degenerate random variable. That is, if $x$ is degenerate then $\sigma_x^2 = 0$ and if $\sigma_x^2 = 0$ then $x$ is degenerate. That means the only random variable where variance is zero is the degenerate random variable. If and only if is usually abbreviated as"iff".

(vi) $\sigma_x^2 > 0$, $\sigma_x > 0$ always except for the degenerate case where $\sigma_x^2 = 0$. That is, the variance, thereby the standard deviation, of a random variable is non-negative.

## Exercises 4.2

**4.2.1.** Compute (i) $E(x)$; (ii) $E[x^3 - 2x + 5]$; (iii) $E[x - 2]^2$; (iv) $\mathrm{Var}(x)$ for the random variable in

$$f(x) = \begin{cases} 0.5, & x = -1 \\ 0.5, & x = 1 \\ 0, & \text{elsewhere.} \end{cases}$$

**4.2.2.** Compute the same items in Exercise 4.2.1 for the random variable in

$$f(x) = \begin{cases} cx, & 0 \le x \le 2 \\ 0, & \text{elsewhere.} \end{cases}$$

Determine the normalizing constant $c$ first.

**4.2.3.** Compute the same items in Exercise 4.2.1 for the random variable in

$$f(x) = \begin{cases} cx^2, & 0 \le x \le 1 \\ 2-x, & 1 < x \le 2 \\ 0, & \text{elsewhere.} \end{cases}$$

Evaluate the normalizing constant $c$ first.

**4.2.4.** Construct two examples where $E(x) = 0$, $\mathrm{Var}(x) = 1$ in discrete case.

**4.2.5.** Construct two examples where $E(x) = 0$, $\mathrm{Var}(x) = 1$ in a continuous case.

**4.2.6.** Compute $E[x-2]^2$ for the Exercises in 4.2.1, 4.2.2, 4.2.3 first directly by taking the expected value of $E[x-2]^2$ and then by expanding $(x-2)^2$, taking the expected values and simplifying. Verify that both procedures give the same results.

**4.2.7.** Construct a probability/density function of $x$ where $E[|x|] = 0$.

**4.2.8.** Let

$$f(x) = \begin{cases} cx^3(1-x)^2, & 0 \le x \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

be a density. Compute (i) the normalizing constant $c$; (ii) the median point; (iii) mode or modes.

**4.2.9.** For Exercise 4.2.3, compute (i) $c$; (ii) the median; (iii) the mode or modes.

**4.2.10.** Construct a density function of a random variable $x$ where $E(x)$ exists whereas $E[x^2]$ does not exist.

**4.2.11.** In Exercises 4.2.1–4.2.3, compute the following probabilities: (i) $\Pr\{-2 \le x \le 0.8\}$; (ii) $\Pr\{0.5 \le x \le 1.8\}$; (iii) mark the areas in a graph in (i) and (ii) for Exercises 4.2.2 and 4.2.3.

**4.2.12.** The monthly income $x$ of households in a city is found to follow the distribution with the density:

$$f(x) = \begin{cases} \frac{c}{x^3}, & 1\,000 \le x \le 30\,000 \\ 0, & \text{elsewhere.} \end{cases}$$

Compute (i) $c$; (ii) the expected income or mean value of the income; (iii) the median income in this city so that 50% of the households have income below that and 50% above that.

**4.2.13.** The waiting time $t$ at a bus stop is found to be exponentially distributed with the density

$$f(t) = \begin{cases} \frac{1}{5}\mathrm{e}^{-\frac{t}{5}}, & 0 \le t < \infty \\ 0, & \text{otherwise,} \end{cases}$$

time being measured in minutes. Compute (i) the expected waiting time $\mu$ (mu) for this bus; (ii) the standard deviation $\sigma$ for this waiting time; (iii) the probability $\Pr\{\mu - \sigma < t < \mu + \sigma\}$ or the waiting time being one standard deviation away from the mean value; (iv) $\Pr\{|x - \mu| \le 2\sigma\}$; (v) $\Pr\{|x - \mu| \le 3\sigma\}$.

**4.2.14.** For the exponential density

$$f(x) = \begin{cases} \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, & 0 \le x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

prove that the mean value of $x$, that is, $E[x]$ is this parameter $\theta$. Compute the variance also. [Observe that $\theta$ is a scaling parameter here.]

**4.2.15.** The life time $t$ of a certain type of electric bulbs is found to be exponentially distributed with expected life time 1 000 hours, time being measured in hours. One bulb from this type of bulbs has already lasted for 1 500 hours (i.e., $t \ge 1500$). Let $A$ be this event. Let $B$ be the event that it will last for at least another 500 hours more. Interpret the events $A, A \cap B$ and then compute the probability that it will last at least another 500 hours more given that it has already lasted for 1 500 hours.

## 4.3 Higher moments

In Section 4.2, we define expected value in general as well as mean value = $E(x)$ and variance = $E[x - E(x)]^2$. Here, we define certain expected values which are called the moments. The concept of moments originally came from physics but used heavily in statistical literature.

**Notation 4.5.** $\mu_r{}'$: $r$-th integer moment about the origin.

**Definition 4.5** (Moments about the origin). The $\alpha$-th moment or the $\alpha$-th moment about the origin of a random variable $x$ is defined as

$$\begin{aligned} E[x^\alpha] &= \int_{-\infty}^{\infty} x^\alpha f(x)\mathrm{d}x \quad \text{when } x \text{ is continuous} \\ &= \sum_{-\infty<x<\infty} x^\alpha f(x) \quad \text{when } x \text{ is discrete} \end{aligned}$$

and when $E(x^\alpha]$ exists. Here, $\alpha$ could be any complex number. When $\alpha$ is a positive integer, $\alpha = r = 1, 2, \ldots$ then the $r$-th moment is denoted by $\mu_r{}'$ or

$$\mu_r{}' = E[x^r], \quad r = 1, 2, \ldots \tag{4.18}$$

In a mixed case, where some probability mass is distributed on some individually distinct points and the remaining on a continuum of points, take $x^\alpha f(x)$, sum up over the discrete points and integrate over the continuum of points to get $E[x^\alpha]$. Observe that $E[x^0] = E[1] = 1$.

**Notation 4.6.** $\mu_r = r$-th central moment for $r = 1, 2, \ldots$

**Definition 4.6** (Central moments). The $\alpha$-th central moment is defined as $E[x - E(x)]^\alpha$ for any complex number $\alpha$, whenever it exists, and when $\alpha = r = 1, 2, \ldots$ then the $r$-th central moment is defined as

$$\mu_r = E[x - E(x)]^r, \quad r = 1, 2, \ldots \tag{4.19}$$

The $\alpha$-th moment about any arbitrary point $a$ is defined as $E[x - a]^\alpha$ whenever this expected value exists. Thus when $a = E(x)$ we have the central moments.

**Notation 4.7.** $\mu_{[r]} = r$-th factorial moment.

**Definition 4.7** (Factorial moments). The $r$-th factorial moment is defined as

$$\mu_{[r]} = E[x(x-1)(x-2)\cdots(x-r+1)] \tag{4.20}$$

whenever it exists.

Factorial moments are usually easier to evaluate compared to moments about the origin or central moments when factorials are involved in the denominator in some probability functions for discrete variables. This will be seen when we evaluate moments in binomial or Poisson probability functions later on. Before proceeding further, let us look into some examples.

**Example 4.7.** Compute (i) $E(x^3)$; (ii) $E[x - E(x)]^4$; (iii) $E|x|$; (iv) $E(x^2 - 3x + 4)$ for the following probability function:

$$f(x) = \begin{cases} 0.5, & x = -2 \\ 0.5, & x = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 4.7.** Since it is a discrete case we sum up. Denoting summation over $x$ by $\sum_x$ we have,

(i) $E(x^3) = \sum_x x^3 f(x) = (-2)^3(0.5) + (2)^3(0.5) + 0 = -4 + 4 = 0$. In this case, it is easily seen that $E(x^r) = 0$, $r = 1, 3, 5, \ldots$, or all odd moments are zeros here, since the probability function is symmetric about $x = 0$.

(ii) Since $E(x) = 0$ here, we have

$$\begin{aligned} E[x - E(x)]^4 &= E[x^4] = \sum_x x^4 f(x) \\ &= (-2)^4(0.5) + (2)^4(0.5) + 0 = (16)(0.5) + (16)(0.5) = 16. \end{aligned}$$

(iii) Here, we take the absolute values

$$E|x| = \sum_x |x| = |(-2)|(0.5) + |(2)|(0.5) + 0 = (2)(0.5) + (2)(0.5) = 2.$$

(iv) Here, we can use the property that the expected value of a sum is the sum of the expected values:

$$E(x^2 - 3x + 4) = E(x^2) - 3E(x) + E(4) = E(x^2) - 3(0) + 4$$

since $E(x) = 0$ and since $E(4) = 4$ and $= 4 + 4 = 8$ since $E(x^2) = (-2)^2(0.5) + (2)^2(0.5) = 4$.

**Example 4.8.** Answer the same questions in Example 4.7 for the following density:

$$f(x) = \begin{cases} x, & 0 \le x < 1 \\ 2 - x, & 1 \le x \le 2 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 4.8.** Since the variable is continuous here, we will integrate. From $-\infty$ to 0, the function $f(x)$ is zero and the integral over zero is zero:

$$E(x^3) = \int_x x^3 f(x)\mathrm{d}x = 0 + \int_0^1 x^3(x)\mathrm{d}x + \int_1^2 [x^3(2 - x)]\mathrm{d}x$$

since $f(x)$ has different forms we integrate on each piece, and then

$$\begin{aligned} E(x^3) &= \left[\frac{x^5}{5}\right]_0^1 + \left[2\frac{x^4}{4} - \frac{x^5}{5}\right]_1^2 \\ &= \frac{1}{5} + \frac{15}{2} - \frac{31}{5} = \frac{3}{2}. \end{aligned}$$

Since $x \ge 0$ here, $|x| = x$ itself. Let us compute $E(x)$.

$$E(x) = 0 + \int_0^1 x(x)\mathrm{d}x + \int_1^2 [x(2 - x)]\mathrm{d}x = \frac{1}{3} + \frac{2}{3} = 1.$$

This may also be seen from the areas in the density function when you graph the density. The mid-value is 1. Thus, (iii) is answered, the value is 1. Let us compute $E(x^2)$:

$$E(x^2) = 0 + \int_0^1 x^2(x)\mathrm{d}x + \int_1^2 x^2(2 - x)\mathrm{d}x = \frac{7}{6}.$$

Then (iv) can be answered:

$$E(x^2 - 3x + 4) = E(x^2) - 3E(x) + 4 = \frac{7}{6} - 3(1) + 4 = \frac{13}{6}.$$

Now, (ii) can be answered in two ways, either by expanding $E[x - E(x)]^4$ first by using the binomial expansion and then taking the expected values or by substituting for $E(x)$ first and then taking the expected values directly. Let us expand after substituting the value for $E(x) = 1$, by using a binomial expansion. That is,

$$\begin{aligned} & E[x - E(x)]^4 \\ & \quad = E[x-1]^4 = E[x + (-1)]^4 \\ & \quad = E\left[x^4 + \binom{4}{1}x^3(-1) + \binom{4}{2}x^2(-1)^2 + x\binom{4}{3}(-1)^3 + \binom{4}{4}(-1)^4\right] \\ & \quad = E[x^4 - 4x^3 + 6x^2 - 4x + 1] = E(x^4) - 4E(x^3) + 6E(x^2) - 4E(x) + 1 \\ & \quad = E(x^4) - 4\left(\frac{3}{2}\right) + 6\left(\frac{7}{6}\right) - 4(1) + 1 = E(x^4) - 2. \end{aligned}$$

Now, we will compute $E(x^4)$.

$$E(x^4) = 0 + \int_0^1 x^4(x)\mathrm{d}x + \int_1^2 x^4(2-x)\mathrm{d}x = \frac{31}{15}.$$

Hence

$$E[x - E(x)]^4 = \frac{31}{15} - 2 = \frac{1}{15}.$$

**Example 4.9.** Prove that the following function is a probability function for a discrete random variable and then compute the second factorial moment:

$$f(x) = \begin{cases} \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, & x = 0, 1, 2, \ldots, \lambda > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

[This probability function is known as *Poisson probability function*, named after its inventor, S. Poisson, a French mathematician.]

**Solution 4.9.** Since exponential function is non-negative, $f(x)$ here is non-negative for all values of $x$ and $\lambda > 0$. Here, $\lambda > 0$ is a parameter. The total probability is

$$\sum_x f(x) = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}.$$

But

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots = \mathrm{e}^{\lambda}$$

by the exponential series. Hence the total probability is

$$e^{-\lambda}e^{\lambda} = e^{-\lambda+\lambda} = e^{0} = 1.$$

Hence $f(x)$ here is a probability function. The second factorial moment

$$\mu_{[2]} = E[x(x-1)] = \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=2}^{\infty} x(x-1)\frac{\lambda^x}{x!}.$$

But note that when $x = 0$ or $x = 1$ the right side is zero, and hence the sum starts only from $x = 2$, going to infinity. When $x$ goes from 2 to $\infty$ or when $x \neq 0$ and $x \neq 1$, we can divide both numerator and denominator by $x(x-1)$ or we can cancel $x(x-1)$. That is,

$$\frac{x(x-1)}{x!} = \frac{1}{(x-2)!}.$$

But this cancellation, or division of numerator and denominator by the same quantity, was not possible if $x$ could be zero or 1 because division by zero is impossible. Now, we can sum up. For convenience, let us take $\lambda^2$ and $e^{-\lambda}$ outside. Hence

$$\begin{aligned} E[x(x-1)] &= \lambda^2 e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{n-2}}{(n-2)!} \\ &= \lambda^2 e^{-\lambda}\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right] \\ &= \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2. \end{aligned}$$

Here, the sum was opened up by putting $x = 2$, $x = 3$, $x = 4$ and so on and writing up the terms. We could have also used the procedure of substitution. Put $y = x - 2$ then, when $x = 2$, we have $y = 0$ and then

$$\sum_{x=2}^{\infty}[\cdot] = \sum_{y=0}^{\infty}[\cdot]$$

which would have also yielded the same result.

In this example, we may note one interesting aspect. Since $x!$ was sitting in the denominator, we could easily cancel factors such as, $x$, $x(x-1)$, $x(x-1)(x-2)$, etc., and thus factorial moments are easy to evaluate. But we cannot cancel $x^2$ because after canceling one $x$ still there is one more $x$ left out and the denominator has become $(x-1)!$. If we want to compute $E[x^2]$ in this case, we can use the identity

$$x(x-1) = x^2 - x \Rightarrow x^2 \equiv x(x-1) + x.$$

Higher moments about the origin can be computed from factorial moments by using such identities in this Poisson case because of the presence of $x!$ in the denominator of the probability function.

### 4.3.1 Moment generating function

There are several types of generating functions which will all generate integer moments about the origin.

**Notation 4.8.** $M(t)$: Moment generating function of a random variable $x$.

**Definition 4.8** (Moment generating function)**.** The moment generating function of a random variable $x$ or of the probability/density function $f(x)$ is defined as the following integral:

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\mathrm{d}x \quad \text{when } x \text{ is continuous} \\ &= \sum_{-\infty<x<\infty} \mathrm{e}^{tx} f(x) \quad \text{when } x \text{ is discrete} \end{aligned}$$

whenever the sum/integral exists. In the mixed case, integrate over the continuum of points and sum up over the discrete points where there are non-zero probability masses.

Why is $M(t)$ called the moment generating function? Let us expand $\mathrm{e}^{tx}$. That is, for example, for the continuous situation,

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left[1 + \frac{tx}{1!} + \frac{t^2x^2}{2!} + \cdots\right] f(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} f(x)\mathrm{d}x + \frac{t}{1!}\int_{-\infty}^{\infty} xf(x)\mathrm{d}x + \cdots \end{aligned}$$

if term by term integration is possible, and

$$M(t) = 1 + \frac{t}{1!}E(x) + \frac{t^2}{2!}E(x^2) + \frac{t^3}{3!}E(x^3) + \cdots \tag{4.21}$$

Thus the coefficient of $\frac{t^r}{r!}$ in the above power series is $E(x^r)$ the $r$-th integer moment of $x$ for $r = 0, 1, 2, \ldots$. Thus all the integer moments about the origin are generated by this function $M(t)$, and hence it is called a moment generating function. Thus, if $M(t)$ exists and admits a power series in $t$ then the coefficient of $\frac{t^r}{r!}$ is the $r$-th integer moment $E(x^r)$. If $M(t)$ is differentiable, then you may differentiate successively and evaluate at $t = 0$. From (4.21), note that

$$\mu_r{}' = E[x^r] = \frac{\mathrm{d}^r}{\mathrm{d}t^r}M(t)\Big|_{t=0} \tag{4.22}$$

$$= \text{coefficient of } \frac{t^r}{r!} \text{ in the series.} \tag{4.23}$$

Differentiation is possible when $M(t)$ is differentiable and series expansion is possible when $M(t)$ can be expanded into a power series.

This $M(t)$ may not exist for many of the probability/density functions. Let us check one example.

**Example 4.10.** Check whether the moment generating function $M(t)$ exists for the following density:

$$f(x) = \begin{cases} \frac{1}{x^2}, & 1 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 4.10.** It was already shown earlier that this $f(x)$ is a density function. Here,

$$M(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\mathrm{d}x = 0 + \int_1^{\infty} \frac{1}{x^2}\mathrm{e}^{tx}\mathrm{d}x.$$

Integrating by parts, taking $u = \mathrm{e}^{tx}$ and $\mathrm{d}v = \frac{1}{x^2}$ so that $v = \int \frac{1}{x^2}\mathrm{d}x = -\frac{1}{x}$ and then using the formula $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$ we have

$$\begin{aligned} M(t) &= \left[-\frac{1}{x}\mathrm{e}^{tx}\right]_1^{\infty} + t\int_1^{\infty} \frac{1}{x}\mathrm{e}^{tx}\mathrm{d}x \\ &= \left[-\frac{1}{x}\mathrm{e}^{tx}\right]_1^{\infty} + t[\ln x\mathrm{e}^{tx}]_1^{\infty} - t^2\int_1^{\infty} \ln x\mathrm{e}^{tx}\mathrm{d}x. \end{aligned}$$

The first term goes to $-\infty$ for $t > 0$ since the exponential term increases faster than $x$. The integral does not exist for $t > 0$. There is another generating function which will exist always when $f(x)$ is a density. This is known as the characteristic function of the random variable $x$.

**Notation 4.9.** $\phi(t)$ = the characteristic function of $x$ or of $f(x)$.

**Definition 4.9** (Characteristic function)**.** The characteristic function $\phi(t)$ is defined as

$$\begin{aligned} \phi(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{itx} f(x)\mathrm{d}x, \quad i = \sqrt{-1} \text{ when } x \text{ is continuous} \\ &= \sum_{-\infty<x<\infty} \mathrm{e}^{itx} f(x) \quad \text{when } x \text{ is discrete.} \end{aligned} \tag{4.24}$$

Mixed cases can be handled as stated before. Since

$$|\mathrm{e}^{itx}| = |\cos tx + i\sin tx| = \sqrt{\cos^2 tx + \sin^2 tx} = 1 \tag{4.25}$$

we have

$$
\begin{aligned}
|\phi(t)| = \left|\int_{-\infty}^{\infty} \mathrm{e}^{itx} f(x)\mathrm{d}x\right| &\leq \int_{-\infty}^{\infty} |\mathrm{e}^{itx}| f(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1
\end{aligned} \tag{4.26}
$$

by definition. Hence the integral is always convergent.

Mathematical concepts corresponding to moments, moment generating function, etc. are given in a note below. Those who are not familiar with complex variables may skip the notes and go directly to Chapter 5.

### 4.3.2 Moments and Mellin transforms

We have defined arbitrary moments. One such moment is called the Mellin transform of a function. Consider a function $f(x)$ defined over $0 \leq x < \infty$ and consider an integral of the following type.

**Notation 4.10.** $M_f(s)$: Mellin transform of the function $f(x)$, with the Mellin parameter $s$.

**Definition 4.10** (Mellin transform). The Mellin transform of a function $f(x)$ is given by

$$
M_f(s) = \int_0^{\infty} x^{s-1} f(x)\mathrm{d}x \tag{4.27}
$$

provided the integral exists, where $s$ is a complex variable. Note that when $f(x)$ is a density function then it is nothing but the $(s-1)$-th moment of $x$. Mellin transform is defined for $f(x)$ where $f(x)$ need not be a density. In Example 4.8, we could have computed arbitrary moments of the type $E[x^{s-1}]$ where $s$ is a complex variable. Thus in this case the Mellin transform of the density function in Example 4.8 exists.

One question that is often asked is that suppose that you know a function of $s$, which is the Mellin transform of some unknown function $f(x)$, can we determine $f(x)$? Then $f(x)$ will be called the inverse Mellin transform. For statisticians, this problem is of great interest. We may be able to come up with a Mellin transform through some procedure. Then the problem is to determine the unknown density which produced that Mellin transform. The formula to recover the unknown function from a given Mellin transform is the following:

$$
f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} M_f(s) x^{-s}\mathrm{d}s \tag{4.28}
$$

where $i = \sqrt{-1}$. This formula holds under some conditions on the unknown function and the known $M_f(s)$. This integral is a contour integral or an integral in the complex plain. Since this area is beyond the scope of this book, we will not elaborate on this aspect here.

**Notation 4.11.** $F_f(t)$: Fourier transform of $f(x)$ with parameter $t$.

**Definition 4.11** (Fourier transform). Fourier transform of a function $f(x)$, with parameter $t$ is defined as

$$F_f(t) = \int_{-\infty}^{\infty} \mathrm{e}^{-itx} f(x)\mathrm{d}x, \quad i = \sqrt{-1} \tag{4.29}$$

This integral is in the complex domain, and hence we will not elaborate here. Note that the characteristic function $\phi(t)$, and the Fourier transform $F_f(t)$ when $f(x)$ is a density function, will also generate the moments. Fourier transform is defined for arbitrary functions, need not be density functions.

Another generating function which is applicable on positive random variables is the Laplace transform. This will also generate moments when $f(x)$ is a density function but the Laplace transform is defined over all $f(x)$ as long as the integral exists.

**Notation 4.12.** $L_f(t)$: Laplace transform of the function $f(x)$, with the Laplace parameter $t$.

**Definition 4.12** (Laplace transform). The Laplace transform of a function $f(x)$ is defined as the following integral:

$$L_f(t) = \int_0^{\infty} \mathrm{e}^{-tx} f(x)\mathrm{d}x \tag{4.30}$$

whenever this integral is convergent, where $t$ is called the Laplace parameter or the parameter in the Laplace transform.

Such an integral need not exist for all functions $f(x)$. Let us expand $\mathrm{e}^{-tx}$. If $f(x)$ is a density function for a positive random variable $x$ and if $L_f(t)$ admits a power series expansion or if all integer moments exist, then as in the case of the moment generating function, one can obtain the integer moments:

$$E(x^r) = (-1)^r \frac{\mathrm{d}^r}{\mathrm{d}t^r} L_f(t)\Big|_{t=0} \quad \text{when } L_f(t) \text{ is differentiable} \tag{4.31}$$

$$= \text{coefficient of } \frac{(-t)^r}{r!} \quad \text{when } L_f(t) \text{ is expansible.} \tag{4.32}$$

Let $M_x(t)$ be the moment generating function of a real random variable $x$ then the following properties follow from the definition itself:

$$(1): \lim_{t\to 0} M_x(t) = 1; \quad (2): M_{ax}(t) = M_x(at); \quad (3): M_{ax+b}(t) = \mathrm{e}^{tb} M_x(at) \tag{4.33}$$

where $a$ and $b$ are constants. Similar properties hold for characteristic function and Laplace transform (for positive variables).

### 4.3.3 Uniqueness and the moment problem

For a given random variable $x$, whether it is continuous, discrete or mixed, there is only one distribution function (cumulative probability or density), one moment generating function, one characteristic function, corresponding to this variable. There cannot be two different density functions or two different moment generating functions, etc. corresponding to a given random variable. What about moments? Suppose that all integer moments are available, that is, suppose that $E(x^r)$, $r = 0, 1, 2, \ldots$ are all fixed or given. Will the random variable $x$ be uniquely determined by this moment sequence? This is known as the *moment problem*. This problem originated in physics. The answer is that these moments, even though countably infinite of them are available, still these need not uniquely determine the random variable or there can be more than one random variable giving rise to the same moment sequence. Some sets of sufficient conditions are available in the literature under which a moment sequence will uniquely determine the random variable. One such condition is that the variable has a finite range with non-zero density/probability. The non-zero part of the density is in the finite range $a \le x \le b$ where $a$ and $b$ are finite. One such example is the uniform density.

What about the Mellin transform, which will be $(s-1)$-th moment when $f(x)$ is a density for $x \ge 0$, will it determine the density? In Mellin transform $M_f(x) = E(x^{s-1})$, $s$ is a complex variable. As long as $s$ is defined on a strip in the complex plain where the function $M_f(s)$ is analytic, then we can show that $f(x)$ is uniquely determined under some minor additional conditions and inverse Mellin transform, given in (4.28), is the unique determination of $f(x)$. Detailed conditions for the existence of Mellin and inverse Mellin transform may be seen from the book [2], which is available at CMS. Similarly, from the given moment generating function, characteristic function, Laplace transform, Fourier transform, the function $f(x)$ can be uniquely determined under specific conditions. These formulae are known as the inverse Laplace transform, inverse Fourier transform, etc. What we will do later is not the evaluation of these inverse transforms by using complex analysis but remembering that these inverse transforms will uniquely determine the original function $f(x)$ we will identify the transforms with known transforms and write up $f(x)$.

## Exercises 4.3

For the following probability/density functions evaluate the moment generating function $M(t)$ and then obtain the $r$-th integer moment by (i) differentiation when $M(t)$ is differentiable, (ii) by series expansion when $M(t)$ can be expanded as a power series:

**4.3.1.** $f(x) = \begin{cases} 0.7, & x = -1 \\ 0.3, & x = 2, \\ 0, & \text{elsewhere.} \end{cases}$

**4.3.2.** $f(x) = \begin{cases} pq^{x-1}, & x = 1, 2, \dots, \ 0 < p < 1, \ q = 1 - p \\ 0, & \text{elsewhere.} \end{cases}$

**4.3.3.** $f(x) = \begin{cases} \frac{1}{3}, & 0 \le x \le 3 \\ 0, & \text{elsewhere.} \end{cases}$

**4.3.4.** $f(x) = \begin{cases} \theta \mathrm{e}^{-\theta x}, & x \ge 0, \ \theta > 0 \\ 0, & \text{elsewhere.} \end{cases}$

**4.3.5.** Compute $E(x^\alpha)$ for $\alpha$ a complex number for Exercises 4.3.1 and 4.3.3.

**4.3.6.** Prove that $E|x - a|$ is a minimum when $a$ is the median of the random variable $x$.

# 5 Commonly used discrete distributions

## 5.1 Introduction

In this chapter, we will deal with some discrete distributions and in the next chapter we will consider continuous distributions. The most commonly appearing discrete distributions are associated with Bernoulli trials. In a random experiment if each outcome consists of only two possibilities, such as in a toss of a coin either head $H$ or tail $T$ can come, only $H$ or $T$ will appear in each trial, only two possibilities are there, then such a random experiment is called *a Bernoulli trial*. If a student is writing an examination and if the final result is to be recorded as either a pass $P$ or a failure $F$, then only one of the two possibilities can occur. Then attempting each examination is a Bernoulli trial. But if the final result is to be recorded as one of the grades $A, B, C, D$, then there are four possibilities in each outcome. Then this is not a Bernoulli trial. When a die is rolled once and if we are looking for either an odd number $O$ or an even number $E$, then there are only two possibilities. It is a Bernoulli trial. But if our aim is to see which number turns up then there are 6 possibilities, that is, one of the numbers $1, 2, 3, 4, 5, 6$ can appear or there are 6 possible items or possibilities in an outcome. It is *a multinomial trial*. It is not a Bernoulli trial.

In a Bernoulli trial, let the possible events in each outcome be denoted by $A$ and $B$. Then $A \cap B = \phi$ and $A \cup B = S =$ the sure event. Let the occurrence of $A$ be called "a success" and the occurrence of $B$ "a failure". Let the probability of $A$ be $p$. Then

$$P(A) = p, \quad P(B) = P(A^c) = 1 - P(A) = 1 - p = q$$

where $1-p$ is denoted by $q$. If a balanced or unbiased coin is tossed once and if getting a head is a success, then $P(A) = \frac{1}{2}$ and if the coin is not unbiased then $P(A) \neq \frac{1}{2}$. When a balanced die is rolled once and if $A$ is the event of getting the numbers 1 or 2, then $B$ is the event of getting 3 or 4 or 5 or 6. In this case,

$$P(A) = \frac{2}{6} = \frac{1}{3} \quad \text{and} \quad P(B) = \frac{4}{6} = \frac{2}{3}.$$

## 5.2 Bernoulli probability law

Let $x$ be the number of successes in one Bernoulli trial. Then $x = 1$ means a success with probability $p$ and $x = 0$ means a failure with probability $q$. These are the only two values $x$ can take with non-zero probabilities here. Then the probability function in this case, denoted by $f_1(x)$, can be written as

$$f_1(x) = \begin{cases} p^x q^{1-x}, & x = 0, 1 \\ 0, & \text{elsewhere.} \end{cases}$$

https://doi.org/10.1515/9783110562545-005

Here, $p$ is the only parameter. This is known as the *Bernoulli probability law*. The mean value $E(x)$, variance $\sigma^2 = E[x - E(x)]^2$ and the moment generating function $M(t)$ are the following: Since it is a discrete case, we sum up:

$$\begin{aligned} E(x) &= \sum_x x f_1(x) = 0 + (0)[p^0 q^{1-0}] + (1)[p^1 q^{1-1}] = p. && (5.1)\\ E(x^2) &= \sum_x x^2 f_1(x) = 0 + (0)^2[p^0 q^{1-0}] + (1)^2[p^1 q^{1-1}] = p. \\ \operatorname{Var}(x) &= E(x^2) - [E(x)]^2 = p - p^2 = p(1-p) = pq. && (5.2)\\ M(t) &= \sum_x \mathrm{e}^{tx} f_1(x) = 0 + \mathrm{e}^{t(0)}[p^0 q^{1-0}] + \mathrm{e}^{t(1)}[p^1 q^{1-1}] \\ &= q + p\mathrm{e}^t. && (5.3) \end{aligned}$$

We may note that this $M(t)$ can be expanded in power series and it can be differentiated also. We can obtain the integer moments by expansion or by differentiation:

$$M(t) = q + p\left[1 + \frac{t}{1!} + \frac{t^2}{2!} + \cdots\right].$$

Therefore, the coefficient of $\frac{t^1}{1!}$ is $p = E(x)$ and the coefficient of $\frac{t^2}{2!}$ is $p = E(x^2)$. Higher integer moments can also obtained from this series. Now, consider differentiation:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M(t)\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}[q + p\mathrm{e}^t]\Big|_{t=0} = [p\mathrm{e}^t]|_{t=0} = p. \\ \frac{\mathrm{d}^2}{\mathrm{d}t^2} M(t)\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{\mathrm{d}}{\mathrm{d}t} M(t)\right\}\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}[p\mathrm{e}^t]\Big|_{t=0} = p. \end{aligned}$$

**Example 5.1.** A gambler gets Rs 5 if the number 1 or 3 or 6 comes when a balanced die is rolled once and he loses Rs 5 if the number 2 or 4 or 5 comes. How much money can he expect to win in one trial of rolling this die once?

**Solution 5.1.** This is nothing but the expected value of a Bernoulli random variable with $p = \frac{1}{2}$ since the die is balanced. Hence

$$E(x) = 0 + (5)\left(\frac{1}{2}\right) + (-5)\left(\frac{1}{2}\right) = 0.$$

It is a fair game. Neither the gambler nor the gambling house has an upper hand, the expected gain or win is zero. Suppose that the die is loaded towards the number 2 or 4 or 5 and suppose that the probability of occurrence of any of these numbers is $\frac{2}{3}$ then the expected gain or win of the gambler is

$$E(x) = 0 + (5)\left(\frac{1}{3}\right) + (-5)\left(\frac{2}{3}\right) = -\frac{5}{3}$$

or the gambler is expected to lose Rs $\frac{5}{3}$ in each game or the gambling house has the upper hand.

## 5.3 Binomial probability law

Suppose that a Bernoulli trial is repeated $n$ times under identical situations or consider $n$ identical independent Bernoulli trials. Let $x$ be the total number of successes in these $n$ trials. Then $x$ can take the values $0, 1, 2, \dots, n$ with non-zero probabilities. In each trial, the probability of success is $p$. A success or failure in a trial does not depend upon what happened before. If $A_2$ is the event of getting a success in the second trial and if $A_1$ is the event of getting a failure in the first trial then

$$P(A_1) = q, \quad P(A_2|A_1) = P(A_2) = p, \quad P(A \cap B) = qp.$$

where $P(A \cap B)$ is the probability of getting the sequence "failure, success". Suppose that the first $x$ trials resulted in successes and the remaining $n - x$ trials resulted in failures. Then the probability of getting the sequence $SS \dots SFF \dots F$, where $S$ denotes a success and $F$ denotes a failure, is

$$pp \cdots pqq \cdots q = p^x q^{n-x}.$$

Suppose that the first three trials were failures, the next $x$ trials were successes and the remaining trials were failures then the probability for this sequence is $qqqp^x q \cdots q = p^x q^{n-x}$. For any given sequence, whichever way $S$ and $F$ appear, the probability is $p^x q^{n-x}$. How many such sequences are possible? It is $\binom{n}{x}$ or $\binom{n}{n-x}$. Hence if the probability of getting exactly $x$ successes in $n$ independent Bernoulli trails is denoted by $f_2(x)$ then

$$f_2(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, \dots, n \\ 0, & \text{elsewhere}, \ 0 < p < 1, \ q = 1 - p, \ n = 1, 2, \dots. \end{cases}$$

Note that $n$ and $p$ are parameters here. What is the total probability in this case?

$$\sum_x f_2(x) = 0 + \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (q + p)^n = 1^n = 1,$$

see equation (3.12) of Section 3.3 for the binomial sum. The total probability is 1 as can be expected when it is a probability law. Since $f_2(x)$ is the general term in a binomial expansion of $(q+p)^n$, this $f_2(x)$ is called a *Binomial probability law*. What are the mean value, variance and the moment generating function in this case?

$$\begin{aligned} E(x) &= \sum_x x f_2(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=1}^{n} x \binom{n}{x} p^x q^{n-x} \end{aligned}$$

since at $x = 0$, $x f_2(x) = 0$. For $x \neq 0$, we can cancel $x$ or divide numerator and denominator by $x$. We can rewrite

$$x \binom{n}{x} = x \left[ \frac{n!}{x!(n-x)!} \right] = \frac{n!}{(x-1)!(n-x)!}$$

since for $x \neq 0$ we can cancel $x$,

$$\frac{n!}{(x-1)!(n-x)!} = n\left[\frac{(n-1)!}{(x-1)!(n-x)!}\right] = n\binom{n-1}{x-1}$$

and

$$p^x q^{n-x} = p[p^{x-1}q^{(n-1)-(x-1)}].$$

Therefore,

$$\begin{aligned}\sum_x x\binom{n}{x}p^x q^{n-x} &= np\sum_{x=1}^{n}\binom{n-1}{x-1}p^{x-1}q^{(n-1)-(x-1)}\\ &= np\sum_{y=0}^{N}\binom{N}{y}p^y q^{N-y}, \quad y = x-1,\ N = n-1\\ &= np(q+p)^{n-1} = np \quad \text{since } q+p=1.\end{aligned}$$

Therefore, the mean value here is

$$E(x) = np. \tag{5.4}$$

For computing the variance, we can use the formula

$$\sigma^2 = E[x - E(x)]^2 = E(x^2) - [E(x)]^2.$$

Let us compute $E(x^2)$ first:

$$E(x^2) = \sum_x x^2 f_2(x) = 0 + \sum_{x=1}^{n} x^2 \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

since at $x = 0$ the right side is zero, and hence the sum starts from $x = 1$. We can cancel one $x$ with $x!$ giving $(x-1)!$ in the denominator. But still there is one more $x$ in the numerator. But we can see that since a factorial is sitting in the denominator it is easier to compute the factorial moments. Hence we may use the identity and write

$$x^2 \equiv x(x-1) + x.$$

Now, we can compute $E[x(x-1)]$, and $E(x)$ which is already computed.

$$\begin{aligned}E[x(x-1)] &= \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x q^{n-x}\\ &= \sum_{x=2}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x q^{n-x}\end{aligned}$$

since at $x = 0$, $x = 1$ the right side is zero. That is,

$$E[x(x-1)] = \sum_{x=2}^{n}\frac{n!}{(x-2)!(n-x)!}p^x q^{n-x}.$$

Now, take out $n(n-1)$ from $n!$, take out $p^2$ from $p^x$, rewrite $n-x=(n-2)-(x-2)$, substitute $y=x-2$, $N=n-2$. Then we have

$$E[x(x-1)] = n(n-1)p^2 \sum_{y=0}^{N} \binom{N}{y} p^y q^{N-y} = n(n-1)p^2(q+p)^N = n(n-1)p^2$$

since $(q+p)=1$. Therefore,

$$\begin{aligned}\sigma^2 &= E[x^2]-[E(x)]^2 = n(n-1)p^2+np-(np)^2\\ &= np-np^2 = np(1-p) = npq. \end{aligned} \tag{5.5}$$

Thus, the mean value in the binomial case $E(x)=np$ and the variance $\sigma^2=npq$. Let us compute the moment generating function:

$$\begin{aligned}M(t) &= \sum_x \mathrm{e}^{tx} f_2(x) = \sum_{x=0}^{n} \binom{n}{x} \mathrm{e}^{tx} p^x q^{n-x}\\ &= \sum_{x=0}^{n} \binom{n}{x} (p\mathrm{e}^t)^x q^{n-x} = (q+p\mathrm{e}^t)^n. \end{aligned} \tag{5.6}$$

For a binomial expansion, see Section 3.3. Note that, the integer moments can be easily obtained by differentiation of this moment generating function. Expansion can be done but it is more involved. Deriving the integer moments by differentiation is left to the students.

Before doing some examples, we will introduce two more standard probability functions associated with Bernoulli trials.

## 5.4 Geometric probability law

Again, let us consider independent Bernoulli trials where the probability of success in every trial is $p$ and $q=1-p$. Let us ask the question: what is the probability that the first success is at the $x$-th trial? The trial number is the random variable here. Let $f_3(x)$ denote this probability function. If the first success is at the $x$-th trial, then the first $x-1$ trials resulted in failures with probabilities $q$ each and then a success with probability $p$. Therefore,

$$f_3(x) = \begin{cases} q^{x-1}p, & x=1,2,\ldots\\ 0, & \text{elsewhere.}\end{cases}$$

Note that $p$ is the only parameter here. The successive terms here are $p, pq, pq^2, \ldots$ which are in geometric progression, and hence the law is called the *geometric probability law*. The graph is shown in Figure 5.1.

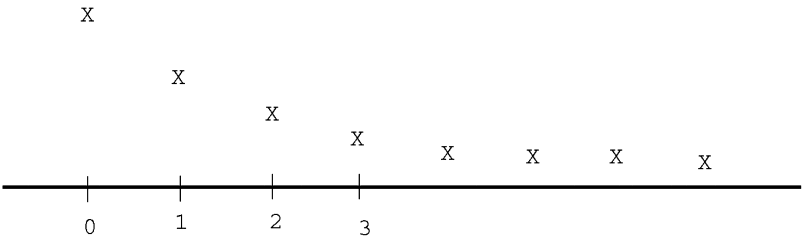


**Figure 5.1:** Geometric probability law.

Let us see the sum of the probabilities here. The total probability

$$
\begin{aligned}
\sum_x f_3(x) &= \sum_{x=1}^{\infty} q^{x-1}p = p\{1+q+q^2+\cdots\} \\
&= p(1-q)^{-1} \quad \text{see the binomial expansion in Section 3.3} \\
&= pp^{-1} = 1
\end{aligned}
$$

as can be expected. Let us compute the mean value $E(x)$, variance $\sigma^2$ and the moment generating function $M(t)$ for this geometric probability law:

$$
\begin{aligned}
E(x) &= \sum_x xf_3(x) = \sum_{x=1}^{\infty} xq^{x-1}p = p\{1+2q+3q^2+\cdots\} \\
&= p(1-q)^{-2}, \quad 0<q<1 \quad \text{see Section 3.3 for the binomial sum} \\
&= pp^{-2} = \frac{1}{p}, \quad 0<p<1. \qquad (5.7)
\end{aligned}
$$

We can also derive this by using the following procedure:

$$
\begin{aligned}
E(x) &= \sum_{x=1}^{\infty} xq^{x-1}p = p\sum_{x=1}^{\infty}\left[\frac{\mathrm{d}}{\mathrm{d}q}q^x\right] \\
&= p\frac{\mathrm{d}}{\mathrm{d}q}\sum_{x=1}^{\infty} q^x \\
&= p\frac{\mathrm{d}}{\mathrm{d}q}[q+q^1+\cdots] = p\frac{\mathrm{d}}{\mathrm{d}q}[q(1-q)^{-1}] \\
&= p[(1-q)^{-1}+q(1-q)^{-2}] = 1+\frac{q}{p} \\
&= \frac{q+p}{p} = \frac{1}{p}.
\end{aligned}
$$

For computing $E(x^2)$, we may observe the following:

$$
x^2q^{x-1}p = p\frac{\mathrm{d}}{\mathrm{d}q}\left[q\frac{\mathrm{d}}{\mathrm{d}q}q^x\right].
$$

Hence

$$
\begin{aligned}
E(x^2) &= \sum_{x=1}^{\infty} x^2 q^{x-1} p = p \sum_{x=1}^{\infty} \left\{ \frac{\mathrm{d}}{\mathrm{d}q} \left[ q \frac{\mathrm{d}}{\mathrm{d}q} q^x \right] \right\} \\
&= \frac{\mathrm{d}}{\mathrm{d}q} q \frac{\mathrm{d}}{\mathrm{d}q} \sum_{x=1}^{\infty} q^x = \frac{\mathrm{d}}{\mathrm{d}q} q \frac{\mathrm{d}}{\mathrm{d}q} [q(1-q)^{-1}] \\
&= p \frac{\mathrm{d}}{\mathrm{d}q} [q(1-q)^{-1} + q^2(1-q)^{-2}] = p(1-q)^{-1} \\
&\quad + p[3q(1-q)^{-2} + 2q^2(1-q)^{-3}] = 1 + \frac{3q}{p} + 2\frac{q^2}{p^2}. 
\end{aligned} \tag{5.8}
$$

We can also obtain this from the moment generating function:

$$
\begin{aligned}
M(t) &= \sum_{x=1}^{\infty} \mathrm{e}^{tx} p q^{x-1} = p\{\mathrm{e}^t + q\mathrm{e}^{2t} + q^2 \mathrm{e}^{3t} + \cdots\} \\
&= p\mathrm{e}^t (1 - q\mathrm{e}^t)^{-1} \quad \text{for } q\mathrm{e}^t < 1.
\end{aligned} \tag{5.9}
$$

Differentiating $M(t)$ with respect to $t$ and then evaluating at $t = 0$, we have

$$
\begin{aligned}
E(x) &= \frac{\mathrm{d}}{\mathrm{d}t} M(t)\Big|_{t=0} = p \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{e}^t (1 - q\mathrm{e}^t)^{-1}\Big|_{t=0} \\
&= \{p\mathrm{e}^t(1-q\mathrm{e}^t)^{-1} + p\mathrm{e}^t(1-q\mathrm{e}^t)^{-2} q\mathrm{e}^t\}|_{t=0} \\
&= pp^{-1} + pp^{-2}q = 1 + \frac{q}{p} = \frac{1}{p}. \\
E(x^2) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2} M(t)\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \{p\mathrm{e}^t(1-q\mathrm{e}^t)^{-1} + pq\mathrm{e}^{2t}(1-q\mathrm{e}^t)^{-2}\}_{t=0} \\
&= \{p\mathrm{e}^t(1-q\mathrm{e}^t)^{-1} + p\mathrm{e}^t q\mathrm{e}^t(1-q\mathrm{e}^t)^{-2} + 2pq\mathrm{e}^{2t}(1-q\mathrm{e}^t)^{-2}\}_{t=0} \\
&= 1 + \frac{q}{p} + 2\frac{q}{p} + 2\frac{q^2}{p^2} = 1 + 3\frac{q}{p} + 2\frac{q^2}{p^2}.
\end{aligned}
$$

## 5.5 Negative binomial probability law

Again, let us consider independent Bernoulli trials with the probability of success $p$ remaining the same. Let us ask the question: what is the probability that the $x$-th trial will result in the $k$-th success for a fixed $k$, something like the 10-th trial resulting in the 7-th success? Let this probability be denoted by $f_4(x)$. The $k$-th success at the $x$-th trial means that there were $k-1$ successes in the first $x-1$ trials; the successes could have occurred any time in any sequence but a total of $k-1$ of them. This is given by a binomial probability law of $x-1$ trials and $k-1$ successes. The next trial should be a success, then one has the $x$-th trial resulting in the $k$-th success. Hence

$$
f_4(x) = \begin{cases} [\binom{x-1}{k-1} p^{k-1} q^{(x-1)-(k-1)}]p = \binom{x-1}{k-1} p^k q^{x-k}, & x = k, k+1, \ldots \\ 0, & \text{elsewhere.} \end{cases}
$$

Note that one has to have at least $k$ trials to get $k$ successes, and hence $x$ varies from $x = k$ onward. Here, $p$ and $k$ are parameters. What is the total probability here?

$$\begin{aligned}\sum_x f_4(x) &= \sum_{x=k}^{\infty}\binom{x-1}{k-1}p^k q^{x-k}\\ &= \sum_{x=k}^{\infty}\binom{x-1}{x-k}p^k q^{x-k} \quad \text{since } \binom{n}{r} = \binom{n}{n-r}\\ &= p^k\left\{\binom{k-1}{0} + \binom{k}{1}q + \binom{k+1}{2}q^2 + \cdots\right\}\\ &= p^k\left\{1 + k\frac{q}{1!} + k(k+1)\frac{q^2}{2!} + \cdots\right\}\\ &= p^k(1-q)^{-k} = p^k p^{-k} = 1\end{aligned}$$

as can be expected. Since $f_4(x)$ is the general term in a binomial expansion with a negative exponent, this probability is known as the *negative binomial probability function*.

Naturally, when $k = 1$ we have the geometric probability law. Thus the geometric probability law is a particular case of the negative binomial probability law. Let us compute $E(x), E(x^2)$ and the moment generating function. The moment generating function:

$$\begin{aligned}M(t) &= p^k\sum_{x=k}^{\infty}\binom{x-1}{k-1}q^{x-k}\mathrm{e}^{tx} = p^k\sum_{x=k}^{\infty}\binom{x-1}{x-k}q^{x-k}\mathrm{e}^{tx}\\ &= p^k\left\{\binom{k-1}{0}\mathrm{e}^{kt} + \binom{k}{1}\mathrm{e}^{(k+1)t}q + \cdots\right\}\\ &= p^k\mathrm{e}^{kt}\left\{1 + k\frac{q\mathrm{e}^t}{1!} + k(k+1)\frac{(q\mathrm{e}^t)^2}{2!}q + \cdots\right\}\\ &= p^k\mathrm{e}^{kt}(1-q\mathrm{e}^t)^{-k} \quad \text{for } q\mathrm{e}^t < 1. \end{aligned} \tag{5.10}$$

This is a differentiable function. Hence

$$\begin{aligned}E(x) &= \frac{\mathrm{d}}{\mathrm{d}t}M(t)\Big|_{t=0} = p^k\{k\mathrm{e}^{kt}(1-q\mathrm{e}^t)^{-k} + k\mathrm{e}^{kt}(1-q\mathrm{e}^t)^{-k-1}q\mathrm{e}^t\}|_{t=0}\\ &= kp^k\{p^{-k} + qp^{-k-1}\} = k\left\{1 + \frac{q}{p}\right\} = \frac{k}{p}. \end{aligned} \tag{5.11}$$

$$\begin{aligned}E(x^2) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}M(t)|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\mathrm{d}}{\mathrm{d}t}M(t)\right]\Big|_{t=0}\\ &= kp^k\{k\mathrm{e}^{kt}(1-q\mathrm{e}^t)^{-k} + kq\mathrm{e}^{(k+1)t}(1-q\mathrm{e}^t)^{-(k+1)}\\ &+ q(k+1)\mathrm{e}^{(k+1)t}(1-q\mathrm{e}^t)^{-(k+1)} + (k+1)q^2\mathrm{e}^{(k+1)t}(1-q\mathrm{e}^t)^{-(k+1)}\}|_{t=0}\\ &= k^2 + \frac{q}{p}(2k^2+k) + \frac{q^2}{p^2}k(k+1). \end{aligned} \tag{5.12}$$

We have given four important probability functions, namely the Bernoulli probability law, the binomial probability law, the geometric probability law and the negative binomial probability law, connected with independent Bernoulli trials. All these are frequently used in statistical literature and probability theory.

**Example 5.2.** In a multiple choice examination, each question is supplied with 3 possible answers of which one is the correct answer to the question. A student, who does not know any of the correct answers, is doing the examination by selecting answers at random. What is the probability that (a) out of the 5 questions answered the student has (i) exactly 3 correct answers, (ii) at least 3 correct answers; (b) (i) the third question answered is the first correct answer; (ii) at least 3 questions out of the 10 questions to be answered are needed to get the first correct answer; (c) the 5th question answered is the 3rd correct answer; (ii) at least 4 questions, out of the 10 questions answered, are needed to get the 3rd correct answer.

**Solution 5.2.** Attempting to answer the questions by selecting answers at random can be taken as independent Bernoulli trials with probability of success $\frac{1}{3}$ because out of the 3 possible answers only one is the correct answer to the question. In our notation, $p = \frac{1}{3}, q = \frac{2}{3}$. For (a), it is a binomial situation with $n = 5$. In (i), we need $\Pr\{x = 3\}$. From the binomial probability law,

$$\begin{aligned}\Pr\{x = 3\} &= \binom{5}{3}p^3q^{5-3} = \binom{5}{2}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 \\ &= \frac{(5)(4)}{2!}\frac{4}{3^5} = \frac{40}{3^5}.\end{aligned}$$

When computing the number of combinations, always use the definition:

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!}.$$

It will be foolish to use all factorials because it will involve unnecessary computations and often big factorials cannot even be handled by computers. In (a)(ii), we need

$$\Pr\{x = 3\} + \Pr\{x = 4\} + \Pr\{x = 5\} = L \quad \text{say.}$$

Then

$$L = \binom{5}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 + \binom{5}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right) + \binom{5}{5}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^0.$$

But

$$\binom{5}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 = \binom{5}{2}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 = \frac{(5)(4)}{2!}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 = \frac{40}{3^5}.$$

$$\binom{5}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^1 = \binom{5}{1}\frac{2}{3^5} = \frac{10}{3^5}.$$

$$\binom{5}{5}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^0 = \frac{1}{3^5}.$$

The total is

$$\frac{51}{3^5} = \frac{17}{81}.$$

For (b), it is a geometric probability law. In (i), $x = 3$ and the answer is

$$q^{3-1}p = \left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) = \frac{4}{27}.$$

For (ii) in (b), we need the sum:

$$\begin{aligned}\sum_{x=3}^{10} q^{x-1}p &= p[q^2 + q^3 + \cdots + q^9] = pq^2[1 + q + \cdots + q^7]\\ &= pq^2\frac{(1-q^8)}{1-q} = q^2(1-q^8) = \left(\frac{2}{3}\right)^2\left[1 - \left(\frac{2}{3}\right)^8\right].\end{aligned}$$

For (c), it is a negative binomial situation with $k = 3$. In (i), $x = 5$, $k = 3$, $p = \frac{1}{3}$, $q = \frac{2}{3}$ and the answer is

$$\binom{x-1}{k-1}p^k q^{x-k} = \binom{4}{2}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2 = \frac{24}{3^5}.$$

In (ii), we need the sum:

$$\begin{aligned}&\sum_{x=4}^{10}\binom{x-1}{k-1}p^k q^{x-k}\\ &= \left(\frac{1}{3}\right)^3\left\{\binom{3}{2}\left(\frac{2}{3}\right)^1 + \cdots + \binom{9}{2}\left(\frac{2}{3}\right)^7\right\}\\ &= \frac{1}{27}\left\{3\left(\frac{2}{3}\right) + 6\left(\frac{2}{3}\right)^2 + 10\left(\frac{2}{3}\right)^3\right.\\ &\quad\left. + 15\left(\frac{2}{3}\right)^4 + 21\left(\frac{2}{3}\right)^5 + 28\left(\frac{2}{3}\right)^6 + 36\left(\frac{2}{3}\right)^7\right\}.\end{aligned}$$

**Example 5.3.** An experiment on rabbits is designed by taking $N = 20$ identical rabbits. But rabbits start dying out before the experiment is completed. Let $n$ be the effective final number. This sample size $n$ has become a random quantity. $n$ could be zero (all rabbits died out), $n$ could be $1, 2, \ldots$ and could be $N = 20$ (no rabbit died). Let 0.1 be the probability of a rabbit dying and suppose that this probability is the same for all rabbits. Construct the probability law for $n$.

**Solution 5.3.** Here, $n$ satisfies all the conditions for a binomial random variable with probability of success $p = 0.1$ and the number of trials $N = 20$. Hence the probability law for $n$ is, denoted by $P_n$,

$$P_n = \binom{N}{n}p^n q^{N-n} = \binom{20}{n}(0.1)^n(0.9)^{20-n}, \quad n = 0, 1, \ldots, 20.$$

**Example 5.4.** A lunch counter in an office building caters to people working in the building. Apart from regular lunches, the counter operator makes an exotic lunch packet every day. If the exotic packet is not sold on that day, then it is a total loss; it cannot be stored or used again. From past experience, the operator found the daily demand for this exotic packet as follows. It costs Rs 5 to make and she can sell it for Rs 10.

$$(\text{Demand}, \text{probability}) = (0, 0.1), (1, 0.2), (2, 0.2), (3, 0.3), (4, 0.1), (5, 0.1).$$

There is no demand for more than 5. That is, the operator can sell, for example, the 3rd packet if the demand is for 3 or more. How many packets she should make so that her expected profit is a maximum?

**Solution 5.4.** If she makes 1 packet, the probability that it can be sold is that the demand on that day is for 1 or more packets. The probability for this is $0.1 + 0.2 + 0.2 + 0.3 + 0.1 + 0.1 = 0.9$. It costs Rs 5 and the expected revenue is Rs $10 \times 0.9 =$ Rs 9 and the expected profit is Rs 4.

[As an expected value of a random variable, this is the following: Let $y$ be her gain or loss on a single packet. Then $y$ takes the value +5 (profit) with probability 0.9 [if the demand on that day is for one or more] and $y$ takes the value −5 (loss) with probability 0.1 [if the demand on that day is for less than one or zero]. Then the expected value of $y$, $E(y) = 5(0.9) - 5(0.1) = 4$. Thus she has an expected profit of Rs 4.]

If she makes 2 packets, then the cost is $2 \times 5 = 10$. She can sell the first packet with probability 0.9 or make the expected revenue of Rs 9. She can sell the second packet if there is demand for 2 or more or with the probability $0.2 + 0.3 + 0.1 + 0.1 = 0.7$ and make the expected revenue $10 \times 0.7 = 7$. Thus the total expected revenue is $9 + 7 = 16$ and the expected profit is $16 - 10 = 6$.

If she makes 3 packets, then she can sell the third packet with probability $0.3 + 0.1 + 0.1 = 0.5$ and the expected revenue is $10 \times 0.5 = 5$. Thus the expected profit is $9 + 7 + 5 = 21 - 15(= 5 \times 3) = 6$.

If she makes 4 packets, then she can sell the 4th packet with probability $0.1 + 0.1 = 0.2$ and the expected revenue is $10 \times 0.2 = 2$ and the expected profit is $9 + 7 + 5 + 2 = 23 - 4 \times 5 = 3$.

If she makes 5 packets, then she can sell the 5th one with probability 0.1 and the expected revenue is Rs 1, the total cost is $5 \times 5 = 25$ and, therefore, there is an expected loss of Rs 1. Hence she should make either 2 or 3 packets to maximize her profit.

## 5.6 Poisson probability law

We will derive this probability law as a limiting form of the binomial as well as a process satisfying some conditions. This law is named after its inventor, S. Poisson, a French mathematician. Consider a binomial probability law where the number of trials $n$ is very large and the probability of success $p$ is very small but $np = \lambda$ (Greek letter lambda), a finite quantity. This situation can be called a situation of rare events, the number of trials is very large and the probability of success in each trial is very small, something like a lightning strike, earthquake at a particular place, traffic accidents on a certain stretch of a highway and so on. Let us see what happens if $n \to \infty$, $p \to 0$, $np = \lambda$. Since we are assuming $\lambda$ to be a finite quantity, we may replace one of $n$ or $p$ in terms of $\lambda$. Let us substitute $p = \frac{\lambda}{n}$. Then

$$\begin{aligned}\lim_{n\to\infty}\binom{n}{x}p^x &= \frac{1}{x!}\lim_{n\to\infty} n(n-1)\cdots(n-x+1)\left(\frac{\lambda}{n}\right)^x\\ &= \frac{\lambda^x}{x!}\lim_{n\to\infty}\frac{n}{n}\frac{(n-1)}{n}\cdots\frac{(n-x+1)}{n}\\ &= \frac{\lambda^x}{x!}1\times\lim_{n\to\infty}\left[1-\frac{1}{n}\right]\times\cdots\times\lim_{n\to\infty}\left[1-\frac{x-1}{n}\right]\\ &= \frac{\lambda^x}{x!}\end{aligned}$$

since $x$ is finite, there are only a finite number of factors, and we can use the formula that the limit of a finite number of products is the product of the limits and each factor here goes to 1. [If it involved an infinite number of factors, then we could not have taken the limits on each factor.] Now let us examine the factor:

$$q^{n-x} = (1-p)^n(1-p)^{-x} = \left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}.$$

But

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \mathrm{e}^{-\lambda}$$

from the definition of e, and

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1$$

since the exponent $-x$ is finite. Therefore,

$$\begin{aligned}\lim_{n\to\infty,p\to 0,np=\lambda} f_2(x) &= \lim_{n\to\infty,p\to 0,np=\lambda}\binom{n}{x}p^x q^{n-x}\\ &= \begin{cases}\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, & x = 0,1,\dots, \lambda > 0\\ 0, & \text{elsewhere.}\end{cases}\end{aligned}$$

Let us call the right side as $f_5(x)$. Let us see whether $f_5(x)$ is a probability function, since we have done a limiting process on a binomial probability function. If you take some sort of limits on a probability function, the limiting form need not remain as a probability function. The total of $f_5(x)$ is given by

$$\sum_x f_5(x) = \sum_{x=0}^{\infty}\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = 1.$$

Hence $f_5(x)$ is a probability function and it is known as the Poisson probability law. $\lambda$ is a parameter here.

Let us evaluate $E(x)$, variance and the moment generating function for the Poisson probability law:

$$E(x) = \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!}$$

since at $x = 0$ the right side is zero. Now we take out one lambda, cancel one $x$, $\frac{x}{x!} = \frac{1}{(x-1)!}$ when $x \neq 0$. Then we have

$$\begin{aligned} E(x) &= \lambda\mathrm{e}^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda\mathrm{e}^{-\lambda}\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right] = \lambda\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = \lambda. \end{aligned} \tag{5.13}$$

Thus, the mean value in the Poisson case is the parameter $\lambda$ sitting in the probability function. Since $x!$ is sitting in the denominator, for computing $E(x^2)$, we will go through the factorial moments or consider the identity

$$x^2 = x(x-1) + x$$

and proceed to evaluate $E[x(x-1)]$. This procedure has already been done several times before. We cancel $x(x-1)$ from $x!$ since at $x = 0, 1$ the function on the right will be zeros, and thus $x$ only goes from $x = 2$ to infinity in the sum. Then we take out $\lambda^2$. That is,

$$\begin{aligned} E[x(x-1)] &= \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} \\ &= \lambda^2\mathrm{e}^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} = \lambda^2\mathrm{e}^{-\lambda}\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right] \\ &= \lambda^2\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = \lambda^2. \end{aligned}$$

Then the variance

$$\begin{aligned} \sigma^2 &= E[x - E(x)]^2 = E[x(x-1)] + E[x] - [E(x)]^2 \\ &= \lambda^2 + \lambda - [\lambda]^2 = \lambda. \end{aligned} \tag{5.14}$$

Thus it is interesting to see that the mean value and the variance are equal to $\lambda$ in the Poisson case. But this is not a unique property or a characterizing property of the Poisson distribution. There are also other distributions satisfying this property that the mean value is equal to the variance.

Let us compute the moment generating function in this case

$$M(t) = E[\mathrm{e}^{tx}] = \sum_{x=0}^{\infty}\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}\mathrm{e}^{tx}$$

$$
\begin{aligned}
&= \mathrm{e}^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda \mathrm{e}^t)^x}{x!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda \mathrm{e}^t} \\
&= \mathrm{e}^{\lambda[\mathrm{e}^t-1]}. \qquad (5.15)
\end{aligned}
$$

The last sum is obtained by observing that it is an exponential series with $\lambda \mathrm{e}^t$ in the exponent.

**Example 5.5.** The monthly traffic accidents on a stretch of a particular highway is seen to be Poisson distributed with expected number of accidents 5. Four months are selected at random. What is the probability (i) in all four months the number of accidents is two or more per month; (ii) in at least one of the months the number of accidents is exactly 3; (iii) the first three months had no accidents and the fourth month had two accidents.

**Solution 5.5.** Let $x$ be the number of monthly accidents on that stretch of the highway. Then the probability function is given as, denoted by $P(x)$,

$$
P(x) = \begin{cases} \frac{5^x}{x!}\mathrm{e}^{-5}, & x = 0, 1, \dots \\ 0, & \text{elsewhere.} \end{cases}
$$

The parameter $\lambda = 5$ because it is given that the mean value is 5. The probability that the number of accidents in a month is 2 or more, denoted by $p_1$, is given by

$$
p_1 = \sum_{x=2}^{\infty}\frac{5^x}{x!}\mathrm{e}^{-5} = 1 - \sum_{x=0}^{1}\frac{5^x}{x!}\mathrm{e}^{-5}
$$

since the total probability is 1. That is,

$$
p_1 = 1 - \frac{5^0}{0!}\mathrm{e}^{-5} - \frac{5^1}{1!}\mathrm{e}^{-5} = 1 - \mathrm{e}^{-5}[1+5] = 1 - 6\mathrm{e}^{-5}.
$$

The answer to (i) is then $p_1^4$.

The probability for the number of accidents in a month is exactly 3, denoted by $p_2$, is given by

$$
p_2 = \frac{5^3}{3!}\mathrm{e}^{-5} = \frac{125}{6}\mathrm{e}^{-5}.
$$

In (ii), it is a binomial situation with the number of trials $n = 4$ and probability of success is $p_2$. Hence the answer to (ii) is

$$
\sum_{x=1}^{4}\binom{4}{x}p_2^x(1-p_2)^{4-x} = 1 - \binom{4}{0}p_2^0(1-p_2)^{4-0} = 1 - \left(1 - \frac{125}{6}\mathrm{e}^{-5}\right)^4.
$$

Probability of having no accidents is

$$
P(0) = \frac{5^0}{0!}\mathrm{e}^{-5} = \mathrm{e}^{-5}.
$$

Probability of having exactly 2 accidents is $\frac{5^2}{2!}e^{-5} = \frac{25}{2}e^{-5}$. Hence the answer to (iii) is

$$[e^{-5}]^3\left[\frac{25}{2}e^{-5}\right] = \frac{25}{2}e^{-20}.$$

### 5.6.1 Poisson probability law from a process

Consider an event taking place over time, such as the arrival of persons into a queue at a checkout counter, arrival of cars into a service station for service, arrival of telephone calls into a phone switchboard, floods in a river during rainy season, earthquakes over the years at a particular place, eruption of a certain volcano over time and so on. Let us assume that our event satisfies the following conditions:

(i) The occurrence of the event from time $t$ to $t+\Delta t$, that is in the interval $[t, t+\Delta t]$, where $\Delta t$ is a small increment in $t$, is proportional to the length of the interval or it is $\alpha\Delta t$, where $\alpha$ is a constant. Here, $\Delta$ (Greek capital letter delta) is not used as a product. $\Delta t$ is a notation standing for a small increment in $t$.

(ii) The probability of more than one occurrence of this event in $[t, t+\Delta t]$ is negligibly small and we take it as zero for all practical purposes, or it is assumed that the interval can be subdivided to the extent that probability of more than one occurrence in this small interval is negligible.

(iii) The occurrence or non-occurrence of this event in $[t, t+\Delta t]$ does not depend upon what happened in the interval $[0, t]$ where 0 indicates the start of the observation period. An illustration is given in Figure 5.2. If the event under observation is a flood in a river during the rainy season, then start of the rainy season is taken as zero and time may be counted in days or hours or in any convenient unit.

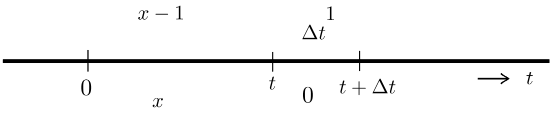


**Figure 5.2:** An event taking place over time.

Under the conditions (i), (ii), (iii) above, what is the probability of getting exactly $x$ occurrences of this event in time $[0, t]$? This probability function depends upon $x$ and the time $t$ and let us denote it by $f(x, t)$. Then the interpretations are the following:

$f(x, t+\Delta t)$ = the probability of getting exactly $x$ occurrences of the event in time $[0, t+\Delta t]$;

$f(x-1, t)$ = the probability of getting exactly $x-1$ occurrences in time $[0, t]$.

Exactly $x$ occurrences in the interval 0 to $t+\Delta t$ can happen in two mutually exclusive ways of (a) exactly $x-1$ occurrences in time 0 to $t$ or in the interval $[0, t]$, and one occurrence from $t$ to $t+\Delta t$ or in the interval $[t,t+\Delta t]$ [probability for one occurrence is $\alpha\Delta t$], or (b) exactly $x$ occurrences in the interval $[0,t]$ and no occurrence in the interval $[t,t+\Delta t]$ [probability for no occurrence is $1-\alpha\Delta t$]. Therefore, from the total probability law,

$$f(x,t+\Delta t)=f(x-1,t)[\alpha\Delta t]+f(x,t)[1-\alpha\Delta t].$$

We can rearrange the terms and write

$$\frac{f(x,t+\Delta t)-f(x,t)}{\Delta t}=\alpha[f(x-1,t)-f(x,t)].$$

Taking the limit when $\Delta t\to 0$, we get a differential equation in $t$ or a partial differential equation in $t$. That is,

$$\frac{\partial}{\partial t}f(x,t)=\alpha[f(x-1,t)-f(x,t)]. \tag{5.16}$$

Here, (5.16) is a differential equation in $t$ whereas it is a difference equation in $x$. We have to solve this difference-differential equation to obtain $f(x,t)$. This can be solved successively by taking values for $x=0$ solving the equation for $t$, then $x=1$ solving the equation for $t$, and so on. The final result will be the following:

$$f(x,t)=\begin{cases}\frac{(\alpha t)^x}{x!}\mathrm{e}^{-\alpha t}, & \alpha>0,\ 0\le t<\infty,\ x=0,1,\dots\\ 0, & \text{elsewhere,}\end{cases}$$

or it is a Poisson probability law with the parameter $\lambda=\alpha t$.

**Example 5.6.** Telephone calls are coming to an office switchboard at the rate of 0.5 calls per minute, time being measured in minutes. What is the probability that (a) in a 10-minute interval (i) there are exactly 2 calls; (ii) there is no call; (iii) at least one call; (b) if two 10-minute intervals are taken at random then (i) in both intervals there are no calls; (ii) in one of the intervals there are exactly 2 calls?

**Solution 5.6.** We will assume that these telephone calls obey the conditions for Poisson arrivals of calls or the Poisson probability law is a good model. We are given $\alpha=0.5$. In (a) $t=10$, then $\lambda=10\times 0.5=5$ and the probability law $P(x)$ is

$$P(x)=\begin{cases}\frac{5^x}{x!}\mathrm{e}^{-5}, & x=0,1,\dots\\ 0, & \text{elsewhere.}\end{cases}$$

In (a)(i), we need the probability $\Pr\{x=2\}$.

$$\Pr\{x=2\}=\frac{5^2}{2!}\mathrm{e}^{-5}=\frac{25}{2}\mathrm{e}^{-5}.$$

In (a)(ii), we need the probability $\Pr\{x = 0\}$.

$$\Pr\{x = 0\} = \frac{5^0}{0!}\mathrm{e}^{-5} = \mathrm{e}^{-5}.$$

In (a)(iii), we need $\Pr\{x \geq 1\}$.

$$\Pr\{x \geq 1\} = 1 - \Pr\{x = 0\} = 1 - \mathrm{e}^{-5}.$$

In (b)(i), it is a case of two Bernoulli trials where the probability of success $p_1$ is the probability of having no arrivals in a 10-minute interval or $p_1 = \mathrm{e}^{-5}$. We want both trials to result in successes, and hence the answer is

$$p_1^2 = [\mathrm{e}^{-5}]^2 = \mathrm{e}^{-10}.$$

In (b)(ii), we have two Bernoulli trials and the probability of success in each trial is $p_2$ where $p_2$ is the probability of having exactly 2 calls in a 10-minute interval. Then

$$p_2 = \frac{5^2}{2!}\mathrm{e}^{-5} = \frac{25}{2}\mathrm{e}^{-5}.$$

In (b)(ii), we need the probability of getting exactly one success in two Bernoulli trials. Then it is given by

$$\binom{2}{1}p_2^1(1-p_2)^{2-1} = 2\left[\frac{25}{2}\mathrm{e}^{-5}\right]\left[1 - \frac{25}{2}\mathrm{e}^{-5}\right].$$

Another probability law which is frequently used in probability and statistics is the discrete hypergeometric law.

## 5.7 Discrete hypergeometric probability law

Let us consider a box containing two types of objects: one type is of $a$ in number, which we will call these a-type objects, and the other type is $b$ in number, which we will call these b-type objects. As an example, we can consider a box containing red and green marbles and suppose that there are 10 green and 8 red marbles then we may consider $a = 10$ and $b = 8$ or vice versa. Suppose that a subset of $n$ items is taken at random from this set of $a + b$ objects. When we say “at random” it means that every subset of $n$ has the same chance of being taken or each subset gets a probability of $\frac{1}{\binom{a+b}{n}}$ because there are $\binom{a+b}{n}$ such subsets possible. This can also be done by taking one by one, at random, without replacement. Both will lead to the same probabilities.

In this experiment, what is the probability that the sample of $n$ items contains $x$ of a-type and $n - x$ of b-type objects. Let this probability be denoted by $f_6(x)$. Note that $x$ of a-type can only come from a-type objects and this can be done in $\binom{a}{x}$ ways. For each such selection of a-type objects, we can select $n-x$ b-type objects in $\binom{b}{n-x}$ ways. Therefore, the number of choices favorable to the event is $\binom{a}{x}\binom{b}{n-x}$. Hence

$$f_6(x) = \begin{cases} \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}}, & x = 0,1,\dots,n \text{ or } a;\ a,b = 1,2,\dots,\ n = 1,2,\dots \\ 0, & \text{elsewhere.} \end{cases}$$

This is known as the *discrete hypergeometric probability law*. $a, b, n$ are parameters here.

First, let us check to see the sum:

$$\sum_x f_6(x) = \frac{\sum_{x=0}^{n,a}\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}} = 1$$

because, from Section 3.3 we have

$$\sum_{x=0}^{n,a}\binom{a}{x}\binom{b}{n-x} = \binom{a+b}{n}. \tag{5.17}$$

Thus the total probability is 1 as can be expected. What are the mean values and variance in this case?

$$E(x) = \frac{1}{\binom{a+b}{n}}\sum_{x=0}^{n,a} x\binom{a}{x}\binom{b}{n-x}.$$

When $x = 0$, the right side is zero, and hence the sum starts only at $x = 1$. Then one may cancel one $x$ from the $x!$. That is,

$$x\binom{a}{x} = x\frac{a!}{x!(a-x)!} = a\frac{(a-1)!}{(x-1)!((a-1)-(x-1))!} = a\binom{a-1}{x-1}.$$

Hence, taking the sum by putting $y = x-1$ so that $y$ goes from 0, and by using (5.17), we have

$$a\sum_{y=0}^{n,a}\binom{a-1}{y}\binom{b}{n-1-y} = a\binom{a+b-1}{n-1}.$$

Now, dividing by $\binom{a+b}{n}$ and simplifying we get

$$E(x) = \frac{na}{a+b}. \tag{5.18}$$

By using the same steps, the second factorial moment is given by

$$E[x(x-1)] = \frac{n(n-1)a(a-1)}{(a+b)(a+b-1)}. \tag{5.19}$$

Now, variance is available from the formula

$$\begin{aligned} \mathrm{Var}(x) &= E[x(x-1)] + E(x) - [E(x)]^2 \\ &= \frac{n(n-1)a(a-1)}{(a+b)(a+b-1)} + \frac{na}{(a+b)} - \frac{n^2a^2}{(a+b)^2}. \end{aligned} \tag{5.20}$$

**Example 5.7.** From a set of 5 women and 8 men a committee of 3 is selected at random [this means all such subsets of 3 are given equal chances of being selected]. What is the probability that the committee consists of (i) no woman; (ii) at least two women; (iii) all women?

**Solution 5.7.** Let $x$ be the number of women in the committee. Then $x$ is distributed according to a discrete hypergeometric probability law. In (i), we need $\Pr\{x = 0\}$:

$$\Pr\{x = 0\} = \frac{\binom{5}{0}\binom{8}{3}}{\binom{13}{3}} = \frac{(8)(7)(6)}{(13)(12)(11)} = \frac{28}{143}.$$

In (ii), we need $\Pr\{x = 2 \text{ or } 3\} = \Pr\{x = 2\} + \Pr\{x = 3\}$. In (iii), we need $\Pr\{x = 3\}$. Let us compute these two probabilities:

$$\Pr\{x = 3\} = \frac{\binom{5}{3}\binom{8}{0}}{\binom{13}{3}} = \frac{(5)(4)(3)}{(13)(12)(11)} = \frac{5}{143}.$$

$$\Pr\{x = 2\} = \frac{\binom{5}{2}\binom{8}{1}}{\binom{13}{3}} = \frac{(5)(4)}{(1)(2)}(8)\frac{(1)(2)(3)}{(13)(12)(11)} = \frac{40}{143}.$$

Hence the answer in (ii) is $\frac{40}{143} + \frac{5}{143} = \frac{45}{143}$.

## 5.8 Other commonly used discrete distributions

Here, we list some other commonly used discrete distributions. Only the non-zero part of the probability function is given and it should be understood that the function is zero otherwise. In some of the probability functions, gamma functions, $\Gamma(\cdot)$ ($\Gamma$ is the Greek capital letter gamma) and beta functions, $B(\cdot,\cdot)$ ($B$ is the Greek capital letter beta) appear. Hence we will list the integral representations of these functions here. Their definitions will be given in the next chapter. Only these integral representations will be sufficient to do problems on the following probability functions where gamma and beta functions appear.

The integral representation for a gamma function, $\Gamma(\alpha)$, is the following:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x, \quad \Re(\alpha) > 0. \tag{5.21}$$

For the integral to converge, we need the condition that the real part of $\alpha$ (alpha) is positive, $\Re(\cdot)$ means the real part of $(\cdot)$.

**Note 5.1.** Usually in statistical problems the parameters are real but the integrals will exist in the complex domain also, and hence the conditions are written in terms of real parts of the complex parameters.

The beta function, $B(\alpha,\beta)$, can be written in terms of the gamma function. In the following, we give the connection to gamma function and integral representations for beta functions:

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x, \tag{5.22}$$

$$=\int_0^\infty y^{\alpha-1}(1+y)^{-(\alpha+\beta)}, \quad \Re(\alpha)>0, \quad \Re(\beta)>0. \tag{5.23}$$

For the convergence of the integrals in (5.22) and (5.23), we need the conditions $\Re(\alpha)>0$ and $\Re(\beta)>0$, ($\beta$ is the Greek small letter beta). It may be noted that

$$B(\alpha,\beta)=B(\beta,\alpha). \tag{5.24}$$

That is, the parameters $\alpha$ and $\beta$ can be interchanged in the integrals:

$$f_7(x)=\binom{n}{x}\frac{\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n+\beta-x)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}$$

for $x=0,1,\dots,n$; $\alpha>0$, $\beta>0$ [Beta-binomial probability function].

$$f_8(x)=\frac{\Gamma(r+s)\Gamma(x+n-r-s)\Gamma(x)\Gamma(n)}{\Gamma(r)\Gamma(s)\Gamma(x-r+1)\Gamma(n-s)\Gamma(x+n)}$$

for $x=r,r+1,\dots$, $s>0$, $n>s$; $r$ a positive integer [Beta-Pascal probability function].

$$f_9(x)=\sum_{i=1}^{m} w_i\binom{n}{x}p_i^x(1-p_i)^{n-x}$$

for $x=0,1,\dots,n$; $0<p_i<1$, $w_i\geq 0$, $i=1,\dots,m$; $\sum_{i=1}^m w_i=1$ [Mixed binomial probability function].

$$f_{10}(x)=\frac{\binom{n}{x}p^x(1-p)^{n-x}}{1-(1-p)^n},$$

for $x=1,2,\dots,n$; $0<p<1$; (truncated below $x=1$) [Truncated binomial probability function].

$$f_{11}(x)=\frac{(x\beta)^{x-1}}{x!}\mathrm{e}^{-x\beta}, \quad x=1,2,\dots;\ \beta>0$$

[Borel probability law].

$$f_{12}(x)=\frac{r}{(x-r)!}x^{x-r-1}\alpha^{x-r}\mathrm{e}^{-\alpha x},$$

for $x=r,r+1,\dots$; $\alpha>0$, where $r$ is a positive integer [Borel–Tanner probability law].

$$f_{13}(x)=\frac{\Gamma(\nu+x)\Gamma(\lambda)}{\Gamma(\lambda+x)\Gamma(\nu)}\frac{\mu^x}{{}_1F_1(\nu;\lambda;\mu)}$$

for $x=0,1,2,\dots$; $\nu>0$, $\lambda>0$, $\mu>0$ where ${}_1F_1$ is a confluent hypergeometric function ($\nu$ is Greek letter nu; $\mu$ is Greek letter mu) [Confluent hypergeometric probability law].

$$f_{14}(x) = w_1 \binom{N}{x} p_1^x (1-p_1)^{N-x} + w_2 \phi(x)$$

for $x = 0, 1, \dots, N$; $w_2 = 1 - w_1$, $0 < w_1 < 1$, $0 < p_1 < 1$ and $\phi(x)$ are some probability functions [Dodge–Romig probability law].

$$f_{15}(x) = \binom{N}{x} p^x \left[1 + \binom{N}{1} p + \dots + \binom{N}{c} p^c\right]^{-1}$$

for $x = 0, 1, \dots, c$; $0 < p < 1$; $N, c$ positive integers [Engset probability law].

$$f_{16}(x) = \frac{a(x)\theta^x}{[\sum_{x \in A} a(x)\theta^x]}, \quad \theta > 0$$

for $a(x) > 0$, $x \in A$ = subset of reals [Generalized power series probability function].

$$f_{17}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + x - 1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+x)}$$

for $x = 1, 2, \dots$; $\alpha > 0$, $\beta > 0$ [Compound geometric probability law].

$$f_{18}(x) = \mathrm{e}^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} \binom{2m}{x} p^x (1-p)^{2m-x}$$

for $x = 0, 1, 2, \dots, 2m$; $\lambda > 0$, $0 < p < 1$ [Hermite probability law].

$$f_{19}(x) = \frac{\Gamma(\lambda)\theta^x}{{}_1F_1(1; \lambda; \theta)\Gamma(\lambda + x)}$$

for $x = 0, 1, 2, \dots$; $\theta > 0$, $\lambda > 0$ [Hyper-Poisson probability function].

$$f_{20}(x) = \frac{\theta^x}{\beta x}, \quad 0 < \theta < 1,$$

for $x = 1, 2, \dots, d$ where $\beta = \sum_{x=1}^{d} \frac{\theta^x}{x}$ [Truncated logarithmic series probability function].

$$f_{21}(x) = \sum_{i=1}^{k} w_i f_i(x), \quad 0 < w_i < 1, \quad \sum_{i=1}^{k} w_i = 1$$

where $f_i(x)$ is a general probability or density function for each $i = 1, \dots, k$ [General mixed probability function].

$$f_{22}(x) = \frac{\Gamma(r+x)}{x!\Gamma(r)} p^r (1-p)^x$$

for $x = 0, 1, \dots$; $0 < p < 1$, $r > 0$ [Negative binomial probability function, model-2].

$$f_{23}(x) = \frac{\alpha^a (\alpha+1)^{p-a} \Gamma(p+x)}{\Gamma(p) x! (\alpha+2)^{p+x}} {}_2F_1\left(a, p+x; p; \frac{1}{\alpha+2}\right)$$

for $x = 0, 1, \dots$; $a > 0$, $p > 0$, $\alpha > 0$ and ${}_2F_1$ is a Gauss' hypergeometric function [Generalized negative binomial probability function].

$$f_{24}(x) = \frac{c^x}{x!} e^{-\lambda} \sum_{k=0}^{\infty} \frac{k^x (\lambda e^{-c})^k}{k!}$$

for $x = 0, 1, \ldots$; $\lambda, c$ positive constants [Neyman type A probability function].

$$f_{25}(x) = \binom{x+r-1}{x} p^r (1-p)^x$$

for $x = 0, 1, \ldots$; $0 < p < 1$; $r$ a positive integer [Pascal probability law].

$$f_{26}(x) = e^{-a} \sum_{m=0}^{\infty} \frac{a^m}{m!} \binom{nm}{x} p^x (1-p)^{nm-x}$$

for $x = 0, 1, \ldots, nm$; $a > 0$, $0 < p < 1$; $n, m$ positive integers [Poisson-binomial probability function].

$$f_{27}(x) = \frac{\mu^x}{x![\exp(\mu) - 1]}$$

for $x = 1, 2, \ldots$; $\mu > 0$ (truncated below $x = 1$) [Truncated Poisson probability function].

$$f_{28}(x) = \binom{N}{x} \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+N-x)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+N)}$$

for $x = 0, 1, \ldots, N$; $\alpha > 0, \beta > 0$ [Polya probability law or Beta-binomial probability function].

$$f_{29}(x) = \frac{1}{x!} \frac{\Gamma(x + \frac{h}{d})}{\Gamma(\frac{h}{d})} \left(1 + \frac{1}{d}\right)^{-(h/d)} \left(\frac{1}{1+d}\right)^x$$

for $x = 0, 1, \ldots$; $d > 0$, $h > 0$ [Polya–Eggenberger probability function].

$$f_{30}(x) = \frac{1}{n}, \quad x = x_1, \ldots, x_n$$

[Discrete uniform probability function].

$$f_{31}(x) = \frac{\exp[-(\lambda + \phi)]}{x!} \sum_{k=0}^{\infty} (k\theta + \phi)^x \frac{[\lambda \exp(-\theta)]^k}{k!}$$

for $x = 0, 1, \ldots$; $\phi > 0$, $\theta > 0$, $\lambda > 0$ [Short's probability law].

$$f_{32}(x) = \frac{1}{[1 - P(0)]} \sum_{t=x+1}^{\infty} \frac{P(t)}{t}$$

for $x = 0, 1, \ldots$; where $P(t)$ is any discrete probability function over the range $t = 0, 1, 2, \ldots$ [Ster's probability function].

$$f_{33}(x) = [\zeta(k) x^k]^{-1}$$

for $x = 1, 2, \ldots$ where $\zeta(k) = \sum_{t=1}^{\infty} t^{-k}$, $k > 1$ ($\zeta$ is the Greek letter zeta) [Zeta probability function].

## Exercises 5

**5.1.** Compute $E(x^2)$ for the geometric probability law by summing up or by using the definition, that is, by evaluating

$$E(x^2)=\sum_{x=1}^{\infty}x^2q^{x-1}p.$$

**5.2.** Compute (i) $E(x)$; (ii) $E(x^2)$; for the negative binomial probability law by using the definition (by summing up).

**5.3.** Compute (i) $E(x)$; (ii) $E(x^2)$; by using the technique used in the geometric probability law by differentiating the negative binomial probability law.

**5.4.** Compute $E(x)$ and $E(x^2)$ by differentiating the moment generating function in the Poisson probability case.

**5.5.** Compute $E(x)$ and variance of $x$ by using the moment generating function in the binomial probability law.

**5.6.** Construct two examples of discrete probability functions where $E(x)=\mathrm{Var}(x)$.

**5.7.** Solve the difference-differential equation in (5.16) and show that the solution is the probability function given therein.

**5.8.** Show that the functions $f_7(x)$ to $f_{33}(x)$ in Section 5.8 are all probability functions, that is, the functions are non-negative and the sum in each case is 1.

**5.9.** For the probability functions in Exercise 5.8, evaluate the first two moments about the origin, that is, $E(x)$ and $E(x^2)$, whenever they exist.

**Truncation.** In some practical problems, the general behavior of the discrete random variable $x$ may be according to a probability function $f(x)$ but certain values may not be admissible. In that case, we remove the total probability masses on the non-admissible points, then re-weigh the remaining points to create a new probability function. For example, in a binomial case suppose that the event of getting zero success is not admissible. In this case, we remove the point $x=0$. At $x=0$, the probability is $\binom{n}{0}p^0(1-p)^{n-0}=(1-p)^n$. Therefore, the remaining mass is $c_0=1-(1-p)^n$. Hence if we divide the remaining probabilities by $c_0$ then the remaining points can produce a truncated binomial probability law, which is

$$g(x)=\begin{cases}\frac{1}{c_0}\binom{n}{x}p^x(1-p)^{n-x}, & x=1,2,\dots,n,\ 0<p<1\\ 0, & \text{elsewhere.}\end{cases}$$

Here, $g(x)$ is called the truncated binomial probability function, truncated below $x=1$ or at $x=0$. Thus truncation is achieved by multiplying the probability function by an appropriate constant $c$. In the above case, it is $c=\frac{1}{c_0}$.

**5.10.** Compute the truncation constant $c$ so that $cf(x)$ is a truncated probability function of $f(x)$ in the following cases:

(i) Binomial probability function, truncated below $x = 1$ (Here, $c = \frac{1}{c_0}$ where $c_0$ is given above);
(ii) Binomial probability, truncated at $x = n$;
(iii) Poisson probability function, truncated below $x = 1$;
(iv) Poisson probability function, truncated below $x = 2$;
(v) Geometric probability function, truncated below $x = 2$;
(vi) Geometric probability function, truncated above $x = 10$.

**Probability Generating Function.** Consider a discrete random variable taking non-zero probabilities at the points $x = 0, 1, \dots$ and let $f(x)$ be the probability function. Consider the expected value of $t^x$ for some parameter $t$. Let us denote it by $P(t)$. Then we have

$$P(t) = E(t^x) = \sum_{x=0}^{\infty} t^x f(x) \tag{5.25}$$

where, for example, the probability that $x$ takes the value 5 is $\Pr\{x = 5\} = f(5)$ or it is the coefficient of $t^5$ on the right side of (5.25). Thus the various probabilities, such as $\Pr\{x = 0\}, \Pr\{x = 1\}, \dots$ are generated by $P(t)$ or they are available from the right side series in (5.25), provided the right side series is convergent. In the case when $x = 0, 1, 2, \dots$ with non-zero probabilities then $P(t)$ in (5.25) is called the generating function for the probability function $f(x)$ of this random variable $x$. We can also notice further properties of this generating function. Suppose that the series on the right side in (5.25) or $P(t)$ is differentiable, then differentiate with respect to $t$ and evaluate at $t = 1$, then we get $E(x)$. For example,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(t)\Big|_{t=1} &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{x=0}^{\infty} t^x f(x)\Big|_{t=1} \\ &= \sum_{x=0}^{\infty} x t^{x-1} f(x)\Big|_{t=1} = \sum_{x=0}^{\infty} x f(x) = E(x).\end{aligned}$$

Successive derivatives evaluated at $t = 1$ will produce $E(x), E[x(x-1)], E[x(x-1)(x-2)]$ and so on, when $P(t)$ series is uniformly convergent and differentiable term by term.

**5.11.** Compute the (a) the probability generating function $P(t)$, (b) $E(x)$ by using $P(t)$, (c) $E(x^2)$ by using $P(t)$ for the following cases: (i) Geometric probability law; (ii) Negative binomial probability law.

**5.12.** A gambler is betting on a dice game. Two dice will be rolled once. The gambler puts in Rs 5 (His bet is Rs 5). If the same numbers turn up on the two dice, then the gambler wins double his bet, that is, Rs 10, otherwise he loses his bet (Rs 5). Assuming that the dice are balanced

(i) What is the gambling house's expected return per game from this gambler?
(ii) What is the probability of the gambler winning exactly five out of 10 such games?
(iii) What is the gambler's expected return in 10 such games?

**5.13.** Cars are arriving at a service station at the rate of 0.1 per minute, time being measured in minutes. Assuming a Poisson arrival of cars to this service station, what is the probability that

(a) in a randomly selected twenty minute interval there are
  (i) exactly 3 arrivals;
  (ii) at least 2 arrivals;
  (iii) no arrivals;
(b) if 5 such 20-minute intervals are selected at random then what is the probability that in at least one of these intervals
  (i) (a)(i) happens;
  (ii) (a)(ii) happens;
  (iii) (a)(iii) happens.

**5.14.** The number of floods in a local river during rainy season is known to follow a Poisson distribution with the expected number of floods 3. What is the probability that

(a) during one rainy season
  (i) there are exactly 5 floods;
  (ii) there is no flood;
  (iii) at least one flood;
(b) if 3 rainy seasons are selected at random, then none of the seasons has
  (i) (a)(i) happening;
  (ii) (a)(ii) happening;
  (iii) (a)(iii) happening;
(c) (i) (a)(i) happens for the first time at the 3rd season;
  (ii) (a)(iii) happens for the second time at the 3rd season.

**5.15.** From a well-shuffled deck of 52 playing cards (13 spades, 13 clubs, 13 hearts, 13 diamonds) a hand of 8 cards is selected at random. What is the probability that the hand contains (i) 5 spades? (ii) no spades? (iii) 5 spades and 3 hearts? (iv) 3 spades 2 clubs, 2 hearts, 1 diamond?

# 6 Commonly used density functions

## 6.1 Introduction

Here, we will deal with the continuous case. Some most commonly used density functions will be discussed here and at the end a few more densities will be listed. The very basic density function is the uniform or rectangular density as shown in Figure 6.1. This was already introduced in Example 4.4 in Chapter 4 and the mean value and variance were evaluated there. For the sake of completeness, we will list here again.

## 6.2 Rectangular or uniform density

$$f_1(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise.} \end{cases}$$

The graph looks like a rectangle, and hence it is also called a *rectangular density*. The total probability mass 1 is uniformly distributed over the interval $[a, b]$, $b > a$, and hence it is called a uniform density.

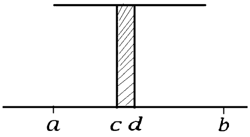


**Figure 6.1:** Uniform or rectangular density.

The probability that the random variable $x$ falls in the interval $a < c \le x \le d < b$ is marked in the graph. It is the area under the curve between the ordinates at $x = c$ and $x = d$. It was shown in Example 4.4. that

$$E(x) = \frac{b+a}{2} \quad \text{and} \quad \text{Var}(x) = \frac{(b-a)^2}{12}. \tag{6.1}$$

The moment generating function in this case is the following:

$$\begin{aligned} M(t) = E[e^{tx}] &= \int_a^b \frac{e^{tx}}{b-a} dx \\ &= \left[\frac{1}{t(b-a)} e^{tx}\right]_a^b = \frac{[e^{bt} - e^{at}]}{t(b-a)}, \quad t \ne 0. \end{aligned} \tag{6.2}$$

One can also obtain the moments by differentiating as well as by expanding $M(t)$ here. For example, let us evaluate the first two moments by expanding $M(t)$:

$$M(t) = \frac{1}{t(b-a)} \left\{ \left[1 + b\frac{t}{1!} + b^2\frac{t^2}{2!} + \cdots\right] - \left[1 + a\frac{t}{1!} + a^2\frac{t^2}{2!} + \cdots\right] \right\}$$

https://doi.org/10.1515/9783110562545-006

$$\begin{aligned}
&= \frac{1}{t(b-a)}\left\{(b-a)\frac{t}{1!} + (b^2-a^2)\frac{t^2}{2!} + (b^3-a^3)\frac{t^3}{3!} + \cdots\right\} \\
&= 1 + \frac{(b^2-a^2)}{(b-a)}\frac{t}{2!} + \frac{(b^3-a^3)}{(b-a)}\frac{t^2}{3!} + \cdots \\
&= 1 + \frac{(b+a)}{2}\frac{t}{1!} + \frac{(b^2+ab+a^2)}{3}\frac{t^2}{2!} + \cdots .
\end{aligned}$$

Since

$$\frac{(b^2-a^2)}{(b-a)} = (b+a) \quad \text{and} \quad \frac{(b^3-a^3)}{(b-a)} = b^2+ab+a^2$$

we have the coefficient of $\frac{t^1}{1!}$ as $\frac{(b+a)}{2}$ and the coefficient of $\frac{t^2}{2!}$ as $\frac{(a^2+ab+b^2)}{3}$. Hence

$$E(x) = \frac{(a+b)}{2} \quad \text{and} \quad E(x^2) = \frac{(a^2+ab+b^2)}{3} \tag{6.3}$$

in the uniform distribution.

**Example 6.1.** Example of a random cut or a point taken at random on a line segment was considered in Chapters 1 and 2. Consider the problem involving areas. A girl is throwing darts at a circular board of diameter 2 meters (2 m). The aim is to hit the center of the board. Assume that she has no experience and she may hit anywhere on the board. What is the probability that she will hit a specified $\frac{1}{2} \times \frac{1}{2}$ square meters region on the board?

**Solution 6.1.** Let d$A$ be an infinitesimal area on the board around the point of hit. Due to her lack of experience, we may assume that the point of hit is uniformly distributed over the area of the board. The area of the board is $\pi r^2 = \pi 1^2 = \pi$ square meters, or $\pi$ m$^2$. Then we have the density

$$f(A)\mathrm{d}A = \begin{cases} \frac{\mathrm{d}A}{\pi}, & 0 \leq A \leq \pi \\ 0, & \text{elsewhere.} \end{cases}$$

The integral over the specified square will give the area of the square, which is, $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ m$^2$. Hence the required probability is

$$\frac{1}{4}\frac{1}{\pi} = \frac{1}{4\pi}.$$

Several densities are connected with a gamma function. Hence we define a gamma function next.

**Notation 6.1.** $\Gamma(\alpha)$: gamma function.

**Definition 6.1** (A gamma function). A gamma function, denoted by $\Gamma(\alpha)$, exists for all values of $\alpha$, positive, negative, rational irrational, complex values of $\alpha$ except for

$\alpha = 0, -1, -2, \ldots$. A gamma function can be defined in many ways. Detailed definitions and properties may be seen from the book [2]. But when defining densities, we will need only an integral representation for a gamma function. Such an integral representation is given below:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \mathrm{d}x, \quad \Re(\alpha) > 0. \tag{6.4}$$

This integral exists only for real value of $\alpha$ greater than zero. In statistical problems, usually $\alpha$ is real and then the condition will become $\alpha > 0$. Other integral representations are available for a gamma function, each with its own conditions. We will make use of only (6.4) in this book. Two basic properties of the gamma function that we will make use of are the following:

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \tag{6.5}$$

when $\Gamma(\alpha - 1)$ is defined. Continuing the process, we have

$$\Gamma(\alpha) = (\alpha - 1)(\alpha - 2) \cdots (\alpha - r)\Gamma(\alpha - r) \tag{6.6}$$

when $\Gamma(\alpha - r)$ is defined. This property can be seen from the integral representation in (6.4) by integrating by parts, by taking $\mathrm{d}v = \mathrm{e}^{-x}$ and $u = x^{\alpha-1}$ and then using the formula $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$ [Verification is left to the student]. From (6.6), it follows that if $\alpha$ is a positive integer, say, $n = 1, 2, 3, \ldots$ then

$$\Gamma(n) = (n - 1)!, \quad n = 1, 2, \ldots. \tag{6.7}$$

The second property that we will use is that

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{6.8}$$

This will be proved only after considering joint distributions of more than one random variable in the coming chapters. Hence the student may take the result for granted for the time being. Note that the above results (6.5), (6.6) are valid for all $\alpha \neq 0, -1, -2, \ldots$, $\alpha$ need not be a positive number or $\Re(\alpha)$, when $\alpha$ is complex, need not be positive. For example,

$$\begin{aligned}
\Gamma\left(\frac{5}{2}\right) &= \left(\frac{5}{2} - 1\right)\left(\frac{5}{2} - 2\right)\Gamma\left(\frac{5}{2} - 2\right) \\
&= \left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) = \left(\frac{3}{4}\right)\sqrt{\pi}. \\
\Gamma\left(\frac{1}{2}\right) &= \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\Gamma\left(-\frac{5}{2}\right) \\
&= -\frac{15}{8}\Gamma\left(-\frac{5}{2}\right) \quad \Rightarrow \\
\Gamma\left(-\frac{5}{2}\right) &= -\frac{8}{15}\sqrt{\pi}.
\end{aligned}$$

$$\begin{aligned}
\Gamma\left(\frac{10}{3}\right) &= \left(\frac{7}{3}\right)\left(\frac{4}{3}\right)\left(\frac{1}{3}\right)\Gamma\left(\frac{1}{3}\right) = \frac{28}{27}\Gamma\left(\frac{1}{3}\right). \\
\Gamma(3.8) &= (2.8)(1.8)(0.8)\Gamma(0.8). \\
\Gamma(1.3) &= (0.3)\Gamma(0.3). \\
\Gamma(1.3) &= (0.3)(-0.7)(-1.7)\Gamma(-1.7) = (0.357)\Gamma(-1.7) \quad \Rightarrow \\
\Gamma(-1.7) &= \frac{\Gamma(1.3)}{0.357}.
\end{aligned}$$

By using the above procedures, one can reduce any gamma function $\Gamma(\beta)$, with $\beta$ real, to a $\Gamma(\alpha)$ where $0 < \alpha \leq 1$, and $\Gamma(\alpha)$ is extensively tabulated when $0 < \alpha < 1$, such numerical tables are available. A density associated with a gamma function of (6.4) is called a gamma density. A two-parameter gamma density is defined next.

## 6.3 A two-parameter gamma density

$$f_2(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\frac{x}{\beta}}, & 0 \leq x < \infty, \beta > 0, \alpha > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

Since the total probability is 1, we have

$$\begin{aligned}
1 &= \int_0^\infty f_2(x)\mathrm{d}x \quad \Rightarrow \\
1 &= \int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} \mathrm{e}^{-\frac{x}{\beta}} \mathrm{d}x \quad \Rightarrow \\
\beta^\alpha &= \int_0^\infty \frac{x^{\alpha-1}\mathrm{e}^{-\frac{x}{\beta}}\mathrm{d}x}{\Gamma(\alpha)}.
\end{aligned} \tag{6.9}$$

This (6.9) is a very useful representation where the only conditions needed are $\beta > 0$ and $\alpha > 0$, then we can replace $\beta^\alpha$ by a gamma integral.

From Figure 6.2, note that the parameter $\beta$ has a scaling effect, and hence $\beta$ is called the *scale parameter* and $\alpha$ is called the *shape parameter* in the gamma case because $\alpha$ throws light on the shape of the density curve. Gamma density is one of the main densities in probability and statistics. Many other densities are associated with it, some of which will be listed later. An extended form of the two-parameter gamma density is available by replacing $x$ by $|x|$ so that the mirror image of the graph is there on the left of the $y$-axis also. An extended form of the gamma density function is then given by

$$f_2^*(x) = \frac{1}{2}\frac{1}{\beta^\alpha \Gamma(\alpha)} |x|^{\alpha-1} \mathrm{e}^{-\frac{|x|}{\beta}}, \quad -\infty < x < \infty, \beta > 0, \alpha > 0.$$

One graph for one set of parameters $\alpha$ and $\beta$ is given in Figure 6.3.

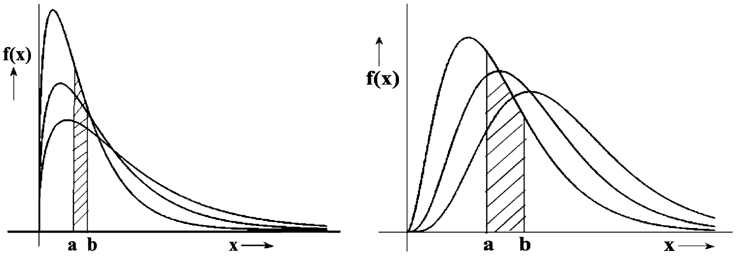


**Figure 6.2:** Gamma density: (a) varying $\beta$, fixed $\alpha$; (b) Varying $\alpha$, fixed $\beta$.

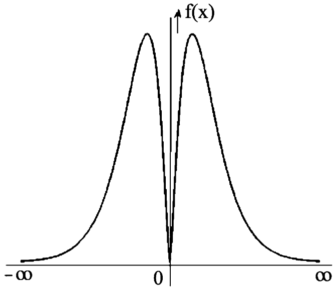


**Figure 6.3:** An extended form of the gamma density.

Let us evaluate arbitrary $(s-1)$-th moment and the moment generating function of the gamma density for $x > 0$. Observe that $(s-1)$-th moment is also the Mellin transform of the density function:

$$E[x^{s-1}] = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} x^{\alpha-1+s-1}\mathrm{e}^{-\frac{x}{\beta}}\mathrm{d}x.$$

Substitute

$$y = \frac{x}{\beta} \quad \Rightarrow \quad \mathrm{d}x = \beta \mathrm{d}y.$$

Then

$$\begin{aligned} E[x^{s-1}] &= \int_0^{\infty} \frac{(\beta y)^{\alpha+s-2}}{\beta^{\alpha}\Gamma(\alpha)}\mathrm{e}^{-y}\beta \mathrm{d}y \\ &= \frac{\beta^{s-1}}{\Gamma(\alpha)}\int_0^{\infty} y^{(\alpha+s-1)-1}\mathrm{e}^{-y}\mathrm{d}y. \end{aligned}$$

But this integral is a gamma function and it is

$$= \beta^{s-1}\frac{\Gamma(\alpha+s-1)}{\Gamma(\alpha)}, \quad \Re(\alpha+s-1) > 0. \tag{6.10}$$

From this general moments, we can obtain all integer moments also. Put $s = 2$ to obtain $E(x)$ and $s = 3$ to get $E(x^2)$:

$$
\begin{aligned}
E(x) &= \beta^{s-1}\frac{\Gamma(\alpha+s-1)}{\Gamma(\alpha)}\Big|_{s=2} \\
&= \beta\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \beta\alpha\frac{\Gamma(\alpha)}{\Gamma(\alpha)} = \alpha\beta
\end{aligned} \tag{6.11}
$$

by using the formula (6.5).

$$
E(x^2) = \beta^2\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \beta^2(\alpha+1)(\alpha)\frac{\Gamma(\alpha)}{\Gamma(\alpha)} = \alpha(\alpha+1)\beta^2. \tag{6.12}
$$

The reduction in the gamma is done by using (6.6). Then the variance for a gamma random variable is given by

$$
\mathrm{Var}(x) = E[x^2] - [E(x)]^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2. \tag{6.13}
$$

Note that the mean value is $\alpha\beta$ and the variance is $\alpha\beta^2$, and hence if $\beta = 1$ then the mean value is equal to the variance in this case also, just like the Poisson case as seen from Chapter 5.

The moment generating function $M(t)$ in the gamma case is the following:

$$
E[\mathrm{e}^{tx}] = \int_0^\infty \mathrm{e}^{tx} f_2(x)\mathrm{d}x = \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty \mathrm{e}^{tx-\frac{x}{\beta}}\mathrm{d}x.
$$

Put the exponent as $-y = -[\frac{1}{\beta} - t]x$, which gives $\mathrm{d}y = [\frac{1}{\beta} - t]\mathrm{d}x$ and integrate to obtain

$$
M(t) = \frac{[\frac{1}{\beta} - t]^{-\alpha}}{\beta^\alpha} = (1-\beta t)^{-\alpha} \quad \text{for } (1-\beta t) > 0. \tag{6.14}
$$

This condition is needed for the integral to be convergent, otherwise the exponent in the integral can become positive and the integral will give $+\infty$ or the integral diverges. When a random variable $x$ is gamma distributed with two parameter gamma density as $f_2(x)$ above then we write $x \sim \mathrm{gamma}(\alpha,\beta)$ where

“$x \sim$ ” stands for “$x$ is distributed as”

What is the probability that the gamma random variable $x$ in $f_2(x)$ lies over the interval $[a,b]$, $a > 0$, $b > a$? This is given by the area under the gamma curve between the ordinates at $x = a$ and $x = b$. This can be obtained by integrating $f_2(x)$ from $a$ to $b$. This is shown in Figure 6.4.

The integral from $a$ to $b$ is the same as the integral from 0 to $b$ minus the integral from 0 to $a$. Let us see what is such an integral. For example, what is the integral from 0 to $a$:

$$
\begin{aligned}
\int_0^a f_2(x)\mathrm{d}x &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^a x^{\alpha-1}\mathrm{e}^{-\frac{x}{\beta}}\mathrm{d}x \\
&= \int_0^{a/\beta} \frac{y^{\alpha-1}}{\Gamma(\alpha)}\mathrm{e}^{-y}\mathrm{d}y, \quad y = \frac{x}{\beta} \\
&= \frac{1}{\Gamma(\alpha)}\gamma\left(\alpha, \frac{a}{\beta}\right)
\end{aligned} \tag{6.15}
$$

where $\gamma(\cdot,\cdot)$ is the *incomplete gamma* and this incomplete gamma is tabulated and numerical tables are available. Thus, for evaluating probabilities one has to use incomplete gamma tables if $\alpha \neq 1$. When $\alpha = 1$, then it is an exponential density which can be integrated and the probabilities can be evaluated directly. If $\alpha$ is a positive integer, then also one can integrate by parts and obtain explicit forms.

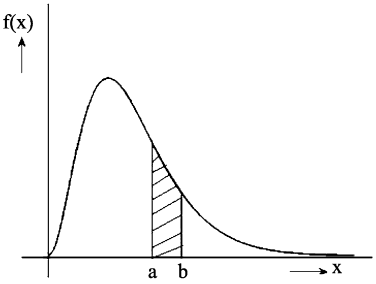


**Figure 6.4:** Probability in the gamma case.

Some special cases of the gamma density are the following.

### 6.3.1 Exponential density

One parameter exponential density was dealt with in Chapters 3 and 4. This is available from the gamma density $f_2(x)$ by putting $\alpha = 1$. That is,

$$f_3(x) = \begin{cases} \frac{1}{\beta}\mathrm{e}^{-\frac{x}{\beta}}, & 0 \leq x < \infty, \beta > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

The moments and moment generating function are available from the corresponding quantities for the gamma case by putting $\alpha = 1$. Exponential density is widely used as a model to describe waiting time, such as waiting in a queue, waiting for a scheduled bus etc. But it may not be a good model for all types of waiting times. If the waiting time consists of several components of waiting times such as waiting for completing a medical examination at a doctor's office, which may consist of blood test, physical examination, checking weight and height, X-ray, etc., then the individual components may be exponentially distributed but the total waiting time is usually gamma distributed.

### 6.3.2 Chi-square density

In a gamma density when $\beta = 2$ and $\alpha = \frac{n}{2}$, $n = 1, 2, \ldots$, then we obtain a chi-square density with $n$ degrees of freedom. The density is the following:

$$f_4(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \mathrm{e}^{-\frac{x}{2}}, & 0 \leq x < \infty,\ n = 1, 2, \ldots \\ 0, & \text{elsewhere.} \end{cases}$$

The meaning of "degrees of freedom" will be given when we consider sampling distributions later on. For the time being, the student may take it as a parameter $n$ in $f_4(x)$, taking integer values $1, 2, \ldots$. Chi-square density is widely used in statistical decision making such as testing of statistical hypotheses, model building, designing of experiments, regression analysis, etc. In fact, this is one of the main distributions in statistical inference. A chi-square random variable with $k$ degrees of freedom is denoted by $\chi_k^2$.

**Notation 6.2.** $\chi_k^2$: chi-square with $k$ degrees of freedom.

**Definition 6.2** (Chi-square random variable). A chi-square random variable, with $k$ degrees of freedom is a gamma random variable with the parameters $\alpha = \frac{k}{2}$ and $\beta = 2$, $k = 1, 2, \ldots$.

**Example 6.2.** The waiting time for the first pregnancy among women in a certain community from the time of marriage or cohabitation is found to be gamma distributed with scale parameter $\beta = 1$ and the expected waiting time 3 months, time being measured in months. (i) If a freshly married woman from this community is selected at random, what is the probability that she has to wait at least eight months before she gets pregnant? (ii) If three freshly married women are selected at random from this community, then what is the probability that at least two of them have to wait at least eight months to get pregnant?

**Solution 6.2.** The expected value in the gamma case is found to be $\alpha\beta$ and if $\beta = 1$ then $\alpha$ is given to be 3. Waiting for at least eight months means that the waiting time $t \geq 8$. Hence we need $p = \Pr\{t \geq 8\}$. That is,

$$\begin{aligned} p = \Pr\{t \geq 8\} &= \int_8^\infty \frac{1}{\Gamma(3)} x^{3-1} \mathrm{e}^{-t} \mathrm{d}t \\ &= \frac{1}{2} \int_8^\infty x^2 \mathrm{e}^{-x} \mathrm{d}x \end{aligned}$$

since $\beta^\alpha = 1$, $\Gamma(3) = 2! = 2$. Since $\alpha$ is a positive integer, we can integrate by parts by taking $\mathrm{d}v = \mathrm{e}^{-x}$ and $u$ as $x^2$. That is,

$$\begin{aligned} p &= \frac{1}{2} \int_8^\infty x^2 \mathrm{e}^{-x} \mathrm{d}x \\ &= \frac{1}{2} \left\{ -[x^2 \mathrm{e}^{-x}]_8^\infty + 2 \int_8^\infty x \mathrm{e}^{-x} \mathrm{d}x \right\} \\ &= \frac{1}{2} \{ 64\mathrm{e}^{-8} - 2[x\mathrm{e}^{-x}]_8^\infty - 2[\mathrm{e}^{-x}]_8^\infty \end{aligned}$$

$$= \frac{1}{2} 82e^{-8} = 41e^{-8}.$$

This answers (i). For answering (ii), we consider three Bernoulli trials with probability of success $p$ above and the required probability is the probability that the number of successes is 2 or 3. That is,

$$\sum_{x=2}^{3} \binom{3}{x} p^x (1-p)^{3-x} = \binom{3}{2} p^2 (1-p) + \binom{3}{3} p^3 = 3p^2(1-p) + p^3 = p^2(3-2p).$$

Another density, having connection to a gamma function is the generalized gamma density.

## 6.4 Generalized gamma density

$$f_5(x) = \begin{cases} cx^{\alpha-1} e^{-bx^\delta}, & b > 0,\ \alpha > 0,\ \delta > 0,\ x \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

where $c$ is the normalizing constant, which can be evaluated by using a gamma integral. Make the substitution

$$y = bx^\delta \quad \Rightarrow \quad x = \left(\frac{y}{b}\right)^{1/\delta} \quad \Rightarrow \quad dx = \frac{1}{\delta} \frac{1}{b^{1/\delta}} y^{(1/\delta)-1} dy.$$

The above transformations are valid since $x > 0$. The limits will remain the same.

$$\begin{aligned} \int_0^\infty x^{\alpha-1} e^{-bx^\delta} dx &= \frac{1}{\delta} \frac{1}{b^{(\alpha/\delta)}} \int_0^\infty y^{\frac{\alpha}{\delta}-1} e^{-y} dy \\ &= \delta^{-1} b^{-\frac{\alpha}{\delta}} \Gamma\left(\frac{\alpha}{\delta}\right). \end{aligned} \tag{6.16}$$

The conditions $\alpha > 0$, $\delta > 0$ are already satisfied. Since the total integral is 1 (total probability) the normalizing constant

$$c = \frac{\delta b^{\frac{\alpha}{\delta}}}{\Gamma(\frac{\alpha}{\delta})}. \tag{6.17}$$

This is one generalized form of the gamma density. One form was introduced in Section 6.4. The form above in Section 6.4, is usually known in the literature as the generalized gamma density. When $\delta = 1$, we have the **two-parameter gamma family**. When $\alpha = \delta$, we have a popular density known as the **Weibull density**, which is also available by using a power transformation in an exponential density. When $\alpha = \delta$, the normalizing constant in (6.17) becomes $c = \delta b$. When $\alpha = 3$, $\delta = 2$, we have one form of **Maxwell–Boltzmann density** in physics. When $\delta = 1$, $\alpha$ an integer we have **Erlang density**. When $\delta = 2$, $\alpha = 2$, we have the **Rayleigh density**. Many more such densities are particular cases of a generalized gamma density.

### 6.4.1 Weibull density

This is the special case of the generalized gamma density where $\delta = \alpha$.

$$f_6(x) = \begin{cases} \delta b x^{\delta-1} \mathrm{e}^{-bx^{\delta}}, & b > 0,\ \delta > 0,\ 0 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

This density is widely used as a model in many practical problems. Over a thousand research papers are available on the application of this model.

Another function associated with a gamma function is the beta function.

**Notation 6.3.** $B(\alpha,\beta)$: Beta function.

**Definition 6.3.** A beta function $B(\alpha,\beta)$ is defined as

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \Re(\alpha) > 0, \quad \Re(\beta) > 0. \tag{6.18}$$

As in the case of gamma function, beta function can also be given integral representations.

$$\begin{aligned} B(\alpha,\beta) &= \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x, \quad \Re(\alpha) > 0,\ \Re(\beta) > 0 \\ &= \int_0^1 y^{\beta-1}(1-y)^{\alpha-1}\mathrm{d}y, \quad \Re(\alpha) > 0, \quad \Re(\beta) > 0 \\ &= B(\beta,\alpha). \end{aligned} \tag{6.19}$$

The second integral is available from the first by putting $y = 1 - x$. These two integrals are called *type-1 beta integrals*.

$$\begin{aligned} B(\alpha,\beta) &= \int_0^\infty z^{\alpha-1}(1+z)^{-(\alpha+\beta)}\mathrm{d}x, \quad \Re(\alpha) > 0, \quad \Re(\beta) > 0 \\ &= \int_0^\infty u^{\beta-1}(1+u)^{-(\alpha+\beta)}\mathrm{d}y, \quad \Re(\alpha) > 0, \quad \Re(\beta) > 0 \\ &= B(\beta,\alpha). \end{aligned} \tag{6.20}$$

These are called *type-2 beta integrals*.

Integrals in (6.20) can also be obtained from (6.19) by using the substitution $z = \frac{x}{1-x}$ and the last integral from (6.20) by the substitution $u = \frac{1}{z}$. Transformation of variables will be discussed after introducing joint distributions in the next chapter. One variable transformation will be discussed at the end of the present chapter. Connections of beta integrals to gamma integral will be considered after introducing joint distributions. We will introduce beta densities associated with these beta integrals.

## 6.5 Beta density

There are two types, type-1, and type-2 which is also called "inverted beta density". But we will use the terminology "type-2 beta" instead of inverted beta. The type-1 beta density is associated with the type-1 beta integral.

$$f_7(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1,\ \alpha > 0,\ \beta > 0 \\ 0, \quad \text{otherwise.} \end{cases}$$

A family of densities is available for various values of the parameters $\alpha$ and $\beta$. When $\beta = 1$, it behaves like $x^{\alpha-1}$ and when $\alpha = 1$ it is of the form $(1-x)^{\beta-1}$, and both of these are power functions. A few of them are shown in Figure 6.5.

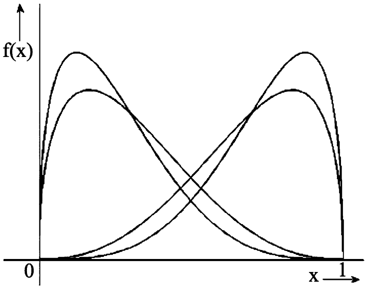


**Figure 6.5:** Type-1 beta densities.

Let us evaluate an arbitrary $h$-th moment of a type-1 beta random variable.

$$E[x^h] = \int_0^1 x^h f_7(x)\mathrm{d}x = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{(\alpha+h)-1}(1-x)^{\beta-1}\mathrm{d}x$$

which is available from type-1 beta integral by replacing $\alpha$ by $\alpha + h$. That is,

$$\begin{aligned} E[x^h] &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+h)\Gamma(\beta)}{\Gamma(\alpha+h+\beta)}, \quad \Re(\alpha+h) > 0 \\ &= \frac{\Gamma(\alpha+h)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+h)}, \quad \Re(\alpha+h) > 0. \end{aligned} \tag{6.21}$$

If $h$ is real, then it is possible to have some negative moments existing such that $\alpha + h > 0$. If $\alpha = 5.8$, then $h$ can be down to $-5.8$ but not equal to $-5.8$. When $h = s - 1$, we have the Mellin transform of the density $f_7(x)$. The moment generating function will go into a series unless $\alpha$ and $\beta$ are positive integers. Hence we will not consider the moment generating function here. [For obtaining the series, expand $\mathrm{e}^{tx}$ and integrate term by term. It will go into a hypergeometric series of the ${}_2F_1$ type.] From (6.21), the first two moments are available by putting $h = 1$ and $h = 2$ and then simplifying by using the formulae (6.5) and (6.6), and they the following, by observing that:

$$\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}=\alpha\frac{\Gamma(\alpha)}{\Gamma(\alpha)}=\alpha$$
$$\text{and}\quad \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}=\alpha(\alpha+1)$$

and similarly for other gamma ratios. That is,

$$\begin{aligned}E[x]&=\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)}\\&=\frac{\alpha}{\alpha+\beta}.\end{aligned}\tag{6.22}$$

$$\begin{aligned}E[x^2]&=\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+2)}\\&=\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.\end{aligned}\tag{6.23}$$

By using (6.22) and (6.23), one can evaluate the variance of a type-1 beta random variable. The type-2 beta density is associated with the type-2 beta integral and it is the following:

$$f_8(x)=\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1+x)^{-(\alpha+\beta)}, & \alpha>0,\ \beta>0,\ x\geq 0\\ 0, & \text{otherwise.}\end{cases}$$

Various shapes are there for various values of $\alpha$ and $\beta$, a few are shown in Figure 6.6.

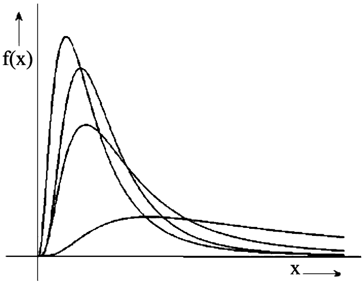


**Figure 6.6:** Family of type-2 beta densities.

In a type-2 beta density if we make a transformation $x=\frac{m}{n}F$, $m,n=1,2,\ldots$, then we get the *F-density* or the variance ratio density, which is one of the main densities in statistical applications. Both type-1 and type-2 beta random variables are connected to gamma random variables. Let us evaluate the $h$-th moment for an arbitrary $h$ in the type-2 beta case.

$$E[x^h]=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^{\infty}x^{\alpha+h-1}(1+x)^{-(\alpha+\beta)}\mathrm{d}x$$

which is available from a type-2 beta integral by replacing $\alpha$ by $\alpha+h$ and $\beta$ by $\beta-h$

$$
\begin{aligned}
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+h)\Gamma(\beta-h)}{\Gamma(\alpha+\beta)} \\
&= \frac{\Gamma(\alpha+h)}{\Gamma(\alpha)} \frac{\Gamma(\beta-h)}{\Gamma(\beta)}, \quad \Re(\alpha+h)>0, \quad \Re(\beta-h)>0. \qquad (6.24)
\end{aligned}
$$

Thus, the effective condition on $h$ is that $-\Re(\alpha)<\Re(h)<\Re(\beta)$. If $\alpha,\beta,h$ are real, then $-\alpha<h<\beta$. Thus only a few moments in this interval will exist. Outside that, the moments will not exist. Let us look into the first two moments. Observe that

$$
\begin{aligned}
&\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha; \\
&\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \alpha(\alpha+1); \\
&\frac{\Gamma(\beta-1)}{\Gamma(\beta)} = \frac{1}{\beta-1}, \quad \beta \neq 1; \\
&\frac{\Gamma(\beta-2)}{\Gamma(\beta)} = \frac{1}{(\beta-1)(\beta-2)}, \quad \beta \neq 1,2.
\end{aligned}
$$

Hence

$$
E[x] = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\Gamma(\beta-1)}{\Gamma(\beta)} = \frac{\alpha}{(\beta-1)}, \quad \beta \neq 1. \qquad (6.25)
$$

$$
E[x^2] = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} \frac{\Gamma(\beta-2)}{\Gamma(\beta)} = \frac{\alpha(\alpha+1)}{(\beta-1)(\beta-2)}, \quad \beta \neq 1,2. \qquad (6.26)
$$

By using (6.25) and (6.26), one can compute the variance of a type-2 beta random variable.

**Example 6.3.** The proportion of people who are politically conscious from village to village in Tamilnadu is seen to be type-1 beta distributed with the parameters $\alpha = 2.5$ and $\beta = 3.5$, that is, $x \sim$ type-1 beta($\alpha = 2.5, \beta = 3.5$). If a village is selected at random, what is the probability that the proportion of politically conscious people in this village is below 0.2.

**Solution 6.3.** The required probability, say $p$, is available from the area under the curve between the ordinates at $x = 0$ and $x = 0.2$ or from the integral

$$
\begin{aligned}
p &= \frac{\Gamma(6)}{\Gamma(2.5)\Gamma(3.5)} \int_0^{0.2} x^{2.5-1}(1-x)^{3.5-1} \mathrm{d}x \\
&= \frac{\Gamma(6)}{\Gamma(2.5)\Gamma(3.5)} \int_0^{0.2} x^{1.5}(1-x)^{2.5} \mathrm{d}x.
\end{aligned}
$$

The gammas can be simplified.

$$\begin{aligned}
&\Gamma(6) = 5! = 120.\\
&\Gamma(2.5) = \Gamma\left(\frac{5}{2}\right) = \left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) = \frac{3}{4}\sqrt{\pi}.\\
&\Gamma(3.5) = \left(\frac{5}{2}\right)\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) = \frac{15}{8}\sqrt{\pi}.
\end{aligned}$$

Therefore,

$$\frac{\Gamma(6)}{\Gamma(2.5)\Gamma(3.5)} = \frac{256}{3\pi}.$$

But the integral part cannot be explicitly evaluated. If $\beta$ was a positive integer, then we could have expanded $(1-x)^{\beta-1}$ by using a binomial expansion and integrate term by term to produce a sum of a finite number of terms. If $\alpha$ was a positive integer, then one could have transformed $x = 1 - y$ [the limits will change] and expanded $(1-y)^{\alpha-1}$, which would have also produced a finite sum. Our exponents here are $\alpha - 1 = 1.5$ and $\beta - 1 = 2.5$, not integers. Then what we can do is either expand $(1-x)^{2.5}$ by using a binomial expansion, then integrate term by term, which will give a convergent infinite series or one can use what is known as the *incomplete beta* tables. Integrals of the type

$$B_a(\alpha,\beta) = \int_0^a x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x \tag{6.27}$$

are tabulated for various values of $a, \alpha, \beta$, called incomplete beta tables, or use a program such as Maple or Mathematica, which will produce the numerical answer also. In (6.27), if we expand $(1-x)^{\beta-1}$, integrate term by term, then we will get a hypergeometric series of the ${}_2F_1$ type. We will leave the final computation as an exercise to the students.

## 6.6 Laplace density

A density which is a good model to describe simple input-output situations, opposing forces, creation of sand dunes etc is the *Laplace density or double exponential density*:

$$f_9(x) = \begin{cases} c\mathrm{e}^{-\theta|x|}, & -\infty < x < \infty,\ \theta > 0\\ 0, & \text{elsewhere}\end{cases}$$

where $c$ is the normalizing constant. Let us evaluate $c$. Note that

$$|x| = \begin{cases} -x & \text{for } x < 0\\ x & \text{for } x \geq 0.\end{cases}$$

The total probability or the total integral should be 1. Therefore,

$$1 = c\int_{-\infty}^{\infty} \mathrm{e}^{-\theta|x|}\mathrm{d}x = c\int_{-\infty}^{0} \mathrm{e}^{-\theta|x|}\mathrm{d}x + c\int_{0}^{\infty} \mathrm{e}^{-\theta|x|}\mathrm{d}x$$

$$= c\int_{-\infty}^{0} \mathrm{e}^{-\theta(-x)}\mathrm{d}x + c\int_{0}^{\infty} \mathrm{e}^{-\theta x}\mathrm{d}x = c\int_{0}^{\infty} \mathrm{e}^{-\theta y}\mathrm{d}y + c\int_{0}^{\infty} \mathrm{e}^{-\theta x}\mathrm{d}x$$

by changing $y = -x$ in the first integral

$$1 = 2c\int_{0}^{\infty} \mathrm{e}^{-\theta t}\mathrm{d}t = 2c\left[-\frac{1}{\theta}\mathrm{e}^{-\theta t}\right]_0^{\infty} = \frac{2c}{\theta}.$$

Therefore, $c = \theta/2$. The graph of the density is given in Figure 6.7.

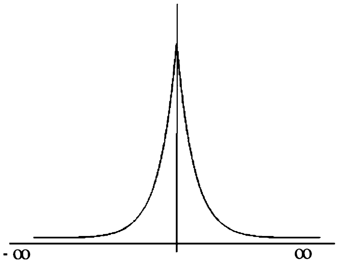


**Figure 6.7:** Laplace density.

Laplace density is widely used in non-Gaussian stochastic processes and time series models also.

**Example 6.4.** For a Laplace density with parameter $\theta = 2$, compute the probability over the interval $[-3, 2]$.

**Solution 6.4.** From the graph in Figure 6.7, note that Laplace density is a symmetric density, symmetric about $x = 0$. Hence

$$\begin{aligned}
\Pr\{-3 \le x \le 2\} &= \int_{-3}^{2} \mathrm{e}^{-2|x|}\mathrm{d}x \\
&= \int_{-3}^{0} \mathrm{e}^{-2|x|}\mathrm{d}x + \int_{0}^{2} \mathrm{e}^{2|x|}\mathrm{d}x \\
&= \int_{0}^{3} \mathrm{e}^{-2x}\mathrm{d}x + \int_{0}^{2} \mathrm{e}^{-2x}\mathrm{d}x \\
&= \left[-\frac{1}{2}\mathrm{e}^{-2x}\right]_0^3 + \left[-\frac{1}{2}\mathrm{e}^{-2x}\right]_0^2 \\
&= \frac{1}{2}[1 - \mathrm{e}^{-6}] + \frac{1}{2}[1 - \mathrm{e}^{-4}] = 1 - \frac{1}{2}[\mathrm{e}^{-6} + \mathrm{e}^{-4}].
\end{aligned}$$

## 6.7 Gaussian density or normal density

The most widely used density is the Gaussian density, which is also called the normal density. [This is another unfortunate technical term. This does not mean that other densities are abnormal or this is some standard density and others are abberations.] The density is the following:

$$f_{10}(x) = c\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

for $-\infty < \mu < \infty$, $\sigma > 0$ where $\mu$ (Greek letter mu) and $\sigma$ (Greek letter sigma) are parameters, and $c$ is the normalizing constant.

Let us compute the normalizing constant $c$, the mean value and the variance. First, make a substitution

$$u = \frac{x-\mu}{\sigma} \quad \Rightarrow \quad \mathrm{d}u = \frac{1}{\sigma}\mathrm{d}x.$$

The total probability is 1 and, therefore,

$$1 = c\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x = c\sigma\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u = 2c\sigma\int_{0}^{\infty} \mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u,$$

since it is an even and exponentially decaying function. Hence

$$1 = c\sigma\sqrt{2}\int_{0}^{\infty} v^{\frac{1}{2}-1}\mathrm{e}^{-v}\mathrm{d}v$$

by putting $v = \frac{u^2}{2} \Rightarrow \mathrm{d}u = \frac{\sqrt{2}}{2}v^{\frac{1}{2}-1}\mathrm{d}v$

$$1 = c\sigma\sqrt{2}\Gamma\left(\frac{1}{2}\right) = c\sigma\sqrt{2\pi} \quad \Rightarrow \quad c = \frac{1}{\sigma\sqrt{2\pi}}.$$

For computing the mean value and variance we make the same substitution, $y = \frac{x-\mu}{\sigma} \Rightarrow \mathrm{d}x = \sigma\mathrm{d}y$ and $x = \mu + \sigma y$. Then the mean value,

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \\ &= \mu\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y + \sigma\int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y \\ &= \mu + 0 = \mu \end{aligned}$$

because the first integral is the total probability in a Gaussian density with $\mu = 0$ and $\sigma^2 = 1$ and the second is an integral over an odd function, where each piece gives convergent integrals, and hence zero. Thus the parameter $\mu$ sitting in the density is the mean value of the Gaussian random variable. Now, let us compute variance, by using the following substitutions $y = \frac{x-\mu}{\sigma}$ and $u = \frac{y^2}{2}$:

$$\begin{aligned} \mathrm{Var}(x) &= E[x - E(x)]^2 = E[x-\mu]^2 \\ &= \int_{-\infty}^{\infty} \frac{[x-\mu]^2}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x = \sigma^2\int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y \end{aligned}$$

by using the substitution $y = \frac{x-\mu}{\sigma}$. Then

$$\mathrm{Var}(x) = \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^\infty y^2 \mathrm{e}^{-\frac{y^2}{2}} \mathrm{d}y$$

by using the property of even functions, and it is

$$= \sigma^2 \int_0^\infty \frac{u^{\frac{1}{2}-1}}{\sqrt{\pi}} \mathrm{e}^{-u} \mathrm{d}u = \sigma^2$$

by using the substitution $u = \frac{y^2}{2}$, by observing that the integral is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Hence the second parameter sitting in the Gaussian density is the variance, and $\sigma^2$ is the standard notation for the variance.

For convenience, we use the following standard notations:

$x \sim N(\mu, \sigma^2)$: $x$ is distributed as a Gaussian or normal distribution with the mean value $\mu$ and variance $\sigma^2$. For example, $x \sim N(3, 5)$ means normal with mean value 3 and variance 5.

$x \sim N(0, \sigma^2)$: $x$ is normally distributed with mean value zero and variance $\sigma^2$;

$x \sim N(0, 1)$: $x$ is normally distributed with mean value zero and variance unity. This is also called a *standard normal distribution* and its density will be of the form:

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}, \quad -\infty < x < \infty. \tag{6.28}$$

The graph of a $N(\mu, \sigma^2)$ will be of the form as depicted in Figure 6.8.

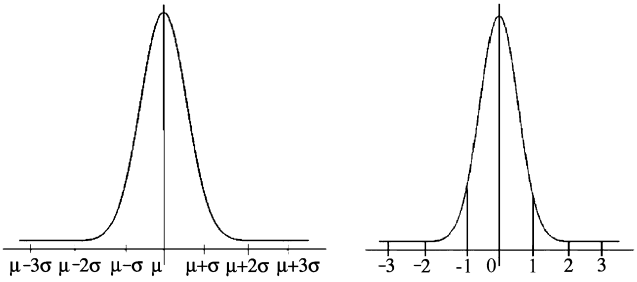


**Figure 6.8:** Left: $N(\mu, \sigma^2)$; Right: $N(0, 1)$.

In the graphs, we have marked the points $1\sigma$ or one standard deviation away from the mean value, two standard deviations away from the mean value and three standard deviations away from the mean value. These intervals are very important in decision making situations. The $N(\mu, \sigma^2)$ curve is symmetric about $x = \mu$ and there are points of inflexion at $x = \mu - \sigma$ and $x = \mu + \sigma$. The $N(0, 1)$ curve is symmetric about $x = 0$ and points of inflexion at $x = \pm 1$. Definite integrals of the form

$$\int_0^a f(x)\mathrm{d}x \quad \text{or} \quad \int_{-\infty}^a f(x)\mathrm{d}x$$

are not available because the indefinite integral $\int \mathrm{e}^{-t^2}\mathrm{d}t$ is not available. But numerical tables of the standard normal density are available. These are called normal tables. You may find tables of the type:

$$\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}}\mathrm{d}z \quad \text{or} \quad \int_{0}^{b} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}}\mathrm{d}z.$$

Probability on any interval in the standard normal case can be computed by using one of the above forms of the normal tables. For example

$$\begin{aligned}\int_{-3}^{2} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x &= \int_{-3}^{0}[\cdot]\mathrm{d}x + \int_{0}^{2}[\cdot]\mathrm{d}x \\ &= \int_{0}^{3}[\cdot]\mathrm{d}x + \int_{0}^{2}[\cdot]\mathrm{d}x\end{aligned}$$

due to symmetry and both these integrals can be read from the tables.

From the normal probability tables, we can see that approximately 65% of the probability is within one standard deviation of the mean value, approximately 95% of the area is within two standard deviations of the mean value and approximately 99% area is within three standard deviations of the mean value. As probability statements, we have the following:

$$\begin{aligned}\Pr\{\mu-\sigma<x<\mu+\sigma\} &= \Pr\{|x-\mu|\le\sigma\} = \Pr\left\{\left|\frac{x-\mu}{\sigma}\right|\le 1\right\} \\ &\approx 0.65. \end{aligned}\tag{6.29}$$

$$\begin{aligned}\Pr\{\mu-2\sigma<x<\mu+2\sigma\} &= \Pr\{|x-\mu|\le 2\sigma\} = \Pr\left\{\left|\frac{x-\mu}{\sigma}\right|\le 2\right\} \\ &\approx 0.95. \end{aligned}\tag{6.30}$$

$$\begin{aligned}\Pr\{\mu-3\sigma<x<\mu+3\sigma\} &= \Pr\{|x-\mu|\le 3\sigma\} = \Pr\left\{\left|\frac{x-\mu}{\sigma}\right|\le 3\right\} \\ &\approx 0.99. \end{aligned}\tag{6.31}$$

These three observations (6.29), (6.30), (6.31) are very important in testing of statistical hypotheses and in making "confidence statements" on the parameter $\mu$.

**Example 6.5.** It is found that the monthly incomes of working females in a city are approximately normally distributed with mean value Rs 10 000 and standard deviation Rs 2 000. (i) What is the range of incomes, around the mean value, where 95% of the working females can be found? (ii) If a working female is picked at random from this city, what is the probability that her income is between Rs 8 000 and Rs 14 000?

**Solution 6.5.** Approximately 95% of incomes, around the mean value, can be found in the range $\mu-2\sigma<x<\mu+2\sigma$ when $x$ denotes the monthly income. The range is

$$[10\,000-2\times2000, 10\,000+2\times2000] = [6\,000, 14\,000].$$

This answers (i). For (ii), we need the probability $\Pr\{8\,000 < x < 14\,000\}$. We will standardize $x$.

Standardization of a random variable $x$ means to consider the random variable $y = \frac{x-E(x)}{\sqrt{\mathrm{Var}(x)}} = \frac{x-\mu}{\sigma}$ so that the new variable $y$ is such that $E(y) = 0$ and $\mathrm{Var}(y) = 1$.

For (ii), denoting the standardized $x$ as $z$ we have

$$\begin{aligned}
&\Pr\{8\,000 < x < 14\,000\}\\
&\quad = \Pr\left\{\frac{8\,000-10\,000}{2\,000} < \frac{x-\mu}{\sigma} < \frac{14\,000-10\,000}{2\,000}\right\}\\
&\quad = \Pr\{-1 < z < 2\} = \Pr\{-1 < z < 0\} + \Pr\{0 < z < 2\}\\
&\quad = \Pr\{0 < z < 1\} + \Pr\{0 < z < 2\} \quad \text{from symmetry}\\
&\quad = 0.325 + 0.475 = 0.8 \quad \text{approximately.}
\end{aligned}$$

The probabilities are read from standard normal tables.

### 6.7.1 Moment generating function of the normal density

Since the Gaussian density is one of the most important densities in statistical and probability literature, the moment generating function in the Gaussian case is also very important. Again, using the same notation

$$M(t) = E[\mathrm{e}^{tx}] = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{tx-\frac{1}{2\sigma^2}(x-\mu)^2}\mathrm{d}x.$$

Make the transformation $y = \frac{x-\mu}{\sigma}$, then $\mathrm{d}y = \frac{1}{\sigma}\mathrm{d}x$ and the limits will remain the same, $-\infty < y < \infty$. Then $x = \mu + \sigma y$. The exponent in the integral above, simplifies to the following:

$$\begin{aligned}
tx - \frac{1}{2\sigma^2}(x-\mu)^2 &= t(\mu+\sigma y) - \frac{y^2}{2}\\
&= t\mu - \frac{1}{2}[y^2 - 2\sigma ty]\\
&= t\mu - \frac{1}{2}[y^2 - 2\sigma ty + \sigma^2t^2 - \sigma^2t^2]\\
&= t\mu - \frac{1}{2}[(y-\sigma t)^2] + \frac{t^2\sigma^2}{2}.
\end{aligned}$$

Substituting these and replacing $\frac{\mathrm{d}x}{\sigma}$ by $\mathrm{d}y$ we have

$$M(t) = \mathrm{e}^{t\mu+\frac{t^2\sigma^2}{2}}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-\frac{1}{2}(y-\sigma t)^2}}{\sqrt{2\pi}}\mathrm{d}y.$$

The integral part can be looked upon as the total probability, which is 1, from a normal density with parameters $t\sigma$ and 1. Therefore,

$$M(t) = \mathrm{e}^{t\mu + \frac{t^2\sigma^2}{2}}. \tag{6.32}$$

That is,

$$\begin{aligned} x \sim N(\mu, \sigma^2) \quad &\Rightarrow \quad M(t) = \exp\left[t\mu + \frac{t^2\sigma^2}{2}\right]. \\ x \sim N(0, \sigma^2) \quad &\Rightarrow \quad M(t) = \exp\left[\frac{t^2\sigma^2}{2}\right]. \\ x \sim N(0, 1) \quad &\Rightarrow \quad M(t) = \exp\left[\frac{t^2}{2}\right]. \end{aligned}$$

The characteristic function of the Gaussian random variable or Gaussian or normal density is obtained by replacing $t$ by $it$, $i = \sqrt{-1}$. Denoting the characteristic function by $\phi(t)$, we have

$$\phi(t) = \exp\left[it\mu - \frac{t^2\sigma^2}{2}\right] \tag{6.33}$$

for the normal or Gaussian case. When $x \sim N(0,1)$, the standard normal, then its characteristic function is

$$\phi(t) = \exp\left[-\frac{t^2}{2}\right]. \tag{6.34}$$

From the moment generating function (6.32) or from the characteristic function (6.33), one property is obvious. What is the distribution of a linear function of a normal random variable? Let $x \sim N(\mu, \sigma^2)$ and let $y = ax + b$, $a \neq 0$, where $a$ and $b$ are constants. Then the moment generating function (sometimes abbreviated as mgf) of $y$, denoted by $M_y(t)$, is given by

$$\begin{aligned} M_y(t) &= E[\mathrm{e}^{ty}] = E[\mathrm{e}^{t(ax+b)}] = \mathrm{e}^{tb} E[\mathrm{e}^{(at)x}] \\ &= \exp\left[tb + \left\{(at)\mu + (at)^2 \frac{\sigma^2}{2}\right\}\right] \\ &= \exp\left[t(a\mu + b) + \frac{t^2}{2}(a^2\sigma^2)\right], \quad a \neq 0. \end{aligned}$$

But this is the mgf of a normal variable with parameters $a\mu + b = E[y] = E[ax + b]$ and $a^2\sigma^2 = \mathrm{Var}(y) = \mathrm{Var}(ax + b)$. Therefore, every linear function of a normal variable is again a normal variable.

**Result 6.1.** *If $x \sim N(\mu, \sigma^2)$ then $y = ax + b$, $a \neq 0$, $\sim N(a\mu + b, a^2\sigma^2)$.*

**Note 6.1.** When $a = 0$, then the mgf is $\mathrm{e}^{tb}$, which is the mgf of a degenerate random variable with the whole probability mass 1 at the point $x = b$ (Observe that $b$ can be zero also).

For example,

$$\begin{aligned} x \sim N(\mu = 2, \sigma^2 = 5) &\quad\Rightarrow\quad y = -3x + 4 \sim N(-2, 45). \\ x \sim N(\mu = -4, \sigma^2 = 1) &\quad\Rightarrow\quad y = -2x + 7 \sim N(15, 4). \\ x \sim N(\mu = 0, \sigma^2 = 1) &\quad\Rightarrow\quad y = 3x + 4 \sim N(4, 9). \end{aligned}$$

Let us see what happens if we take the $n$-th power of this mgf for a normal random variable. That is,

$$\begin{aligned} M(t) &= \mathrm{e}^{t\mu + \frac{t^2\sigma^2}{2}} \quad \Rightarrow \\ [M(t)]^n &= [\mathrm{e}^{t\mu + \frac{t^2\sigma^2}{2}}]^n = \mathrm{e}^{tn\mu + \frac{t^2}{2}n(\sigma)^2}. \end{aligned}$$

But this is the mgf of a normal or Gaussian variable with the parameters $n\mu$ and $n\sigma^2$. If the corresponding random variable is denoted by $u$ then $u \sim N(n\mu, n\sigma^2)$. This is a property associated with "infinite divisibility" of a random variable, which will be discussed after considering independence of random variables in the next chapter. Observe that if $M_x(t)$ is the mgf of $x$ then $[M_x(t)]^n$ does not mean the mgf of $x^n$.

$$[M_x(t)]^n \neq M_{x^n}(t).$$

We have seen the infinite divisibility property for a gamma random variable also. If $z$ is a two parameter gamma random variable, then we have seen that its mgf, denoted by $M_z(t)$, is $(1 - \beta t)^{-\alpha}$. Therefore,

$$[M_z(t)]^n = [(1 - \beta t)^{-\alpha}]^n = (1 - \beta t)^{-n\alpha},$$

which is the mgf of a gamma variable with the shape parameter $n\alpha$ and scale parameter $\beta$.

## 6.8 Transformation of variables

Here, we consider the problem of finding the probability or density of a function $g(x)$ of a random variable $x$, given the probability or density function of the random variable $x$. As an example, suppose that we know that the waiting time at a certain queue is exponentially distributed with the expected waiting time one hour. Suppose that for a person waiting at this queue it costs him Rs 500 per hour of time lost plus the transportation cost of Rs 40. This means, if $t$ is the actual waiting time then his loss is $g(t) = 40 + 500t$. Knowing the distribution of $t$ [known to be exponential here] we want the distribution of $40 + 500t$. As another example, suppose that a working girl is appearing for an interview for promotion. If $x$ is the number of correct answers given, then her salary is likely to be $x^2$ + Rs 2 000 (fringe benefits at the next position). Here, $x$ is a binomial random variable and we want the distribution of $y = 2\,000 + x^2$. Problems of this type will be examined here. First, let us examine discrete cases. If the probability function of $x$ is given and if we need the probability function of $y = g(x)$, some function of $x$, then the problem is answered if we can compute the probability for each value $g(x)$ takes, knowing the probability function of $x$. Substitute the values of $x$, for which there are non-zero probabilities, into $g(x)$ and evaluate the corresponding probabilities. Then we have the answer. This will be illustrated with an example.

**Example 6.6.** Suppose $x$ has the probability function:

$$f(x) = \begin{cases} 0.25, & x = -1 \\ 0.25, & x = 1 \\ 0.5, & x = 0 \\ 0, & \text{elsewhere.} \end{cases}$$

Compute the probability function of (i) $y = x^2$; (ii) $y = 3 + 2x + 5x^2$.

**Solution 6.6.** (i) When $x = -1$, $y = x^2 = 1$ with probability 0.25. When $x = 1$, $y = x^2 = 1$ with probability 0.25. No other $x$-value gives $y = 1$. Hence the probability associated with $y = 1$ is $0.25 + 0.25 = 0.5$. When $x = 0$, $y = x^2 = 0$ with probability 0.5. This completes the computations and hence the probability function of $y$, denoted by $h_1(y)$ is the following:

$$h_1(y) = \begin{cases} 0.5, & y = 1 \\ 0.5, & y = 0 \\ 0, & \text{elsewhere.} \end{cases}$$

For (ii) also the procedure is the same. For $x = -1$, $y = 3 + 2x + 5x^2 = 3 + 2(-1) + 5(-1)^2 = 6$ with probability 0.25. When $x = 1$, $y = 3 + 2x + 5x^2 = 3 + 2(1) + 5(1)^2 = 10$ with probability 0.25. When $x = 0$, $y = 3 + 2x + 5x^2 = 3 + 2(0) + 5(0)^2 = 3$ with probability 0.5. Hence the probability function of $y$, denoted by $h_2(y)$, is the following:

$$h_2(y) = \begin{cases} 0.25, & y = 6 \\ 0.25, & y = 10 \\ 0.5, & y = 3 \\ 0, & \text{elsewhere.} \end{cases}$$

Whatever be the function $g(x)$ of $x$ the procedure is the same in the discrete case.

For the continuous case, the procedure is different. We have to look into the Jacobian of the transformation, which means that we need a one-to-one function. Let $x$ be a continuous random variable and let $g(x)$ be a one-to-one function of $x$ over a given interval. Some situations are shown in Figure 6.9. In (a), we have an increasing function, which is one-to-one. In (b), we have a decreasing function, which is one-to-one.

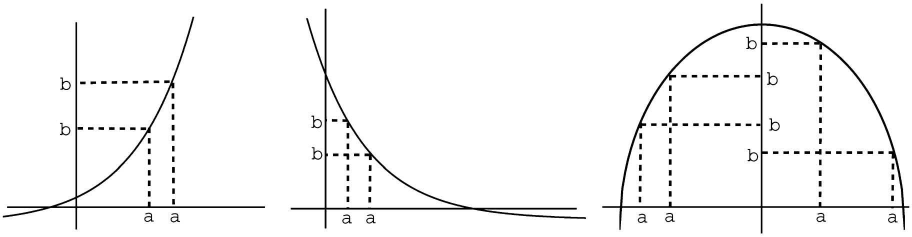


**Figure 6.9:** (a) increasing, (b) decreasing, (c) increasing/decreasing in each sub-interval.

In (c), we can subdivide the interval into two pieces where in each piece the function is one-to-one, and hence we can apply the procedure for each piece separately.

Let the distribution functions (cumulative density) of $x$ and $y = g(x)$ be denoted by $F_x(x)$ and $F_y(y)$, respectively. Then

$$\Pr\{x \le a\} = F_x(a) \quad \text{and} \quad \Pr\{y \le b\} = F_y(b).$$

If the function $y = g(x)$ is increasing as in Figure 6.9 (a), then as the point on the $x$-axis moves to the right or as $x$ increases the corresponding point on the $y$-axis moves up or $y$ also increases. In this case,

$$\Pr\{x \le a\} = \Pr\{y \le b = g(a)\} = \Pr\{y \le g(a)\} \quad \Rightarrow \quad F_x(a) = F_y(g(a)). \tag{6.35}$$

We can differentiate to get the density function. Observe that $\frac{\mathrm{d}}{\mathrm{d}x}F_x(x) = f_1(x)$ where $f_1(x)$ is the density function of $x$ and $\frac{\mathrm{d}}{\mathrm{d}y}F_y(y) = f_2(y)$ where $f_2(y)$ is the density of $y$. Let us differentiate (6.35) on both sides with respect to $a$. Then on the left side we should get the density of $x$ evaluated at $x = a$. That is,

$$\begin{aligned}
f_1(a) &= \frac{\mathrm{d}}{\mathrm{d}a}F_x(a) = \frac{\mathrm{d}}{\mathrm{d}a}F_y(g(a)) \\
&= \frac{\mathrm{d}}{\mathrm{d}g}F_y(g(a)) \times \frac{\mathrm{d}}{\mathrm{d}a}g(a) \quad \Rightarrow \\
f_1(a) &= f_2(g(a))g'(a) \quad \Rightarrow \\
f_1(x) &= f_2(y) \times \frac{\mathrm{d}y}{\mathrm{d}x}.
\end{aligned} \tag{6.36}$$

This is the connection between the densities of $x$ and $y$ when $y$ is an increasing function of $x$. Now, let us see what happens if $y$ is a decreasing function of $x$ as in Figure 6.9 (b). Observe that when a point is moving to the right on the $x$-axis (increasing) the corresponding point on the $y$-axis is decreasing. Hence the connection between the probabilities is the following:

$$\begin{aligned}
&\Pr\{x \le a\} = \Pr\{y \ge b = g(a)\} = 1 - \Pr\{y \le g(a)\} \quad \Rightarrow \\
&F_x(a) = 1 - F_y(g(a)).
\end{aligned}$$

Differentiating both sides with respect to $a$ we have

$$\begin{aligned} f_1(a) &= \frac{\mathrm{d}}{\mathrm{d}a}F_x(a) = -\frac{\mathrm{d}}{\mathrm{d}a}F_y(g(a)) \\ &= -\frac{\mathrm{d}}{\mathrm{d}g}F_y(g(a)) \times \frac{\mathrm{d}}{\mathrm{d}a}g(a) \quad \Rightarrow \\ f_1(a) &= -f_2(g(a))\frac{\mathrm{d}}{\mathrm{d}a}g(a). \end{aligned}$$

That is,

$$f_1(x) = -f_2(y)\frac{\mathrm{d}y}{\mathrm{d}x}. \tag{6.37}$$

Thus, when $y$ is a decreasing function of $x$ then the formula is (6.37). Note that when $y$ is decreasing, $\frac{\mathrm{d}y}{\mathrm{d}x}$ will be negative, and thus the right side will be positive all the time. Thus, in general, the formula is (6.36), and when $y$ is decreasing then multiply the right side by –1. In practical computations, this minus sign is automatically taken care of in the limits of integration and hence the formula to be remembered is (6.36) and take the correct limits of integration.

**Example 6.7.** Let $x$ be a continuous random variable with density $f(x)$ and distribution function $F_x(x)$, which is a function of $x$. Consider the transformation $y = F_x(x)$ [This transformation is called the *probability integral transformation*, which is the basis for the area called *statistical simulation* and also the basis for generating *random numbers* or taking a random sample from a given distribution. Evaluate the density of $y$.

**Solution 6.7.** This is a one-to-one transformation. $y$ is an increasing (non-decreasing) function of $x$. Applying the formula (6.36), we have

$$\begin{aligned} f_1(x) &= f_2(y)\frac{\mathrm{d}y}{\mathrm{d}x} \quad \Rightarrow \\ f_1(x) &= f_2(y)\frac{\mathrm{d}}{\mathrm{d}x}F_x(x) = f_2(y)f_1(x) \quad \Rightarrow \\ 1 &= f_2(y). \end{aligned}$$

That is, $y$ is uniformly distributed on the interval $[0,1]$. Thus, through a probability integral transformation any density can be brought to a uniform density over $[0,1]$. This is the importance of this transformation.

**Example 6.8.** Let $x$ be exponentially distributed with mean value 5. Let $y = 2 + 3x$. Compute the density of $y$.

**Solution 6.8.** When $x$ goes from 0 to $\infty$, $y = 2 + 3x$ is an increasing function of $x$. Hence we can use the formula (6.36). $\frac{\mathrm{d}y}{\mathrm{d}x} = 3$ and $x = \frac{y-2}{3}$. Also when $x = 0$, $y = 2$ and when $x \to \infty$, $y \to \infty$. Therefore,

$$f_1(x) = f_2(y) \times 3 \quad \Rightarrow \quad f_2(y) = \frac{1}{3} f_1\left(\frac{y-2}{3}\right) = \frac{1}{15} \mathrm{e}^{-\frac{(y-2)}{15}} \quad \Rightarrow$$

$$f_2(y) = \begin{cases} \frac{1}{15} \mathrm{e}^{-\frac{(y-2)}{15}}, & 2 \le y < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

Thus $y$ is a relocated re-scaled exponential random variable.

**Example 6.9.** Let $x$ be a standard normal variable and let $y = x^2$. Compute the density of $y$.

**Solution 6.9.** Here, $x$ goes from $-\infty$ to $\infty$, and hence $y = x^2$ is not a one-to-one function in the whole range. But in the interval $-\infty < x < 0$ (the function is strictly decreasing), and in the interval $0 < x < \infty$ (the function is strictly increasing) the function $y = x^2$ is one-to-one in each of these intervals. Hence we can apply formula (6.36) in the interval $0 < x < \infty$ and (6.37) in the other interval $-\infty < x < 0$. The curve $y = x^2$ is shown in Figure 6.10.

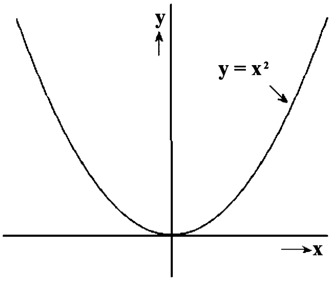


**Figure 6.10:** $y = x^2$.

For positive $x$, $y = x^2 \Rightarrow x = y^{\frac{1}{2}} \Rightarrow \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{2} y^{-\frac{1}{2}}$. Let the piece of the standard normal density in the interval $0 < x < \infty$ be denoted by $f_{11}(x)$, that is,

$$f_{11}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}, & 0 \le x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

so that $f_1(x) = f_{11}(x) + f_{12}(x)$, where $f_{12}(x)$ is the corresponding piece of $N(0,1)$ density over $-\infty < x < 0$, and the corresponding piece of the density of $y$ be denoted by $f_{21}(y)$. Then from (6.36)

$$\begin{aligned} f_{21}(y) &= \frac{1}{2} y^{\frac{1}{2}-1} \frac{\mathrm{e}^{-\frac{y}{2}}}{\sqrt{2\pi}}, \quad 0 \le y < \infty \\ &= \frac{1}{2} \frac{1}{2^{\frac{1}{2}} \Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} \mathrm{e}^{-\frac{y}{2}}. \end{aligned}$$

From the symmetry of $y$ and due to $f_1(x)$ being an even function, $f_{22}(y)$ corresponding to $f_{12}(x)$ also gives exactly the same function from (6.37). Hence

$$f_2(y) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} \mathrm{e}^{-\frac{y}{2}}, & 0 \le y < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

But this is a gamma density with parameters $\alpha = \frac{1}{2}$ and $\beta = 2$ or it is a chi-square density with one degree of freedom; see Section 6.3.2.

**Result 6.2.** *When $x$ is standard normal, then $y = x^2$ is a chi-square with one degree of freedom or*

$$x \sim N(0,1) \quad \Rightarrow \quad y = x^2 \sim \chi_1^2.$$

Another method of showing that when $x \sim N(0,1)$ then $y = x^2 \sim \chi_1^2$ or a chi-square variable with one degree of freedom is to use the distribution function of $y$ itself. The distribution function of $y$, denoted by $F_y(z) = \Pr\{y \le z, z > 0\}$ is such that $g_y(z) = \frac{\mathrm{d}}{\mathrm{d}z} F_y(z)$ where $g_y(z)$ is the density of $y$, evaluated at $y = z$. Note that

$$\begin{aligned} F_y(z) &= \Pr\{y \le z, z > 0\} = \Pr\{x^2 \le z\} = \Pr\{|x| \le \sqrt{z}\} \\ &= \Pr\{-\sqrt{z} \le x \le \sqrt{z}\} = F_x(\sqrt{z}) - F_x(-\sqrt{z}) \end{aligned}$$

where $F_x(\cdot)$ is the distribution function of $x$, and the density of $x$, denoted by $f_x(z) = \frac{\mathrm{d}}{\mathrm{d}z} F_x(z)$. Therefore, differentiating the above with respect to $z$ we have

$$\begin{aligned} g_y(z) &= \frac{\mathrm{d}}{\mathrm{d}z} F_y(z) = \frac{\mathrm{d}}{\mathrm{d}z} [F_x(\sqrt{z}) - F_x(-\sqrt{z})] \\ &= f_x(\sqrt{z}) \frac{1}{2} z^{\frac{1}{2}-1} + f_x(-\sqrt{z}) \frac{1}{2} z^{\frac{1}{2}-1} = \frac{z^{\frac{1}{2}-1}}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z}{2}} \\ &= \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} z^{\frac{1}{2}-1} \mathrm{e}^{-\frac{z}{2}}, \quad 0 \le z < \infty \end{aligned}$$

and zero elsewhere, which is the density of a chi-square random variable with one degree of freedom or a gamma variable with the parameters $(\alpha = \frac{1}{2}, \beta = 2)$.

**Example 6.10.** Let $x \sim N(0,1)$ and let $y = 5x^2 - 3$. Compute the density of $y$.

**Solution 6.10.** Let $u = x^2$. Then we have from Example 6.9 that $u$ is a chi-square variable with one degree of freedom. Let the density of $u$ be $f_u(u)$. Then

$$f_u(u) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} u^{\frac{1}{2}-1} \mathrm{e}^{-\frac{u}{2}}, & 0 \le u < \infty \\ 0, & \text{elsewhere} \end{cases}$$

But $y$ to $u$ is a one to one transformation. $y = 5u - 3 \Rightarrow u = \frac{y+3}{5}$, $\mathrm{d}y = 5\mathrm{d}u$, $0 \le u < \infty \Rightarrow -3 \le y < \infty$. Therefore, if the density of $y$ is denoted by $f_y(y)$ then

$$f_y(y) = \begin{cases} \frac{1}{(5)2^{\frac{1}{2}}\Gamma(\frac{1}{2})}(\frac{y+3}{5})^{\frac{1}{2}-1}\mathrm{e}^{-\frac{1}{2}(\frac{y+3}{5})}, & -3 \le y < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

## 6.9 A note on skewness and kurtosis

Skewness is often misinterpreted as asymmetry in a distribution. Skewness is associated with the median. In a continuous case, $\Pr\{x \le M\} = \Pr\{x \ge M\} = \frac{1}{2}$ or probability to the left of the median point $M$ is the same as the probability to the right of $M$. If the range of $x$ to the right of $M$ is not equal to the range of $x$ to the left of $M$, then there is possibility of skewness. If the probability 0.5 on one side of the point $M$ is stretched out compared to the other side, then the density is skewed to the stretched outer side. A density curve can be asymmetric but need not be skewed. Some possibilities are marked in Figure 6.11.

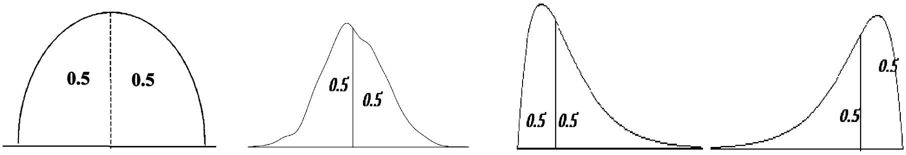


**Figure 6.11:** Symmetric, asymmetric, skewed to right, skewed to left densities.

A scale-free measure based on the third central moment, such as

$$s = \frac{\mu_3}{[\mu_2]^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3} \tag{6.38}$$

where $\mu_3 = E[x - E(x)]^3$, $\mu_2 = E[x - E(x)]^2 = \sigma^2$, is often used to measure skewness. But $s$ can only measure asymmetry rather than skewness. But

$$s_1 = \frac{E[x - M]^3}{E[x - M]^{2\frac{3}{2}}} \tag{6.39}$$

can measure skewness to some extent where $M$ is the median. If $s_1 > 0$, then one may say that the density is skewed to the right and if $s_1 < 0$ then skewed to the left. We shall not elaborate on this aspect further because the measures $s$ or $s_1$ is not a unique property associated with any shape. Kurtosis has something to do with peakedness or flatness of a density curve or probability function. When we say more flat or more peaked then there has to be a standard item to compare with. The normal (Gaussian) density curve is taken as the standard for comparison purposes. For a Gaussian density when we compute the ratio,

$$k = \frac{E[x - E(x)]^4}{[E(x - E(x))^2]^2} = \frac{\mu_4}{\mu_2^2} \tag{6.40}$$

then it is $k = 3$ for the Gaussian case. Hence the comparison is made with this number 3. For a given distribution if $k > 3$, then we say that the distribution if lepto-kurtic (more peaked) and if $k < 3$, then we say that the distribution is plati-kurtic (more flat) as shown in Figure 6.12.

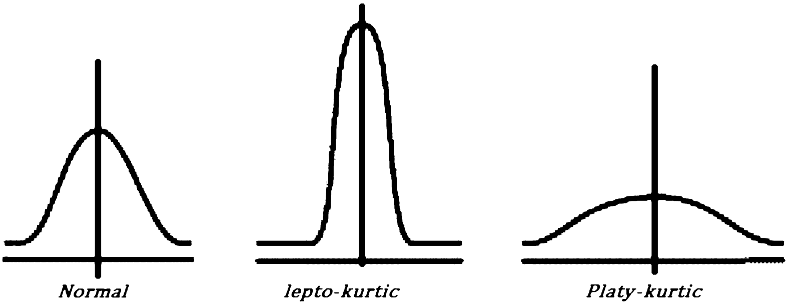


**Figure 6.12:** Left: normal; Middle: lepto-kurtic; Right: plati-kurtic.

But the items $s, s_1, k$ given above are not characterizing quantities for distributions. In other words, these measures do not uniquely determine (characterize) distributions or shapes. Hence no unique conclusions can be made by using these measures. We have already seen that a property such as mean value $\mu$ being equal to variance $\sigma^2$ is enjoyed by the Poisson random variable as well as by a gamma random variable with the scale parameter $\beta = 1$. Hence it is not a characteristic property of any random variable. Similarly, for a Gaussian random variable $k$ in (6.40) is 3. But $k = 3$ is not a characteristic property of the Gaussian density. Hence the concepts behind it and comparison with 3 do not have much significance. Hence, nowadays, the students are unlikely to find discussion of skewness and kurtosis in modern probability/statistics books.

**Note 6.2.** In Chapter 5, we derived most of the discrete probability functions by looking at experimental situations satisfying some conditions. When it came to densities, for continuous random variables, we could not list experimental situations, other than for the case of uniform distribution. Are there experimental situations from where densities could be derived? The answer is in the affirmative. The derivation of the densities from the basic assumptions to the final densities involve the concepts of joint distributions, statistical independence, etc. of random variables. Hence we will consider such problems after discussing joint distributions.

## 6.10 Mathai's pathway model

A very general density with a switching mechanism, introduced by Mathai [5] has the following form in the particular case of real scalar variables: [The above paper [5] is on rectangular matrix variate functions.]

$$g_x(x) = c|x|^{\gamma-1}[1-a(1-q)|x|^{\delta}]^{\frac{\eta}{1-q}} \tag{6.41}$$

for $a > 0$, $\delta > 0$, $\eta > 0$, $\gamma > 0$, $-\infty < x < \infty$, and $1 - a(1-q)|x|^{\delta} > 0$, and zero elsewhere, where $c$ is the normalizing constant. A particular case of (6.41) for $x > 0$ is the following:

$$g_1(x) = c_1 x^{\gamma-1}[1-a(1-q)x^{\delta}]^{\frac{\eta}{1-q}} \tag{6.42}$$

for $a > 0$, $\delta > 0$, $\gamma > 0$, $\eta > 0$, $1 - a(1-q)x^{\delta} > 0$. Here, $x$ will be in a finite range with a non-zero function for $q < 1$. Observe that when $q < 1$ then the density in (6.42) stays in the generalized type-1 beta family. Generalized type-1 beta in the sense that if $y$ is type-1 beta as described in Section 6.5 then consider a transformation $y = a(1-q)x^{\delta}$ then the density of $x$ will reduce to the form in (6.42).

When $q > 1$, then $1 - q = -(q-1)$ and the density, denoted by $g_2(x)$, has the form

$$g_2(x) = c_2 x^{\gamma-1}[1+a(q-1)x^{\delta}]^{-\frac{\eta}{q-1}} \tag{6.43}$$

for $0 \le x < \infty$, $a > 0$, $\delta > 0$, $\eta > 0$, $q > 1$, $\gamma > 0$, and zero elsewhere, where $c_2$ is the normalizing constant. The form in (6.43) is a generalized type-2 beta family in the sense if $y$ is type-2 beta as in Section 6.5 then consider a transformation $y = a(q-1)x^{\delta}$ then $x$ will have the density of the form in (6.43).

Now considering the limiting process of $q$ going to 1 either from the left or from the right. From the property, coming from the definition of the mathematical constant e, that

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \mathrm{e}^x$$

we have

$$\lim_{q\to 1_-} g_1(x) = \lim_{q\to 1_+} g_1(x) = g_3(x)$$

where

$$g_3(x) = c_3 x^{\gamma-1}\mathrm{e}^{-a\eta x^{\delta}}, \tag{6.44}$$

for $0 \le x < \infty$, $a > 0$, $\eta > 0$, $\delta, \gamma > 0$ and zero elsewhere. This is the generalized gamma density. In the pathway model, $q$ is called the *pathway parameter* because through $q$ one can go from a generalized type-1 beta family of densities to a generalized type-2 beta family of densities to a generalized gamma family. Thus all these families of functions are connected through this pathway parameter $q$. By making the substitution $y = a(1-q)x^{\delta}$ in (6.42), $y = a(q-1)x^{\delta}$ in (6.43) and $y = ax^{\delta}$ in (6.44) and then integrating out by using type-1 beta, type-2 beta and gamma integrals respectively, one can compute the normalizing constants $c_1, c_2, c_3$. [This evaluation is given as an exercise to the students.]

### 6.10.1 Logistic model

A very popular density in industrial applications is logistic model or logistic density. Let us denote it by $f(x)$.

$$f(x) = \frac{\mathrm{e}^{x}}{(1+\mathrm{e}^{x})^2} = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2}, \quad -\infty < x < \infty. \tag{6.45}$$

Note that the shape of the curve corresponds to a Gaussian density but with thicker tails at both ends. In situations where the tail probabilities are bigger than the corresponding areas from a standard normal density, then this logistic model is used. We will look at some interesting connections to other densities. Let us consider a type-2 beta density with the parameters $\alpha > 0$ and $\beta > 0$ or with the density

$$g(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1+x)^{-(\alpha+\beta)}, \quad 0 \le x < \infty,\ \alpha > 0,\ \beta > 0 \tag{6.46}$$

and zero elsewhere. Let us make the transformation $x = \mathrm{e}^{y}$ or $x = \mathrm{e}^{-y}$ then $-\infty < y < \infty$ and the density of $y$, denoted by $g_1(y)$, is given by

$$g_1(y) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\mathrm{e}^{\alpha y}}{(1+\mathrm{e}^{y})^{\alpha+\beta}} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\mathrm{e}^{-\beta y}}{(1+\mathrm{e}^{-y})^{\alpha+\beta}} \tag{6.47}$$

for $\alpha > 0$, $\beta > 0$, $-\infty < y < \infty$. Note that for $\alpha = 1$, $\beta = 1$, (6.47) reduces to (6.45) the logistic density. This (6.47) is the generalized logistic density introduced by this author and his co-workers and available from a type-2 beta density by a simple transformation. If we put $x = \frac{m}{n}z$, $\alpha = \frac{m}{2}$, $\beta = \frac{n}{2}$ in (6.46), then the density of $z$ is the F-density or variance ratio density. If a power transformation is used in (6.46), that is, if we replace $x$ by $at^{\rho}$, $\rho > 0$, $a > 0$, then (6.46) will lead to a particular case of the pathway density in Section 6.10 for $q > 1$. For all the models described in Sections 6.1 to 6.11, one can look at power transformations and exponentiation. That is, replace $x$ by $ay^{\rho}$, $a > 0, \rho > 0$ or $x$ by $\mathrm{e}^{-y}$ , then we end up with very interesting models which are useful when models are constructed for given data, see the effects of such transformations from [7]. There are other classes of densities associated with Mittag–Leffler functions. These functions naturally arise in fractional calculus, especially in the solutions of fractional differential equations. Such models may be seen from [6].

## 6.11 Some more commonly used density functions

In the following list, only the non-zero parts of the densities are given. It is understood that the functions are zeros outside the ranges listed therein.

$$f_{11}(x) = \frac{4}{\pi}\sin^{-1}\sqrt{x}, \quad 0 < x < 1$$

[Arc-sine density].

$$f_{12}(\theta) = \frac{2\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)}(\sin\theta)^{2p_1-1}(\cos\theta)^{2p_2-1}$$

for $0 < \theta < \pi/2$, $p_1 > 0$, $p_2 > 0$ [Beta type-1 polar density].

$$f_{13}(x) = cx^{\frac{s(2m+s+1)}{2}-1}\left(1-\frac{x}{s}\right)^{\frac{s(2n+s-1)}{2}-1}$$

for $0 < x < s$, $m > 0$, $n > 0$,

$$c = \frac{\Gamma(s(2m+n+s+1)}{s^{\frac{s}{2}(2m+s+1)}\Gamma(\frac{s(2m+s+1)}{2})\Gamma(\frac{s(2n+s+1)}{2})}$$

[Beta type-1, three-parameter density].

$$\begin{aligned} f_{15}(x) = {} & \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mathrm{e}^{-\frac{\lambda^2}{2}}x^{\alpha-1}(1-x)^{\beta-1} \\ & \times {}_1F_1\left(\alpha+\beta;\alpha;\frac{\lambda^2 x}{2}\right) \end{aligned}$$

for $0 < x < 1$, $\lambda > 0$, $\alpha > 0$, $\beta > 0$ and ${}_1F_1$ is a confluent hypergeometric function [Beta type-1 non-central density].

$$f_{16}(x) = cx^{\frac{s(2m+s+1)}{2}-1}\left(1+\frac{x}{s}\right)^{-\frac{s(2m+2n+s+1)}{2}-1}$$

for $0 \le x < \infty$, $s > 0$, $m > 0$, $n > 0$,

$$c = \frac{\Gamma(\frac{s(2m+2n+s+1)}{2}+1)}{s^{\frac{s(2m+s+1)}{2}}\Gamma(\frac{s(2m+s+1)}{2})\Gamma(sn+1)}$$

[Beta type-2 three-parameter density].

$$\begin{aligned} f_{17}(x) = {} & \mathrm{e}^{-\frac{\lambda^2}{2}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1+x)^{-(\alpha+\beta)} \\ & \times {}_1F_1\left(\alpha+\beta;\alpha;\frac{\lambda^2}{2}\frac{x}{1+x}\right) \end{aligned}$$

for $0 \le x < \infty$, $\alpha > 0$, $\beta > 0$, $\lambda > 0$ [Beta type-2 non-central density].

$$f_{18}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{x^{\alpha-1}b^{\beta}}{(x+b)^{\alpha+\beta}}$$

for $\alpha > 0, \beta > 0, b > 0, 0 \le x < \infty$ [Beta type-2 inverted density].

$$f_{19}(x) = [2\pi I_0(k)]^{-1} e^{k\cos 2(x-\beta)}$$

for $0 < x \le \pi$, $0 \le \beta \le \pi$, $0 \le k < \infty$ where $I_0(k)$ is a Bessel function [Circular normal density; also see bimodal density].

$$f_{20}(x) = \frac{(1-\rho)^{\frac{1}{2}}}{2\pi(1-\rho\sin 2x)}$$

for $\rho^2 < 1, 0 < x \le \pi$ [Bimodal density].

$$f_{21}(x) = \frac{1}{c[-1+\exp(\alpha+\beta x)]}$$

for $0 \le x < \infty, \beta > 0, e^{\alpha} > 1, c = \frac{1}{\beta}\ln(\frac{e^{\alpha}}{e^{\alpha}-1})$ [Bose–Einstein density].

$$f_{22}(x) = \frac{\Delta}{\pi[\Delta^2 + (x-\mu)^2]}$$

for $-\infty < x < \infty, \Delta > 0, -\infty < \mu < \infty$ [Cauchy density].

$$f_{23}(x) = \frac{(1-\sigma^2)}{2\pi[1+\sigma^2 - 2\sigma\cos x]}$$

for $0 < x \le 2\pi, 0 < \sigma \le 1$ [Cauchy wrapped up density].

$$f_{24}(x) = \frac{2}{\pi[1+x^2]}$$

for $0 < x < \infty$ [Cauchy folded density].

$$f_{25}(x) = \frac{2(\frac{n}{2})^{\frac{n}{2}}}{\sigma^n\Gamma(\frac{n}{2})} x^{n-1} e^{-\frac{nx^2}{2\sigma^2}}$$

for $x \ge 0, \sigma > 0$, $n$ a positive integer [Chi density].

$$f_{26}(x) = e^{-\frac{\mu^2}{2\sigma^2}} \sum_{r=0}^{\infty} \frac{1}{r!}\left(\frac{\mu^2}{2\sigma^2}\right) x^{\frac{k}{2}+r-1} \frac{e^{-\frac{x}{2}}}{\Gamma(\frac{k}{2}+r)}$$

for $x \ge 0, -\infty < \mu < \infty, \sigma > 0$, $k$ a positive integer [Non-central chi-square density; $\mu = 0$ gives chi-square density; see also gamma density].

$$f_{27}(x) = \frac{x^{p-1}}{a^p\Gamma(p)} e^{-\frac{x}{a}}$$

for $x \ge 0, a > 0$, $p$ a positive integer [Erlang density; see also gamma density].

$$f_{28}(x) = \frac{\alpha}{2}\mathrm{e}^{-\alpha|x|}, \quad -\infty < x < \infty,\ \alpha > 0$$

[Exponential – bilateral density or Laplace density ; with $x$ replaced by $x - c$ we get exponential-double density].

$$f_{29}(x) = \alpha^{a}(\alpha+1)^{p-a}[\Gamma(p)]^{-1}x^{p-1}\exp[-(\alpha+1)x]{}_1F_1(a;p;x)$$

for $0 \le x < \infty$, $a > 0$, $p > 0$, $\alpha > 0$ [Exponential – generalized density].

$$f_{30}(x) = \frac{\mathrm{e}^{-x}}{(\mathrm{e}^{-a} - \mathrm{e}^{-b})}$$

for $a \le x \le b$, $a > 0$, $b > 0$ [Exponential – truncated density].

$$f_{31}(x) = \frac{1}{\beta}\mathrm{e}^{-\frac{1}{\beta}(x-\lambda)} \quad \text{for } \lambda \le x < \infty,$$

$0 < \beta < \infty$, $-\infty < \lambda < \infty$ [Exponential – two-parameter density].

$$f_{32}(x) = \frac{1}{\beta}\exp[-y - \exp(-y)] \quad \text{for } -\infty < x < \infty,\ 0 < \beta < \infty,$$

$-\infty < \lambda < \infty$, $y = \frac{x-\lambda}{\beta}$ [Extreme value, first asymptotic density].

$$f_{33}(x) = \frac{k}{v}\left(\frac{v}{x}\right)^{k+1}\exp\left[-\left(\frac{v}{x}\right)^{k}\right]$$

for $x \ge 0$, $v, k > 0$ [Extreme value, second asymptotic density].

$$f_{34}(x) = \frac{k}{(-v)}\left(\frac{x}{v}\right)^{k-1}\mathrm{e}^{-(\frac{x}{v})^{k}}$$

for $x < 0$, $v > 0$, $k > 1$ [Extreme value, third asymptotic density].

$$f_{35}(x) = \frac{1}{\lambda}\exp\left[-\left(\frac{\exp(x) - 1}{\lambda}\right) + x\right]$$

for $x \ge 0$, $\lambda > 0$ [Extreme value, modified density].

$$f_{36}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{\frac{m}{2}}\frac{x^{\frac{m}{2}-1}}{(1+\frac{m}{n}x)^{\frac{m+n}{2}}}$$

for $x \ge 0$, $m, n$ positive integers [Fisher's $F$-density].

$$f_{37}(x) = \mathrm{e}^{-\frac{\lambda^2}{2}}\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\frac{y^{\frac{m}{2}-1}}{(1+y)^{\frac{m+n}{2}}{}_1F_1(\frac{m+n}{2};\frac{m}{2};\frac{\lambda^2}{2}\frac{y}{1+y})}$$

for $x \ge 0$, $\lambda > 0$, $m, n$ positive integers, $x = \frac{n}{m}y$ [F–non-central density].

$$f_{37}(x) = \left[c\{1 + \exp(\alpha + \beta x)\}\right]^{-1}$$

for $0 \le x < \infty$, $\alpha \ne 0$, $\beta > 0$, $c = \ln(\frac{1+e^{\alpha}}{e^{\alpha}})$ [Fermi–Dirac density].

$$f_{38}(x) = \frac{k}{2\sinh k} e^{k\cos\theta} \sin\theta,$$

for $0 < \theta \le \pi$, $k > 0$ [Fisher's angular density].

$$f_{39}(x) = \frac{\sqrt{k}}{\sqrt{2\pi}} \frac{\exp[-\frac{k}{2}(b - \ln(x-a))^2]}{(x-a)} \quad \text{for } k > 0,\ a < x < \infty,\ a > 0,$$

$-\infty < b < \infty$ [Galton's density or Log-normal density].

$$f_{40}(x) = \sum_{j=1}^{\infty} c_j H_j(x)\phi(x)$$

for $-\infty < x < \infty$, where $\phi(x)$ is the standard normal density and $H_j(x)$ is the Chebyshev–Hermite polynomial of degree $j$ in $x$ defined by $(-1)^j \frac{d^j}{dx^j}\phi(x) = H_j(x)\phi(x)$ and $c_j = \frac{1}{j!}\int_{-\infty}^{\infty} H_j(x)f(x)dx$ [Gram–Charlier type A density].

$$f_{41}(x) = \left(\frac{mg}{KT}\right) e^{-\frac{(mgx)}{(KT)}}$$

for $x \ge 0$, $m > 0$, $g > 0$, $T > 0$ [Helley's density].

$$f_{42}(x) = \frac{n^{\frac{1}{2}(n-1)}}{\sigma 2^{\frac{1}{2}(n-3)}\Gamma(\frac{n-1}{2})} \left(\frac{x}{\sigma}\right)^{n-2} e^{-(nx^2/(2\sigma^2))}$$

for $0 \le x < \infty$, $\sigma > 0$, $n$ a positive integer [Helmert density].

$$f_{43}(x) = [\pi \cosh x]^{-1}$$

for $-\infty < x < \infty$ [Hyperbolic cosine density].

$$f_{44}(x) = p^{-x}$$

for $e^{-p} \le x \le 1$, $p > 0$ [Hyperbolic truncated density].

$$f_{45}(x) = \frac{(\lambda)^{\frac{1}{2}}}{(2\pi x^3)^{\frac{1}{2}}} \exp\left[-\frac{\lambda(x-\mu)^2}{(2x\mu)^2}\right]$$

for $x > 0$, $\lambda, \mu > 0$ [Inverse Gaussian density].

$$f_{46}(x) = \frac{\exp[-\frac{1}{\beta}(x-\alpha)]}{\beta[1 + \exp(-\frac{1}{\beta}(x-\alpha)]^2}$$

for $-\infty < x < \infty$, $\beta > 0$, $-\infty < \alpha < \infty$ [Logistic density].

$$f_{47}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp[-(\ln x - \mu)^2/(2\sigma^2)]$$

for $x > 0$, $\sigma > 0$, $-\infty < \mu < \infty$ [Log normal density].

$$f_{48}(x) = \frac{4}{\sqrt{\pi}}\beta^{\frac{3}{2}}x^2\exp(-\beta x^2)$$

for $x \geq 0$, $\beta > 0$ [Maxwell–Boltzmann density].

$$f_{49}(x) = \frac{\alpha}{x_0}\left(\frac{x_0}{x}\right)^{\alpha+1}$$

for $x \geq x_0$, $\alpha > 0$ [Pareto density].

$$\frac{\mathrm{d}}{\mathrm{d}x}f_{50}(x) = \frac{(x-a)f_{50}(x)}{b_0 + b_1x + b_2x^2}$$

where $f_{50}(x)$ is the density function. The explicit solutions are classified into types I–XII according to the nature of roots of $b_0 + b_1x + b_2x^2 = 0$ [Pearson family].

$$f_{51}(x) = \frac{x}{\alpha^2}\exp\left[-\frac{1}{2}(x/\alpha)^2\right]$$

for $x \geq 0$, $\alpha > 0$ [Rayleigh density].

$$f_{52}(x) = \frac{\beta\exp[\alpha + \beta x]}{[1 + \exp(\alpha + \beta x)]^2}$$

for $-\infty < x < \infty$, $\beta > 0$, $-\infty < \alpha < \infty$ [Sech square density].

$$f_{53}(x) = \frac{\Gamma(\frac{1}{2}(\nu+1))}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$$

for $-\infty < x < \infty$, $\nu$ a positive integer [Student-t density].

$$f_{54}(x) = \frac{1}{a}\left(1 - \frac{|x|}{a}\right) \quad \text{for } |x| \leq a,\ a > 0$$

[Triangular density; there are several modifications of this density].

$$f_{55}(x) = [2\pi I_0(k)]^{-1}\mathrm{e}^{k\cos(x-\beta)} \quad \text{for } 0 < x \leq 2\pi,$$

$0 \leq \beta \leq 2\pi$, $0 \leq k < \infty$ where $I_o(k)$ is a Bessel function [Von Mises density].

$$f_{56}(x) = m(x-\alpha)^{m-1}\theta^{-1}\exp\left[-\frac{1}{\theta}(x-\alpha)^m\right]$$

for $x \geq \alpha$, $\theta > 0$, $m > 0$, $\alpha > 0$ [Weibull three parameter density; for $\alpha = 0$, we have the usual Weibull density, which is a special case of generalized gamma].

## Exercises 6

**6.1.** Evaluate $E(x)$ and $E(x^2)$ for the uniform density by differentiating the moment generating function in (6.2).

**6.2.** Obtain $E(x)$, $E(x^2)$, thereby the variance of the gamma random variable by using the moment generating function $M(t)$ in (6.14), (i) by expanding $M(t)$; (ii) by differentiating $M(t)$.

**6.3.** Expand the exponential part in the incomplete gamma integral $\gamma(\alpha; a)$, integrate term by term and obtain the series as a ${}_1F_1$ hypergeometric series.

**6.4.** Expand the factor $(1-x)^{\beta-1}$ in the incomplete beta integral in (6.27), integrate term by term and obtain a ${}_2F_1$ hypergeometric series.

**6.5.** Show that the functions $f_{11}(x)$ to $f_{55}(x)$ given in Section 6.11 are all densities, that is, show that the functions are non-negative and the total integral is 1 in each case.

**6.6.** For the functions in Exercise 6.5, compute (1) $E(x)$, (2) $\mathrm{Var}(x)$; (3) the moment generating function of $x$, whenever these exist.

**6.7.** For the functions in Exercise 6.5 compute (1) the Mellin transform; (2) the Laplace transform wherever they exist and wherever the variable is positive. Give the conditions of existence.

**6.8.** Let $f(x)$ be a real-valued density function of the real random variable $x$. Let $y$ be another real variable. Consider the functional equation

$$f(x)f(y) = f(\sqrt{x^2+y^2})$$

where $f$ is an arbitrary function. By solving this functional equation, show that $f(x)$ is a Gaussian density with $E(x) = \mu = 0$.

**6.9.** For the Exercise in 6.8, let $z$ be a real variable and let the functional equation be

$$f(x)f(y)f(z) = f(\sqrt{x^2+y^2+z^2}).$$

Show that the solution gives a Gaussian density with $E(x) = \mu = 0$.

**6.10.** Shannon's entropy, which is a measure of uncertainty in a distribution and which has wide range of applications in many areas, especially in physics, is given by

$$S = -c\int_{-\infty}^{\infty}[\ln f(x)]f(x)\mathrm{d}x$$

where $c$ is a constant and $f(x)$ is a non-negative integrable function. Show that if $S$ is maximized over all densities under the conditions (1) (i) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ then the resulting density is a uniform density; (2) Show that under the conditions (i) and (ii) $E(x) =$ is given or fixed over all functional $f$ then the resulting density is the exponential density; (3) Show that under the conditions (i), (ii) and (iii) $E(x^2) =$ a given quantity, then the resulting density is a Gaussian density. [Hint: Use calculus of variations].

**6.11.** The residual effect $x$ of two opposing forces behaves like small positive or negative residual effect having high probabilities and larger residual effect having smaller probabilities. A Laplace density is found to be a good model. For a Laplace density of the form $f(x) = c\mathrm{e}^{-\beta|x|}$, $-\infty < x < \infty$, where $c$ is the normalizing constant, compute (1) $\Pr\{-5 < x \le 3\}$; (2) $E(x)$; (3) moment generating function of $x$.

**6.12.** If $x \sim N(\mu = 2, \sigma^2 = 5)$ derive the densities of (1) $y_1 = 2x$; (2) $y_2 = 2x + 5$; (3) $y_3 = 3x^2$ when $\mu = 0$, $\sigma^2 = 1$; (4) $y_4 = 2 + 3x^2$ when $\mu = 0$, $\sigma^2 = 1$, by using transformation of variables.

**6.13.** If $x \sim \text{gamma}(\alpha = 3, \beta = 2)$ derive the density of (1) $y_1 = 4x^3$; (2) $y_2 = 4x^3 + 3$ by using transformation of variables.

**6.14.** Under a probability integral transformation, an exponential variable $x$ with expected value $E(x) = 3$ makes $y$ a uniformly distributed variable over $[0, 1]$. What is the $x$ value corresponding to (1) $y = 0.2$; (2) $y = 0.6$.

**6.15.** If $M_x(t)$ is the moment generating function of a random variable $x$, is $[M_x(t)]^{\frac{1}{n}}$, $n = 2, 3, \ldots$ a moment generating function for some random variable? List at least two random variables where this happens, from among the random variables discussed in this chapter.

**6.16.** Let $x$ be a type-1 beta random variable with parameters $(\alpha, \beta)$. Let (1) $y = \frac{x}{1-x}$; (2) $z = \frac{1-x}{x}$; (3) $u = 1 - x$. By using transformation of variables, show that $y$ and $z$ are type-2 beta distributed with parameters $(\alpha, \beta)$ and $(\beta, \alpha)$, respectively, and that $u$ is type-1 beta distributed with parameters $(\beta, \alpha)$.

**6.17.** Let $x$ be a type-2 beta random variable with parameters $(\alpha, \beta)$. Let (1) $y = \frac{x}{1+x}$; (2) $z = \frac{1}{1+x}$; (3) $u = \frac{1}{x}$. By using transformation of variables show that $y$ and $z$ are type-1 beta distributed with the parameters $(\alpha, \beta)$ and $(\beta, \alpha)$, respectively, and that $u$ is type-2 beta distributed with the parameters $(\beta, \alpha)$.

**6.18.** By using the moment generating function show that if $x$ is normally distributed, that is, $x \sim N(\mu, \sigma^2)$, then $y = ax + b$ is normally distributed, where $a$ and $b$ are constants. What is the distribution of $y$ when $a = 0$?

**6.19.** If $x$ is binomial with parameters $(n = 10, p = 0.6)$ and $y$ is Poisson with parameter $\lambda = 5$, evaluate the probability functions of $u = 2x + 3$ and $v = 3y - 5$.

**6.20.** Evaluate the probability generating function, $E(t^x)$, for (1) Poisson probability law, (2) geometric probability law; (3) negative binomial probability law.

**6.21.** Consider a Poisson arrival of points on a line or occurrence of an event over time according to a Poisson probability law as in (5.16) with the rate of arrival $\alpha$. Let $x$ denote the waiting time for the first occurrence. Then the probability $\Pr\{x > t\}$ = probability that the number of occurrence is zero $= \mathrm{e}^{-\alpha t}$, $t > 0$. Hence $\Pr\{x \le t\} = 1 - \mathrm{e}^{-\alpha t}$, $t > 0$, which means $x$ has an exponential distribution. If $y$ is the waiting time before the $r$-th occurrence of the event, then show that $y$ has a gamma density with parameter $(r, \frac{1}{\alpha})$.

**6.22.** Let a random variable $x$ have density function $f(x)$ and distribution function $F(x)$ then the hazard rate $\lambda(x)$ is defined as $\lambda(x) = \frac{f(x)}{1-F(x)}$. Compute the hazard rate when $x$ has (i) exponential density; (ii) Weibull density; (iii) logistic density.

# 7 Joint distributions

## 7.1 Introduction

There are many practical situations where we have pairs of random variables. Examples are many in nature. $(x,y)$ with $x$ = blood pressure of a patient before administering a drug, $y$ = blood pressure after administering that drug; $(x,y)$ with $x$ = height, $y$ = weight of an individual; $(x,y)$ with $x$ = waiting time for an interview, $y$ = the interview time; $(x,y)$ with $x$ = amount of fertilizer applied, $y$ = yield of tapioca, etc. In these examples, both variables $x$ and $y$ are continuous variables and they have some joint distributions. Let us consider the following cases: $(x,y)$ with $x$ = number of questions attempted in an examination, $y$ = number of correct answers; $(x,y)$ with $x$ = the number of local floods, $y$ = number of houses damaged; $(x,y)$ with $x$ = number of monthly traffic accidents on a particular road, $y$ = the number of associated injuries, etc. These are all pairs where both $x$ and $y$ are discrete. Now, consider the situations such as the following: $(x,y)$ with $x$ = number of monthly accidents, $y$ = the amount of compensation paid; $(x,y)$ with $x$ = number of computer breakdowns, $y$ = the duration of working hours lost, etc. These are pairs of variables where one of them is discrete and the other is continuous. We will consider joint distributions involving pairs of variables first and then we will extend the theory to joint distributions of many real scalar variables.

We will introduce the functions as mathematical quantities first and then we will look into experimental situations where such probability models will be appropriate.

**Definition 7.1** (Joint probability/density function). A function $f(x,y)$ is called a joint probability/density function of the random variables $x$ and $y$ if the following two conditions are satisfied:

(i) $f(x,y) \geq 0$ for all real values of $x$ and $y$;

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{d}x \wedge \mathrm{d}y = 1$ if $x$ and $y$ are continuous, where $\wedge$ product of differentials is explained in Module 4;

$\sum_{-\infty<x<\infty}\sum_{-\infty<y<\infty} f(x,y) = 1$ if $x$ and $y$ are discrete. (Sum up the discrete variable and integrate the continuous variable in the mixed case.)

**Note 7.1.** The wedge product or skew symmetric product of differentials is defined by the equation

$$\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$$

so that $\mathrm{d}x \wedge \mathrm{d}x = 0$. Applications in computing Jacobians of transformations and more details are given in Module 4.

$f(x_1,\dots,x_k)$ is the joint probability/density function of the random variables $x_1,\dots,x_k$ if the following conditions are satisfied:

https://doi.org/10.1515/9783110562545-007

(i) $f(x_1,\dots,x_k) \geq 0$ for all real values of $x_1,\dots,x_k$;
(ii) $\int_{x_1}\cdots\int_{x_k} f(x_1,\dots,x_k)\mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_k = 1$ if all variables are continuous;
$\sum_{x_1}\cdots\sum_{x_k} f(x_1,\dots,x_k) = 1$ if all variables are discrete. (Sum up the discrete variables and integrate the continuous variables in the mixed case.)

**Example 7.1.** Check whether the following are joint probability functions: $f(x,y)$ is given in the table for non-zero values and $f(x,y) = 0$ elsewhere.

|  | $x = 0$ | $x = 1$ | Sum |
|---|---|---|---|
| $y = 1$ | $\frac{1}{10}$ | $\frac{1}{10}$ | $\frac{2}{10}$ |
| $y = -1$ | $\frac{2}{10}$ | $\frac{1}{10}$ | $\frac{3}{10}$ |
| $y = 2$ | $\frac{3}{10}$ | $\frac{2}{10}$ | $\frac{5}{10}$ |
| Sum | $\frac{6}{10}$ | $\frac{4}{10}$ | 1 |

This can also be stated as follows:

$$f(x=0,y=1) = f(0,1) = \frac{1}{10},$$
$$f(0,-1) = \frac{2}{10}, \quad f(0,2) = \frac{3}{10}, \quad f(1,1) = \frac{1}{10},$$
$$f(1,-1) = \frac{1}{10}, \quad f(1,2) = \frac{2}{10} \quad \text{and} \quad f(x,y) \text{ is zero elsewhere.}$$

This can also be written as

$$\Pr\{x=0,y=1\} = \frac{1}{10}, \quad \Pr\{x=0,y=-1\} = \frac{2}{10},$$
$$\Pr\{x=0,y=2\} = \frac{3}{10}, \quad \Pr\{x=1,y=1\} = \frac{1}{10},$$
$$\Pr\{x=1,y=-1\} = \frac{1}{10}, \quad \Pr\{x=1,y=2\} = \frac{2}{10}$$

and $f(x,y) = 0$ for all other $x$ and $y$.

**Solution 7.1.** Since $f(x,y)$ here is non-negative for all $x$ and $y$ and since the total $\sum_x\sum_y f(x,y) = 1$, $f(x,y)$ is a joint probability function of two random variables $x$ and $y$.

**Example 7.2.** In Example 7.1, compute the following: (1) The probability function of $x$ alone, which is also called the *marginal probability function of* $x$; (2) the marginal probability function of $y$; (3) $\Pr\{x > 0, y > 1\}$; (4) $\Pr\{x > 0, y \leq -1\}$; (5) $\Pr\{x + y = 2\}$.

**Solution 7.2.** (1) $\Pr\{x = 0\}$ means all probabilities where this condition $x = 0$ is satisfied. This is available as the sum of the probabilities in the column corresponding to $x = 0$ or from the marginal sum, which is $\frac{6}{10}$. Similarly, $\Pr\{x = 1\} = \frac{4}{10}$. Thus the marginal probability function of $x$, namely $f_1(x)$, is available from the marginal sum as

$$f_1(x) = \begin{cases} 6/10, & x = 0 \\ 4/10, & x = 1 \\ 0, & \text{elsewhere.} \end{cases}$$

(2) Similarly, the probability function of $y$ is available from the marginal sum, given by

$$f_2(y) = \begin{cases} 2/10, & y = 1 \\ 3/10, & y = -1 \\ 5/10, & y = 2 \\ 0, & \text{elsewhere.} \end{cases}$$

(3)

$$\Pr\{x > 0, y > -1\} = \Pr\{x = 1, y = 1\} + \Pr\{x = 1, y = 2\} = \frac{1}{10} + \frac{2}{10} = \frac{3}{10}.$$

(4)

$$\Pr\{x > 0, y \le -1\} = \Pr\{x = 1, y = -1\} = \frac{1}{10}.$$

For computing the probability for $x + y$, first compute the possible values $x + y$ can take with non-zero probabilities. Possible values of $x + y$ are $1, -1, 2, 0, 3$.

$$\begin{aligned} \Pr\{x + y = 1\} &= \Pr\{x = 0, y = 1\} = \frac{1}{10} \\ \Pr\{x + y = -1\} &= \Pr\{x = 0, y = -1\} = \frac{2}{10} \\ \Pr\{x + y = 0\} &= \Pr\{x = 1, y = -1\} = \frac{1}{10} \\ \Pr\{x + y = 3\} &= \Pr\{x = 1, y = 2\} = \frac{2}{10} \end{aligned}$$

Similarly, for (5),

$$\Pr\{x + y = 2\} = \Pr\{x = 1, y = 1\} + \Pr\{x = 0, y = 2\} = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}.$$

**Example 7.3.** For the probability function in Example 7.1 compute the following: (1) Graph the function $f(x, y)$; (2) What is the probability function of $x$ given that $y = -1$?

**Solution 7.3.** (1) It is a 3-dimensional graph.

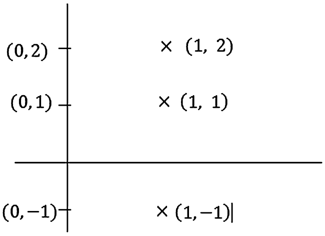


**Figure 7.1:** Joint probability function.

At the points marked in Figure 7.1, there are points up at heights equal to the probabilities or the $z$-coordinates are the probabilities.

(2) At $y = -1$, there are two points $(y = -1, x = 0)$ and $(y = -1, x = 1)$, with the respective probabilities $\frac{2}{10}$ at $x = 0$ and $\frac{1}{10}$ at $x = 1$, thus a total of $\frac{3}{10}$. Hence a probability function can be created by dividing by the total.

$$f(x \text{ given } y = -1) = \begin{cases} (\frac{2}{10})/(\frac{3}{10}), & x = 0 \\ (\frac{1}{10})/(\frac{3}{10}), & x = 1 \\ 0, & \text{elsewhere} \end{cases}$$

$$= \begin{cases} 2/3, & x = 0 \\ 1/3, & x = 1 \\ 0, & \text{elsewhere.} \end{cases}$$

**Example 7.4.** Evaluate $c$ so that $f(x,y)$, given below, is a density function and then compute the marginal density functions of $x$ and $y$.

$$f(x,y) = \begin{cases} c(x+y), & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 7.4.** Since $f(x,y)$ is defined on a continuum of points in the square $\{(x,y) \mid 0 \le x \le 1, 0 \le y \le 1\}$ it is a non-negative function if $c > 0$. The total integral is given by

$$\int_{x=0}^{1}\int_{y=0}^{1} c(x+y)\mathrm{d}x \wedge \mathrm{d}y = c\int_{x=0}^{1}\left[\int_{y=0}^{1}(x+y)\mathrm{d}y\right]\mathrm{d}x$$

$$= c\int_{x=0}^{1}\left[xy + \frac{y^2}{2}\right]_0^1 \mathrm{d}x = c\int_{x=0}^{1}\left(x + \frac{1}{2}\right)\mathrm{d}x$$

$$= c\left[\frac{x^2}{2} + \frac{x}{2}\right]_0^1 = 1 \quad \Rightarrow \quad c = 1.$$

(2) Density of $x$ alone is available by integrating out $y$ from the joint density, which is available from the above steps. That is, the marginal density of $x$ is given by

$$f_1(x) = \begin{cases} x + \frac{1}{2}, & 0 \le x \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

From symmetry, it follows that the marginal density of $y$ is given by

$$f_2(y) = \begin{cases} y + \frac{1}{2}, & 0 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

## Exercises 7.1

**7.1.1.** Check whether the following are probability functions:

$$\begin{aligned} &(1)\quad f(-1,1) = \frac{c}{20}, \quad f(-1,2) = \frac{1}{20}, \quad f(-1,3) = \frac{2}{20}, \\ &\qquad f(0,1) = \frac{2}{20}, \quad f(0,2) = \frac{3}{20}, \quad f(0,3) = \frac{8}{20} \quad \text{and} \\ &\qquad f(x,y) = 0, \quad \text{elsewhere.} \\ &(2)\quad f(-2,0) = 1 \quad \text{and} \\ &\qquad f(x,y) = 0, \quad \text{elsewhere.} \\ &(3)\quad f(3,1) = \frac{1}{5}, \quad f(3,2) = -\frac{2}{5}, \quad f(0,1) = \frac{3}{5}, \\ &\qquad f(0,2) = \frac{3}{5} \quad \text{and} \\ &\qquad f(x,y) = 0, \quad \text{elsewhere.} \end{aligned}$$

**7.1.2.** Check whether the following are density functions:

$$(1)\quad f(x,y) = \begin{cases} \frac{e^{-\frac{1}{2}(y-2x-3)^2}}{\sqrt{2\pi}}, & -\infty < y < \infty,\ 0 \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

$$(2)\quad f(x,y) = \begin{cases} \frac{c}{x^2} e^{-3(y-2x-3)^2}, & -\infty < y < \infty,\ 1 \le x < \infty,\ c = \text{constant} \\ 0, & \text{elsewhere.} \end{cases}$$

$$(3)\quad f(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-x-\frac{1}{2}(y-2x-3)^2}, & -\infty < y < \infty,\ 0 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

**7.1.3.** If $f(x,y) = c e^{-(\alpha x^2 + \beta xy + \gamma y^2)}$, $-\infty < x < \infty$, $-\infty < y < \infty$ is a density then find the conditions on $\alpha, \beta, \gamma$ and then evaluate $c$.

**7.1.4.** Can the following function be a joint density function (give reasons):

$$f(x,y) = \begin{cases} e^{x-y}, & 0 \le x < \infty,\ 0 \le y < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

**7.1.5.** Can the following be a density function (give reasons):

$$f(x,y) = c e^{-(2x^2 + 6xy + 3y^2)}, \quad -\infty < x < \infty,\ -\infty < y < \infty$$

where $c$ is a positive constant.

**7.1.6.** For Exercise 7.1.1 (1) compute the following: (1) $\Pr\{x=-1, y=1\}$; (2) $\Pr\{x<0, y>2\}$; (3) $\Pr\{x\le -2, y\ge 0\}$.

**7.1.7.** For Exercise 7.1.1 (2) compute the following probabilities: (1) $\Pr\{x=-2\}$; (2) $\Pr\{y=0\}$; (3) $\Pr\{x\le -2\}$; (4) $\Pr\{x=-2, y\ge 0\}$.

**7.1.8.** A balanced die is rolled 3 times. Let $x_1$ be the number of times 1 appears and $x_2$ be the number of times 2 appears. Work out the joint probability function of $x_1$ and $x_2$.

**7.1.9.** A box contains 10 red, 12 green and 14 white identical marbles. Marbles are picked at random, one by one, without replacement. 8 such marbles are picked. Let $x_1$ be the number of red marbles, $x_2$ be the number of green marbles obtained out of these 8 marbles. Construct the joint probability function of $x_1$ and $x_2$.

**7.1.10.** In Exercise 7.1.9, compute the probability $\Pr\{x_1\ge 8, x_2=5\}$.

## 7.2 Marginal and conditional probability/density functions

**Definition 7.2** (Marginal Probability/Density Functions). If we have $f(x_1,\dots,x_k)$ as a joint probability/density function of the random variables $x_1,\dots,x_k$, then the joint marginal probability/density function of any subset of $x_1,\dots,x_k$, for example, $x_1,\dots,x_r$, $r<k$, is available by summing up/integrating out the other variables. The marginal probability/density of $x_1,\dots,x_k$, denoted by $f_{1,\dots,r}(x_1,\dots,x_r)$, is given by

$$f_{1,\dots,r}(x_1,\dots,x_r)=\sum_{x_{r+1}}\cdots\sum_{x_k}f(x_1,\dots,x_k)$$

when $x_{r+1},\dots,x_k$ are discrete, and

$$=\int_{x_{r+1}}\cdots\int_{x_k}f(x_1,\dots,x_k)\mathrm{d}x_{r+1}\wedge\cdots\wedge\mathrm{d}x_k$$

when $x_{r+1},\dots,x_k$ are continuous. In the mixed cases, sum over the discrete variables and integrate over the continuous variables.

**Notation 7.1.** $x|y$ or $x|(y=b)\Rightarrow x$ given $y$ or $y$ is fixed at $y=b$. The notation is a vertical bar and not $x/y$ or $\frac{x}{y}$.

$$g_1(x|y)=\text{density of } x \text{ given } y;$$
$$g_1(x|y=b)=\text{density of } x \text{ given } y=b.$$

**Definition 7.3.** The conditional probability/density function of $x$, given $y=b$, is defined by the following:

$$g_1(x|y=b) = \frac{f(x,y)}{f_2(y)}\bigg|_{y=b}$$
$$= \frac{\text{joint density}}{\text{marginal density of } y}\bigg|_{y=b} \tag{7.1}$$

provided $f_2(b) \neq 0$. Similarly, the conditional density of $y$ given $x$ is given by

$$g_2(y|x=a) = \frac{f(x,y)}{f_1(x)}\bigg|_{x=a} \tag{7.2}$$

provided $f_1(a) \neq 0$.

**Example 7.5.** Verify that the following $f(x,y)$ is a density. Then evaluate (1) the marginal densities; (2) the conditional density of $x$ given $y = 0.8$; (3) the conditional density of $y$, given $x = 0.3$, where

$$f(x,y) = \begin{cases} 2, & 0 \le x \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 7.5.** The density is a flat plane over the triangular region as given in Figure 7.2.

The triangular region can be represented as one of the following:

$$\{(x,y) \mid 0 \le x \le y \le 1\} = \{(x,y) \mid 0 \le x \le y \;\&\; 0 \le y \le 1\}$$
$$= \{(x,y) \mid x \le y \le 1 \;\&\; 0 \le x \le 1\}.$$

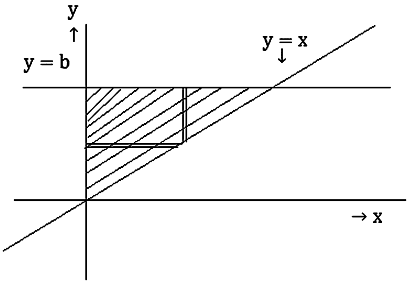


**Figure 7.2:** Triangular region for a joint density.

The elementary strips of integration are shown in Figure 7.2. The marginal density of $f(x,y)$ is given by

$$f_1(x) = \int_y f(x,y)\mathrm{d}y = \int_{y=x}^{1} 2\mathrm{d}y$$
$$= \begin{cases} 2(1-x), & 0 \le x \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

The marginal density of $y$ is given by

$$f_2(y) = \int_x f(x,y)\mathrm{d}x = \int_{x=0}^{y} 2\mathrm{d}x$$
$$= \begin{cases} 2y, & 0 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Note that

$$\int_0^1 f_1(x) = \int_0^1 2(1-x)\mathrm{d}x = 2\Big[x - \frac{x^2}{2}\Big]_0^1 = 1.$$

Hence

$$\int_x \int_y f(x,y)\mathrm{d}x \wedge \mathrm{d}y = 1.$$

Therefore, $f(x,y)$ is a joint density function. Here, the ordinate $f(x,y) = 2$, which is bigger than 1. But the probabilities are the volumes under the surface $z = f(x,y)$ (it is a 3-dimensional surface) and the ordinate does not matter. The ordinate can be bigger or less than 1 but the volumes have to be less than or equal to 1. The total volume here, which is the total integral, is 1, and hence it is a density function.

(2) The conditional density of $x$ given $y$, $g_1(x|y)$, is then

$$g_1(x|y) = \frac{f(x,y)}{f_2(y)}$$

and

$$g_1(x|y = 0.8) = \frac{f(x,y)}{f_2(y)}\Big|_{(y=0.8)} = \frac{2}{2y}\Big|_{(y=0.8)}$$
$$= \begin{cases} \frac{1}{0.8}, & 0 \le x \le 0.8 \\ 0, & \text{elsewhere.} \end{cases} \tag{1}$$

The conditional density of $y$ given $x$, is

$$g_2(y|x = 0.3) = \frac{f(x,y)}{f_1(x)}\Big|_{x=0.3}$$
$$= \frac{2}{2(1-x)}\Big|_{x=0.3} = \frac{1}{1-0.3}$$
$$\begin{cases} \frac{1}{0.7}, & 0.3 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Observe that in the conditional space $x$ varies from 0 to 0.8, whereas in the marginal space $x$ varies from 0 to 1. Similarly, in the conditional space $y$ varies from 0.3 to 1, whereas in the marginal space $y$ varies from 0 to 1.

### 7.2.1 Geometrical interpretations of marginal and conditional distributions

What are the geometrical interpretations of joint density, conditional densities, marginal densities, etc.?

When $f(x,y) \geq 0$ for all $x$ and $y$ and continuous, it means a surface sitting over the $(x,y)$-plane, something like a hill sitting on a plain ground. If $f(x,y)$ could be negative also, dipping below the $(x,y)$-plane then it will be like a ship in the ocean, taking the ocean surface as the $(x,y)$-plane. Then the portion of the ship's hull dipping below the water level represents $f(x,y) < 0$. Our densities are non-negative for all $(x,y)$, and hence when there are only two variables $x$ and $y$ then it is like a hill $z = f(x,y)$ sitting on the plain ground.

What is the geometry of the conditional density of $y$ given $x = a$?

Note that $x = a$ is a point in 1-space (line), it is a line in 2-space (plane) but a plane in 3-space, parallel to the $(y,z)$-plane. When this plane cuts the hill the whole hill tops will be traced on this plane. If the total area under this curve is made unity, then this is the conditional density of $y$ given $x = a$ or $g_2(y|x = a)$. In Figure 7.3, we have $z = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}$, $-\infty < x,y < \infty$; $z = (1+x)e^{-(1+x)y}$, $0 \leq x \leq 1$, $0 \leq y < \infty$.

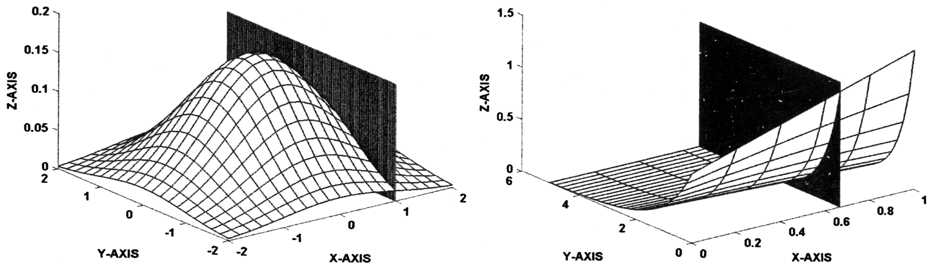


**Figure 7.3:** Surface cut by a plane.

What is the geometry of marginal densities?

Suppose that you use a giant bulldozer and push all the earth and stones of the hill from both sides of the $(x,z)$-plane and then pile up on the $(x,z)$-plane like a "pappadam". If we assume the total area under this pile-up as one unit, then we have the marginal density of $x$. Similar interpretation for the marginal density of $y$, pileup on the $(y,z)$-plane. In the discrete case, such pile-ups are already available from the marginal totals, as seen earlier.

In higher dimensions, we cannot see the geometry or visualize but algebraic evaluations are possible. In fact, we can only see 3-dimensional objects. We cannot see zero-dimensional (point), one-dimensional (line), 2-dimensional (plane), 4 and higher dimensional objects. But 0, 1, 2 dimensional objects can be visualized. You see a point, line or the surface of the blackboard only because of the thickness or as 3-dimensional objects.

**Example 7.6.** For the following function $f(x_1,x_2)$, (1) evaluate the normalizing constant; (2) the conditional density of $x_1$ given $x_2 = 0.7$; (3) the conditional probability for $x_1 > 0.2$ given that $x_2 = 0.7$; (4) the conditional probability for $x_1 > 0.2$ given that $x_2 < 0.7$, where

$$f(x_1,x_2) = \begin{cases} c(x_1+x_2), & 0 \le x_1 \le 1,\ 0 \le x_2 \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

**Solution 7.6.** (1) For evaluating $c$, we should compute the total probability, which is the total integral, and equate to 1.

$$\begin{aligned} 1 &= c\int_{x_1=0}^{1}\int_{x_2=0}^{1}(x_1+x_2)\mathrm{d}x_1 \wedge \mathrm{d}x_2 \\ &= c\int_{x_1=0}^{1}\left[\int_{x_2=0}^{1}(x_1+x_2)\mathrm{d}x_2\right]\mathrm{d}x_1 \\ &= c\int_0^1\left[x_1x_2+\frac{x_2^2}{2}\right]_0^1\mathrm{d}x_1 = c\int_0^1\left[x_1+\frac{1}{2}\right]\mathrm{d}x_1 \\ &= c\left[\frac{x_1^2}{2}+\frac{x_1}{2}\right]_0^1 = c \quad \Rightarrow \quad c = 1. \end{aligned}$$

Hence $f(x_1,x_2)$ is a density since it is already a non-negative function.

(2) The marginal density of $x_2$ is available by integrating out $x_1$ from the joint density. That is,

$$f_2(x_2) = \int_{x_1=0}^{1}(x_1+x_2)\mathrm{d}x_1 = \begin{cases} x_2+\frac{1}{2}, & 0 \le x_2 \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Hence the conditional density of $x_1$ given $x_2$, is given by

$$g_1(x_1|x_2) = \frac{f(x_1,x_2)}{f_2(x_2)} = \frac{(x_1+x_2)}{x_2+\frac{1}{2}}$$

for $0 \le x_1 \le 1$ and for all given $x_2$. Therefore,

$$g_1(x_1|x_2 = 0.7) = \left.\frac{x_1+x_2}{x_2+\frac{1}{2}}\right|_{x_2=0.7} = \begin{cases} \frac{x_1+0.7}{1.2}, & 0 \le x_1 \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Note that the ranges of $x_1$ and $x_2$ do not depend on each other, not like in Example 7.5, and hence in the conditional space also $x_2$ ranges over $[0,1]$. [The student may verify that $g_1(x_1|x_2 = 0.7)$, as given above, is a density function by evaluating the total probability and verifying it to be 1.]

(3) $\Pr\{x_1 > 0.2|x_2 = 0.7\}$ is the probability for $x_1 > 0.2$ in the conditional density. That is,

$$\begin{aligned}\Pr\{x_1 > 0.2|x_2 = 0.7\} &= \int_{0.2}^{1} g_1(x_1|x_2 = 0.7)\mathrm{d}x_1 = \int_{0.2}^{1} \frac{x_1 + 0.7}{1.2}\mathrm{d}x_1 \\ &= \frac{1}{1.2}\left[\frac{x_1^2}{2} + 0.7x_1\right]_{0.2}^{1} \approx 0.87.\end{aligned}$$

(4) This does not come from the conditional density $g_1(x_1|x_2)$. Let $A$ be the event that $x_1 > 0.2$ and $B$ be the event that $x_2 < 0.7$. Then

$$A \cap B = \{(x_1, x_2) \mid 0.2 < x_1 < 1, 0 \le x_2 < 0.7\}.$$

These events are marked in Figure 7.4.

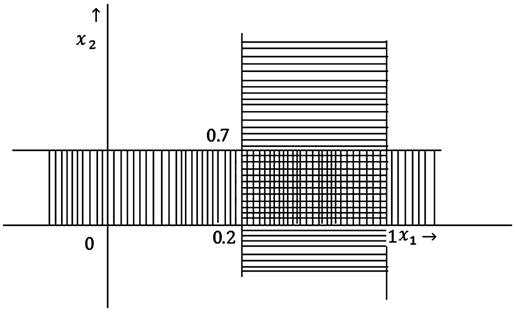


**Figure 7.4:** Events $A, B, A \cap B$.

Probabilities of $A$ and $B$ can be computed either from the marginal densities or from the joint density.

$$\begin{aligned}P(A \cap B) &= \int_{x_1=0.2}^{1} \int_{x_2=0}^{0.7} (x_1 + x_2)\mathrm{d}x_1 \wedge \mathrm{d}x_2 \\ &= \int_{x_1=0.2}^{1} \left[\int_{x_2=0}^{0.7} (x_1 + x_2)\mathrm{d}x_2\right]\mathrm{d}x_1 \\ &= \int_{x_1=0.2}^{1} \left[x_1x_2 + \frac{x_2^2}{2}\right]_0^{0.7} \mathrm{d}x_1 = \int_{x_1=0.2}^{1} \left[0.7x_1 + \frac{0.49}{2}\right]\mathrm{d}x_1 \\ &= \left[0.7\frac{x_1^2}{2} + \frac{0.49}{2}x_1\right]_{0.2}^{1} = 0.532.\end{aligned}$$

$$\begin{aligned}P(B) = \Pr\{0 \le x_2 < 0.7\} &= \int_{x_1=0}^{1} \int_{x_2=0}^{0.7} (x_1 + x_2)\mathrm{d}x_1 \wedge \mathrm{d}x_2 \\ &= \int_{x_2=0}^{0.7} \left(x_2 + \frac{1}{2}\right)\mathrm{d}x_2 = \left[\frac{x_2^2}{2} + \frac{x_2}{2}\right]_0^{0.7} \\ &= 0.595.\end{aligned}$$

Therefore,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.532}{0.595} \approx 0.89.$$

From the definition of conditional probability/density function, we may note one interesting property:

$$g_1(x|y) = \frac{f(x,y)}{f_2(y)}, \quad f_2(y) \neq 0.$$

This shows that we can always have a decomposition of the joint density as product of conditional and marginal densities.

$$\begin{aligned} f(x,y) &= g_1(x|y)f_2(y), \quad f_2(y) \neq 0 \\ &= g_2(y|x)f_1(x), \quad f_1(x) \neq 0 \end{aligned} \tag{7.3}$$

where $f_1(x)$ and $f_2(y)$ are the marginal densities and $g_1(x|y)$ and $g_2(y|x)$ are the conditional densities. If there are $k$ variables $x_1, \dots, x_k$ and if we look at the conditional joint density of $x_1, \dots, x_r$ given $x_{r+1}, \dots, x_k$, then also we have the decomposition

$$f(x_1, \dots, x_k) = g(x_1, \dots, x_r | x_{r+1}, \dots, x_k) f_2(x_{r+1}, \dots, x_k) \tag{7.4}$$

where $f_2(x_{r+1}, \dots, x_k) \neq 0$. Another observation one can make from (7.3) and (7.4) is that if the conditional probability/density functions do not depend on the conditions or free of $y$ and $x$, that is, if $g_1(x|y) = f_1(x) =$ the marginal probability/density of $x$ itself and $g_2(y|x) = f_2(y) =$ the marginal probability/density of $y$ itself, then we have the product probability property

$$f(x,y) = f_1(x)f_2(y)$$

and this is also called statistical independence of the random variables $x$ and $y$.

## Exercises 7.2

**7.2.1.** Compute (1) marginal probability functions; conditional probability functions of (2) $x$ given $y = 1$; (3) $y$ given $x = -1$, in Exercise 7.1.1 (1).

**7.2.2.** Compute the marginal and conditional probability functions, for all values of the conditioned variables, in Exercise 7.1.1 (2).

**7.2.3.** Construct the conditional density of $y$ given $x$ and the marginal density of $x$ in Exercise 7.1.2 (1). What is the marginal density of $y$ here?

**7.2.4.** Repeat Exercise 7.2.3 for the function in Exercise 7.1.2 (2), if possible.

**7.2.5.** Repeat Exercise 7.2.3 for the function in Exercise 7.1.2 (3), if possible.

## 7.3 Statistical independence of random variables

**Definition 7.4** (Statistical independence). Let $f(x_1, \dots, x_k)$ be the joint probability/density function of the real random variables $x_1, \dots, x_k$ and let $f_1(x_1), f_2(x_2), \dots, f_k(x_k)$ be the marginal probability/density functions of the individual variables. Let them satisfy the condition

$$f(x_1,\dots,x_k)=f_1(x_1)\cdots f_k(x_k) \tag{7.5}$$

for all $x_1,\dots,x_k$ for which $f_i(x_i)\neq 0$, $i=1,\dots,k$ then $x_1,\dots,x_k$ are said to be *statistically independently distributed* or $x_1,\dots,x_k$ are said to satisfy the *product probability property*.

Because of the term "independent" this concept is often misused in applications. The variables depend on each other heavily through the product probability property (7.5). The phrase "independent" is used in the sense that if we look at the conditional probability/density function then the function does not depend on the conditions imposed. In this sense "independent of the conditions" or does not depend on the observation made. Variables may not satisfy the property in (7.5) individually but in two sets or groups the variables may have the property in (7.5). That is, suppose that

$$f(x_1,\dots,x_k)=f_1(x_1,\dots,x_r)f_2(x_{r+1},\dots,x_k) \tag{7.6}$$

then we say that the two sets $\{x_1,\dots,x_r\}$ and $\{x_{r+1},\dots,x_k\}$ of variables are statistically independently distributed. If the property (7.6) holds in two such sets, that does not imply that the property (7.5) holds on individual variables. But if (7.5) holds then, of course, (7.6) will hold, not vice versa.

**Example 7.7.** If the following function

$$f(x_1,x_2,x_3)=\begin{cases} ce^{-2x_1-x_2-4x_3}, & 0\le x_j<\infty,\ j=1,2,3\\ 0, & \text{elsewhere}\end{cases}$$

is a density function then compute (1) $c$; (2) density of $x_1$ given $x_2=5$, $x_3=2$; (3) probability that $x_1\le 10$ given that $x_2=5$, $x_3=2$.

**Solution 7.7.**

$$\begin{aligned}1&=c\int_{x_1=0}^{\infty}\int_{x_2=0}^{\infty}\int_{x_3=0}^{\infty}\mathrm{e}^{-2x_1-x_2-4x_3}\mathrm{d}x_1\wedge\mathrm{d}x_2\wedge\mathrm{d}x_3\\ &=c\int_{x_1=0}^{\infty}\int_{x_2=0}^{\infty}\mathrm{e}^{-2x_1-x_2}\left[\int_{x_3=0}^{\infty}\mathrm{e}^{-4x_3}\mathrm{d}x_3\right]\mathrm{d}x_1\wedge\mathrm{d}x_2.\end{aligned}$$

But

$$\int_0^{\infty}\mathrm{e}^{-4x_3}\mathrm{d}x_3=\left[-\frac{1}{4}\mathrm{e}^{-4x_3}\right]_0^{\infty}=\frac{1}{4}.$$

Similarly, the integral over $x_1$ gives $\frac{1}{2}$ and the integral over $x_2$ gives 1. Hence

$$1=\frac{c}{4\times 2}\quad\Rightarrow\quad c=8.$$

(2) Here, we want the conditional density of $x_1$ given $x_2$ and $x_3$. The joint marginal density of $x_2$ and $x_3$ is available by integrating out $x_1$ from $f(x_1,x_2,x_3)$. Denoting by

$$f_{23}(x_2,x_3) = \int_{x_1=0}^{\infty} 8e^{-2x_1-x_2-4x_3}dx_1 = 8e^{-x_2-4x_3}\int_0^{\infty} e^{-2x_1}dx_1$$

$$= \begin{cases} 4e^{-x_2-4x_3}, & 0 \le x_j < \infty,\ j=2,3 \\ 0, & \text{elsewhere.} \end{cases}$$

Hence the conditional density of $x_1$, given $x_2, x_3$, denoted by $g_1(x_1|x_2,x_3)$, is available as

$$g_1(x_1|x_2,x_3) = \frac{f(x_1,x_2,x_3)}{f_{23}(x_2,x_3)}\Big|_{x_2=5,x_3=2}$$

$$= \frac{8\exp[-2x_1-x_2-4x_3]}{4\exp[-x_2-4x_3]}\Big|_{x_2=5,x_3=2}$$

$$= 2e^{-2x_1}, \quad 0 \le x_1 < \infty.$$

In other words, this function is free of the conditions $x_2 = 5$, $x_3 = 2$. In fact, this conditional density of $x_1$ is the marginal density of $x_1$ itself. It is easily seen here that

$$f(x_1,x_2,x_3) = f_1(x_1)f_2(x_2)f_3(x_3)$$

the product of the marginal densities or the variables $x_1, x_2, x_3$ are statistically independently distributed.

Statistical independence can also be defined in terms of joint distribution function, joint moment generating function, etc. We have defined the concept assuming the existence of the probability/density functions. In some cases, the moment generating function may be defined but the density may not exist. Lévy distribution and singular normal are examples.

**Definition 7.5** (Joint distribution function). The cumulative probability/density function in $x_1,\dots,x_k$ is denoted by $F_{x_1,\dots,x_k}(a_1,\dots,a_k)$ and it is the following:

$$F_{x_1,\dots,x_k}(a_1,\dots,a_k) = \Pr\{x_1 \le a_1,\dots,x_k \le a_k\} \tag{7.7}$$

for all real values of $a_1,\dots,a_k$.

If the right side of (7.7) can be evaluated, then we say that we have the joint *distribution function* for the random variables $x_1,\dots,x_k$. In terms of probability/density functions

$$F_{x_1,\dots,x_k}(a_1,\dots,a_k) = \sum_{-\infty<x_1\le a_1}\cdots\sum_{-\infty<x_k\le a_k} f(x_1,\dots,x_k)$$

when $x_1,\dots,x_k$ are discrete, and it is

$$= \int_{-\infty}^{a_1}\cdots\int_{-\infty}^{a_k} f(x_1,\dots,x_k)dx_1\wedge\cdots\wedge dx_k \tag{7.8}$$

when $x_1,\dots,x_k$ are continuous. For the mixed cases, sum up over discrete variables and integrate over the continuous variables.

Note that when the random variables $x_1, \dots, x_k$ are independently distributed, that is, when the joint density is the product of the marginal densities then the joint distribution function factorizes into product of individual distribution functions. That is,

$$F_{x_1,\dots,x_k}(a_1,\dots,a_k) = F_{x_1}(a_1)\cdots F_{x_k}(a_k) \tag{7.9}$$

where $F_{x_i}(a_i)$ is the distribution function or cumulative probability/density function of $x_i$, evaluated at $x_i = a_i$, $i = 1,\dots,k$. One can use (7.9) as the definition of statistical independence then the joint distribution function becomes the product of the individual distribution functions. Then from (7.9), (7.5) will follow. Example 7.7 is a case where the variables are independently distributed.

## Exercises 7.3

**7.3.1.** If $f(x_1,x_2,x_3) = c(x_1 + x_2 + x_3)$, $0 \le x_i \le 1$, $i = 1,2,3$ and $f(x_1,x_2,x_3) = 0$ elsewhere is a density function then (1): evaluate $c$; (2): show that the variables here are not independently distributed.

**7.3.2.** If the following is the non-zero part of a joint density then evaluate (1) the normalizing constant $c$ and list the conditions needed on the parameters; (2) the marginal densities of $x_1$ and $x_2$; (3) show that $x_1$ and $x_2$ are not independently distributed.

$$f(x_1,x_2) = cx_1^{\alpha-1}x_2^{\alpha_2-1}(1 - x_1 - x_2)^{\alpha_3-1}$$

for $0 \le x_1 \le 1$, $0 \le x_2 \le 1$, $x_1 + x_2 \le 1$.

**7.3.3.** Do the same Exercise 7.3.2 if the function is the following:

$$f(x_1,x_2) = cx_1^{\alpha_1-1}x_2^{\alpha_2-1}(1 + x_1 + x_2)^{-(\alpha_1+\alpha_2+\alpha_3)}$$

for $0 \le x_1 < \infty$, $0 \le x_2 < \infty$.

**7.3.4.** Do the same Exercise 7.3.2 if the function is the following:

$$f(x_1,x_2) = cx_1^{\alpha-1}(x_1 + x_2)^{\beta_1}(1 - x_1 - x_2)^{\alpha_3-1}$$

for $0 \le x_1 \le 1$, $0 \le x_2 \le 1$, $x_1 + x_2 \le 1$.

**7.3.5.** Do the same Exercise 7.3.2 if the function is the following:

$$f(x_1,x_2) = cx_1^{\alpha_1-1}(x_1 + x_2)^{\beta_1}(1 + x_1 + x_2)^{-\gamma}$$

for $0 \le x_1 < \infty$, $0 \le x_2 < \infty$.

## 7.4 Expected value

The definition of expected values in the joint distribution is also parallel to that in the one variable case. Let $f(x_1,\dots,x_k)$ be the joint probability/density function of real random variables $x_1,\dots,x_k$. Let $\psi(x_1,\dots,x_k)$ be a function of $x_1,\dots,x_k$.

**Notation 7.2.** $E[\psi(x_1,\dots,x_k)]$: expected value of $\psi(x_1,\dots,x_k)$.

**Definition 7.6** (Expected value).

$$E[\psi(x_1,\dots,x_k)] = \sum_{-\infty<x_1<\infty} \cdots \sum_{-\infty<x_k<\infty} \psi(x_1,\dots,x_k)f(x_1,\dots,x_k)$$

if $x_1,\dots,x_k$ are discrete;

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(x_1,\dots,x_k)f(x_1,\dots,x_k)\mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_k$$

if $x_1,\dots,x_k$ are continuous. For the mixed cases, integrate over the continuous variables and sum up over the discrete variables.

**Example 7.8.** Compute (1) $E(xy)$; (2) $E\{[x-E(x)][y-E(y)]\}$; (3) $E[x]$ for the following probability function:

$$f(0,-1) = \frac{1}{5}; \quad f(0,1) = \frac{2}{5};$$
$$f(1,-1) = \frac{1}{5}; \quad f(1,1) = \frac{1}{5};$$

and $f(x,y) = 0$ elsewhere.

**Solution 7.8.** Since both the variables $x$ and $y$ here are discrete we will sum up.

(1) Expected value of $xy$ means, take the values that $x$ can take with non-zero probabilities and the corresponding $y$ values, multiply together and then multiply by the corresponding probabilities and add up all such sums. For example, when $x$ takes the value 0 and $y$ takes the value $-1$ the corresponding probability is $\frac{1}{5}$. Hence this term is $(0)(-1)(\frac{1}{5}) = 0$. That is, $E(xy) = (0)(-1)(\frac{1}{5}) + (0)(1)(\frac{2}{5}) + (1)(-1)(\frac{1}{5}) + (1)(1)(\frac{1}{5}) + 0 = 0$. Thus $E[xy]$ in this example is zero.

(2) For computing the second expected value, we need $E[x]$ and $E[y]$. We can compute these from either the marginal probability functions or from the joint probability function. The marginal probability function of $x$ is given by

$$f_1(x) = \sum_y f(x,y) = \begin{cases} 3/5, & x = 0 \\ 2/5, & x = 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Therefore,

$$E[x] = (0)\left(\frac{3}{5}\right) + (1)\left(\frac{2}{5}\right) = \frac{2}{5}.$$

If this is to be computed by using the joint probability function, then take one $x$ value and then add up all the corresponding probabilities, multiply and add up. That is,

$$
\begin{aligned}
E[x] &= \sum_x \sum_y x f(x,y) = \sum_x \Big[\sum_y x f(x,y)\Big] = \sum_x x f_1(x) \\
&= (0)\Big(\frac{1}{5}+\frac{2}{5}\Big) + (1)\Big(\frac{1}{5}+\frac{1}{5}\Big) = \frac{2}{5}.
\end{aligned}
$$

Similarly, $E[y] = (-1)(\frac{2}{5}) + (1)(\frac{3}{5}) = \frac{1}{5}$. Now, we can compute

$$
E\{[x-E(x)][y-E(y)]\} = E\Big\{\Big[x-\frac{2}{5}\Big]\Big[y-\frac{1}{5}\Big]\Big\}.
$$

Now put a value $x$ takes, put the corresponding $y$ value and multiply by the corresponding probability and add up all such quantities.

$$
\begin{aligned}
E\Big\{\Big[x-\frac{2}{5}\Big]\Big[x-\frac{1}{5}\Big]\Big\} &= \sum_x \sum_y \Big[x-\frac{2}{5}\Big]\Big[y-\frac{1}{5}\Big] f(x,y) \\
&= \Big[0-\frac{2}{5}\Big]\Big[-1-\frac{1}{5}\Big]\Big(\frac{1}{5}\Big) + \Big[0-\frac{2}{5}\Big]\Big[1-\frac{1}{5}\Big]\Big(\frac{2}{5}\Big) \\
&\quad + \Big[1-\frac{2}{5}\Big]\Big[-1-\frac{1}{5}\Big]\Big(\frac{1}{5}\Big) + \Big[1-\frac{2}{5}\Big]\Big[1-\frac{1}{5}\Big]\Big(\frac{1}{5}\Big) \\
&= \frac{12}{5^3} - \frac{16}{5^3} - \frac{18}{5^3} + \frac{12}{5^3} = -\frac{2}{25},
\end{aligned}
$$

which is also equal to

$$
E[xy] - E(x)E(y) = 0 - \Big(\frac{2}{5}\Big)\Big(\frac{1}{5}\Big) = -\frac{2}{25}
$$

in this case. In fact, this is a general result.

**Notation 7.3.** Cov$(x,y)$: covariance between $x$ and $y$.

**Definition 7.7** (Covariance)**.** The covariance between two real scalar random variables $x$ and $y$ is defined as

$$
\operatorname{Cov}(x,y) = E\{[x-E(x)][y-E(y)]\} \equiv E[xy] - E(x)E(y). \tag{7.10}
$$

The equivalence can be seen by observing the following: Once the expected value is taken it is a constant (does not depend on the variables any more); expected value of a constant times a function is constant times the expected value; expected value of a sum is the sum of the expected values as long as the expected values exist. These properties, analogous to the corresponding properties in the one variable case, will follow from the definition itself. Opening up

$$
[x-\mu_1][y-\mu_2] = xy - \mu_1 y - \mu_2 x + \mu_1\mu_2
$$

where $\mu_1 = E(x)$ and $\mu_2 = E(y)$ are constants. Now taking the expected values we have

$$\begin{aligned}\mathrm{Cov}(x,y) &= E\{[x-E(x)][y-E(y)]\}\\ &= E\{[x-\mu_1][y-\mu_2]\} = E\{xy-\mu_1 y-\mu_2 x+\mu_1\mu_2\}\\ &= E(xy)-\mu_1 E(y)-\mu_2 E(x)+\mu_1\mu_2\\ &= E(xy)-\mu_1\mu_2-\mu_1\mu_2+\mu_1\mu_2\\ &= E(xy)-\mu_1\mu_2 = E(xy)-E(x)E(y).\end{aligned}$$

This result holds when both $x$ and $y$ are continuous, both are discrete, one is continuous and the other is discrete, whenever the expected values exist.

What is this measure of covariance? We have seen that in the one variable case, standard deviation can be interpreted as a "measure of scatter" or "dispersion" in the single variable $x$, and the square of the standard deviation is the variance. Analogous to variance, **covariance measures the scatter or joint dispersion or angular dispersion or joint scatter in the point** $(x,y)$ **or the joint variation of the coordinates** $x$ **and** $y$**, so that when** $x=y$ **then the covariance becomes the variance in** $x$. For example, if $(x,y)=(x_1,y_1),\dots,(x_p,y_p)$ with probabilities $\frac{1}{n}$ at each point then $E(x)=\bar{x}$ and $E(y)=\bar{y}$, $\mathrm{Cov}(x,y)=\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y})/n$ but if the angle between the two vectors is $\theta$ then

$$\cos\theta = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)}\sqrt{\mathrm{Var}(y)}}. \tag{7.11}$$

Thus, covariance measures the angular dispersion between the random variables $x$ and $y$

**Example 7.9.** For the following continuous case evaluate (1) $\mathrm{Cov}(x,y)$; (2) $E(x^2+xy-y^2)$:

$$f(x,y)\begin{cases}2, & 0\le x\le y\le 1\\ 0, & \text{elsewhere.}\end{cases}$$

**Solution 7.9.** This density function was handled before and the marginal densities were already evaluated. For computing $E(x^2)$ and $E(y^2)$, we will use the marginal densities, which are

$$f_1(x)=2(1-x),\quad 0\le x\le 1$$

and zero elsewhere, and

$$f_2(y)=2y,\quad 0\le y\le 1$$

and zero elsewhere. Hence

$$\begin{aligned}E(x^2) &= \int_x x^2 f_1(x)\mathrm{d}x = \int_0^1 (x^2)2(1-x)\mathrm{d}x\\ &= 2\left[\frac{x^3}{3}-\frac{x^4}{4}\right]_0^1 = \frac{1}{6}\end{aligned}$$

and

$$E(y^2) = \int_y (y^2) f_2(y)\mathrm{d}y = \int_0^1 (y^2)2(y)\mathrm{d}y$$
$$= 2\left[\frac{y^4}{4}\right]_0^1 = \frac{1}{2}.$$

For computing $E(xy)$, we have to use double integrals. Either we can integrate out $y$ first from $x \le y \le 1$ and then $x$ from $0 \le x \le 1$ or integrate out $x$ first from $0 \le x \le y$ and then $y$ from $0 \le y \le 1$. Therefore,

$$E(xy) = \int_{y=0}^1 (y)\left[\int_{x=0}^y 2(x)\mathrm{d}x\right]\mathrm{d}y$$
$$= \int_0^1 y[y^2]\mathrm{d}y = \left[\frac{y^4}{4}\right]_0^1 = \frac{1}{4}.$$

Now we need $E(x)$ and $E(y)$.

$$E(x) = \int_x x f_1(x)\mathrm{d}x = \int_0^1 (x)2(1-x)\mathrm{d}x = \frac{1}{3}$$

and

$$E(y) = \int_y (y) f_2(y)\mathrm{d}y = \int_0^1 y(2y)\mathrm{d}y = \frac{2}{3}.$$

Therefore, (1)

$$\mathrm{Cov}(x,y) = E(xy) - E(x)E(y) = \frac{1}{4} - \left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{1}{36}.$$

and (2)

$$E[x^2 + xy - y^2] = E(x^2) + E(xy) - E(y^2) = \frac{1}{6} + \frac{1}{4} - \frac{1}{2} = -\frac{1}{12}.$$

### 7.4.1 Some properties of expected values

Some properties parallel to the ones in the one variable case are the following. These follow from the definition itself.

**Result 7.1.**

$$E(c) = c$$

*where c is a constant, with respect to the variables $x_1, \dots, x_k$ for which the joint distribution is used to compute the expected values.*

**Result 7.2.**

$$E[c\psi(x_1,\dots,x_k)] = cE[\psi(x_1,\dots,x_k)]$$

*whenever the expected value exists, where c is a constant.*

**Result 7.3.**

$$E[a\psi_1(x_1,\dots,x_k) + b\psi_2(x_1,\dots,x_k)] = aE[\psi_1(x_1,\dots,x_k)] + bE[\psi_2(x_1,\dots,x_k)]$$

*whenever the expected values exist, where a and b are constants, and $\psi_1$ and $\psi_2$ are two functions of $x_1,\dots,x_k$.*

[Here, we have a linear combination of two functions. The result holds for a finite number of such linear combinations but need not hold for an infinite sum.]

**Result 7.4.** *Let $\psi_1(x_1),\psi_2(x_2),\dots,\psi_k(x_k)$ be functions of $x_1,\dots x_k$ alone, then the expected value of a product of finite number of such factors is the product of the expected values, when the variables are independently distributed. That is,*

$$E[\psi_1(x_1)\psi_2(x_2)\cdots\psi_k(x_k)] = E[\psi_1(x_1)]E[\psi_2(x_2)]\cdots E[\psi_k(x_k)]$$

*whenever $x_1,\dots,x_k$ are independently distributed.*

The proof is trivial because when the variables are independently distributed the product probability property will hold and the joint probability/density function will factorize into product of individual probability/density functions then the sum or integral will apply to each factor.

One consequence of this result is that if the variables are independently distributed then the covariance between them is zero but the converse need not be true. Covariance being zero does not imply that the variables are independently distributed.

**Result 7.5.** *When x and y are independently distributed,* $\mathrm{Cov}(x,y) = 0$, *but the converse need not be true. That is,*

$$\mathrm{Cov}(x,y) = 0$$

*when x and y are independently distributed.*

The proof is trivial. When the variables are independently distributed the expected value of a product is the product of the expected values. Therefore,

$$E[(x - E(x))(y - E(y))] = E[x - E(x)]E[y - E(y)].$$

But for any random variable $E[x - E(x)] = E(x) - E(x) = 0$, and hence the result.

**Example 7.10.** For the following joint probability function, show that $\mathrm{Cov}(x,y)=0$ but the variables are not independently distributed.

$$f(0,-1)=\frac{3}{10},\quad f(0,1)=\frac{2}{10},\quad f(1,-1)=\frac{1}{10},\quad f(2,1)=0,$$
$$f(2,-1)=\frac{1}{10},\quad f(1,1)=\frac{3}{10}$$

and $f(x,y)=0$ elsewhere.

**Solution 7.10.** The marginal probability function of $y$ is given by the marginal sum, that is,

$$f_2(y)=\begin{cases}1/2, & y=-1\\ 1/2, & y=1\\ 0, & \text{elsewhere.}\end{cases}$$

Hence $E[y]=(-1)\frac{1}{2}+(1)\frac{1}{2}=0$. Also

$$E(xy)=(0)(-1)\Big(\frac{3}{10}\Big)+(0)(1)\Big(\frac{2}{10}\Big)+(1)(-1)\Big(\frac{1}{10}\Big)$$
$$+(2)(1)(0)+(2)(-1)\Big(\frac{1}{10}\Big)+(1)(1)\Big(\frac{3}{10}\Big)=0.$$

But, for example,

$$\Pr\{x=1,y=-1\}=\frac{1}{10};\quad \Pr\{x=1\}=\frac{4}{10};\quad \Pr\{y=-1\}=\frac{1}{2}.$$

Therefore, $\Pr\{x=1,y=-1)\neq\Pr\{x=1\}\Pr\{y=-1\}$, and hence the variables $x$ and $y$ are not independently distributed.

**Note 7.4.** If the product probability property does not hold even for one point, then the variables are not independently distributed. Hence to show that two variables are not independently distributed, one has to come up with at least one point where the product probability property does not hold.

### 7.4.2 Joint moment generating function

**Notation 7.4.** $M(T)=M(t_1,\dots,t_k)$: Joint moment generating function.

**Definition 7.8** (Joint moment generating function). The joint moment generating function is defined as the expected value of a linear function in the exponent. That is,

$$M(t_1,\dots,t_k)=E[\mathrm{e}^{t_1x_1+\cdots+t_kx_k}]$$

where $t_1,\dots,t_k$ are arbitrary parameters, provided the expected value exists.

From the joint moment generating function $M(t_1,\dots,t_k)$, the individual moment generating functions (mgf) are available by putting the other $t_j$'s zeros. For example, the joint mgf of the first $r$ variables, $r<k$, is available by putting $t_{r+1}=0=\cdots=t_k$, or it is $M(t_1,\dots,t_r,0,\dots,0)$. Also the joint integer moments, usually known as *product moments*, of the type $E[x_1^{r_1}\cdots x_k^{r_k}]$ where $r_j=0,1,2,\dots$, $j=1,\dots,k$ are available from $M(t_1,\dots,t_k)$ by expansion, when $M(t_1,\dots,t_k)$ admits a power series expansion or by differentiation when $M(t_1,\dots,t_k)$ is differentiable. That is,

$$\begin{aligned}E(x_1^{r_1}\cdots x_k^{r_k})&=\frac{\partial^{r_1+\cdots+r_k}}{\partial t_1^{r_1}\cdots\partial t_k^{r_k}}M(t_1,\dots,t_k)\Big|_{t_1=0,\dots,t_k=0}\\&=\text{coefficient of }\frac{t_1^{r_1}\cdots t_k^{r_k}}{r_1!\cdots r_k!}\end{aligned}$$

in the expansion of $M(t_1,\dots,t_k)$ around $(t_1=0,\dots,t_k=0)$, for $r_j=0,1,\dots$, $j=1,\dots,k$.

Sometimes it is convenient to use the vector, matrix notation. Let the prime denote a transpose and let $T'=(t_1,\dots,t_k)$ then we may also represent the joint moment generating function as either $M(T)$ or $M(T')$ according to convenience. One immediate consequence of statistical independence is that the joint moment generating function will factorize into individual moment generating functions. This will be stated as a result but it follows from the fact that when the variables are independently distributed the joint probability/density or joint distribution function factorizes into individual probability/density or distribution functions.

**Result 7.6.** *When the real random variables $x_1,\dots,x_k$ are independently distributed, then the joint moment generating function, when it exists, factorizes into the product of individual moment generating functions. That is,*

$$M(t_1,\dots,t_k)=M_{x_1}(t_1)\cdots M_{x_k}(t_k) \tag{7.12}$$

*where $M_{x_j}(t_j), j=1,\dots,k$ are the individual moment generating functions.*

One may use (7.12) as the definition of statistical independence. Then the product property of probability/density functions and distribution functions can be shown to follow from (7.12). Thus either one can define statistical independence through probability/density functions, distribution functions, moment generating functions, characteristic functions, etc.

**Note 7.5.** The joint characteristic function is available from the joint moment generating function formula by replacing $t_j$ by $it_j$, $i=\sqrt{-1}$, $j=1,\dots,k$. Moment generating function needs not exist all of the time but the characteristic function will always exist.

**Note 7.6.** The joint Laplace transform is available by replacing $t_j$ of the joint moment generating function by $-t_j$, $j = 1, \dots, k$ for positive variables $x_j > 0$, $j = 1, \dots, k$. The joint Mellin transform for positive variables is available as the expected value $E[x_1^{s_1-1} \cdots x_k^{s_k-1}]$, whenever it exists, when $x_1, \dots, x_k$ are positive random variables, where $s_1, \dots, s_k$ are complex parameters.

**Example 7.11.** Let $x_1, \dots, x_k$ be independently distributed normal variables with parameters $(\mu_1, \sigma_1^2), \dots, (\mu_k, \sigma_k^2)$. Let $a_1, \dots, a_k$ be real constants. Find the distribution of the linear function $u = a_1x_1 + \cdots + a_kx_k$.

**Solution 7.11.** Let us compute the moment generating function of $u$ if it exists.

$$\begin{aligned} M_u(t) &= E[\mathrm{e}^{tu}] = E[\mathrm{e}^{t(a_1x_1+\cdots+a_kx_k)}] \\ &= \prod_{j=1}^{k} M_{x_j}(a_jt) \end{aligned}$$

due to product probability property (PPP) or statistical independence of the variables, where $M_{x_j}(t)$ is the moment generating function of $x_j$. Since $x_j$ is normally distributed we have $a_jx_j \sim N(a_j\mu_j, a_j^2\sigma_j^2)$ and, therefore,

$$M_{x_j}(a_jt) = \exp\left[ta_j\mu_j + \frac{t^2}{2}a_j^2\sigma_j^2\right].$$

Hence the product becomes

$$\begin{aligned} M_u(t) &= \prod_{j=1}^{k}\left\{\exp\left[ta_j\mu_j + \frac{t^2}{2}a_j^2\sigma_j^2\right]\right\} \\ &= \exp\left[t\left(\sum_{j=1}^{k} a_j\mu_j\right) + \frac{t^2}{2}\left(\sum_{j=1}^{k} a_j^2\sigma_j^2\right)\right] \end{aligned}$$

which is the moment generating function of a normal variable. Therefore, $u$ is normally distributed with the parameters $\sum_{j=1}^{k}\mu_ja_j = E[u]$ and $\sum_{j=1}^{k} a_j^2\sigma_j^2 = \mathrm{Var}(u)$.

**Definition 7.9** (Independently and identically distributed variables (iid) or a simple random sample). A collection $x_1, \dots, x_n$ of real random variables, which are independently and identically distributed with the common probability/density function $f(x)$, is called a simple random sample of size $n$ from the *population* designated by $f(x)$.

A population can be described by a random variable, such as “a normal population”, “an exponential population”, “a Bernoulli population”, etc., or by a probability/density function or by a distribution (cumulative probability/density) function or by a characteristic function, etc. Thus a simple random sample is a collection of random variables and not a set of numbers, as often misinterpreted. The phrase “simple” is used because there are also other types of random samples such as systematic samples, multistage samples, stratified samples, etc. Here, we are considering only simple random samples.

When the variables satisfy PPP or are independently distributed the joint probability/density function (if it exists) is the product of the marginal probability/density functions, or the joint distribution function (our random variables are defined in terms of distribution functions) factorizes into product of individual distribution functions or the joint characteristic function factorizes into product of individual characteristic functions (always exist) or the joint moment generating function (if it exists) factorizes into individual moment generating functions. Identically distributed means that the functional forms and the parameters are the same for all the probability/density functions, distribution functions, etc.

As an example, if $x_1$ and $x_2$ are iid exponentially distributed variables then the joint density, denoted by $f(x_1,x_2)$, is given by

$$f(x_1,x_2)=\frac{1}{\theta^2}\mathrm{e}^{-\frac{x_1}{\theta}-\frac{x_2}{\theta}}$$

for $0\le x_1<\infty$, $0\le x_2<\infty$, $\theta>0$, and zero elsewhere. Thus the functional forms and the parameters in $x_1$ and $x_2$ are the same. Of course, the variables $x_1$ and $x_2$ are two different variables and do not put $x_1=x_2=x$.

**Example 7.12.** Let $x_1,\dots,x_n$ be iid variables. Let $u=x_1+\cdots+x_n$ and $v=\bar{x}=\frac{x_1+\cdots+x_n}{n}$. Then show that (1) if $x_j$ is gamma random variable then $u$ and $v$ are gamma random variables; (2) if $x_j$ is normal random variable then $u$ and $v$ are normal random variables, or gamma and normal variables are *infinitely divisible*.

**Solution 7.12.** Let $M(t)$ be the moment generating function of $x_j$. If the variables are iid, then all have the same moment generating function $M(t)$. Then

$$\begin{aligned}M_u(t)&=\prod_{j=1}^{n}M_{x_j}(t)\quad\text{due to statistical independence}\\&=[M(t)]^n\quad\text{due to identical distribution.}\end{aligned}$$

When $x_j$ is gamma distributed with the shape parameter $\alpha$ and scale parameter $\beta$, then the moment generating function

$$M(t)=(1-\beta t)^{-\alpha}$$

and, therefore,

$$[M(t)]^n = (1-\beta t)^{-n\alpha}$$

which means that $u$ is gamma distributed with parameters $n\alpha$ and $\beta$. Also note that if $y$ is a gamma random variable with the moment generating function

$$M_y(t) = (1-\beta t)^{-\alpha}$$

then

$$M_y(t) = [\{M_y(t)\}^{\frac{1}{n}}]^n$$

where $M_y(t) = (1-\beta t)^{-\alpha}$, $\{M_y(t)\}^{1/n} = (1-\beta t)^{-\frac{\alpha}{n}}$ are both moment generating functions of gamma variables. In other words, a given gamma variable $y$ can be looked upon as the sum of iid gamma variables. In other words, a gamma variable is infinitely divisible. From the same procedure above, note that $v$ is also gamma distributed.

(2) If $x_j \sim N(\mu,\sigma^2)$ (normally distributed) then

$$M_u(t) = [M_{x_j}(t)]^n = \mathrm{e}^{t(n\mu)+\frac{t^2}{2}(n\sigma^2)}$$

which means that $u$ is normally distributed with the parameters $n\mu$ and $n\sigma^2$. On the other hand,

$$M_{x_j}(t) = [\{M_{x_j}(t)\}^{\frac{1}{n}}]^n$$

where $\{M_{x_j}(t)\}^{\frac{1}{n}}$ is again the moment generating function of a normal with parameters $\frac{\mu}{n}$ and $\frac{\sigma^2}{n}$. In other words, a normal variable can be decomposed into sum of iid normal variables and hence a normal variable is “infinitely divisible”. From the same procedure above, note that $v$ is also normally distributed.

**Definition 7.10** (Infinite divisibility). If a real random variable can be decomposed into the sum of iid variables belonging to the same family, then the variable is said to be infinitely divisible. Equivalently, let $\phi_x(t)$ be the characteristic function of $x$ and if $[\phi_x(t)]^{\frac{1}{n}}$ is also a characteristic function belonging to the same family of functions as $\phi_x(t)$, then $x$ is said to be infinitely divisible.

The concept of infinite divisibility of random variables is very important in probability theory, stochastic process, time series, etc.

### 7.4.3 Linear functions of random variables

In many applications, we need the distribution of linear functions of random variables, and hence we will consider variances of linear functions and covariance between linear functions. Let $x_1,\dots,x_k$ be $k$ random variables with $\mathrm{Var}(x_j) = \sigma_j^2 = \sigma_{jj}$ and covariance between $x_i$ and $x_j$ denoted by $\sigma_{ij}$ and the matrix of variances and covariances by $\Sigma = (\sigma_{ij})$. Let $u = a_1x_1 + \cdots + a_mx_m$ and $v = b_1x_1 + \cdots + b_nx_n$ where $a_1, \ldots, a_m$, $b_1, \ldots, b_n$ are constants, something like $a_1 = 1$, $a_2 = -3$, $a_3 = 2$, $m = 3$ so that $u = x_1 - 3x_2 + 2x_3$; $b_1 = 1$, $b_2 = -1$, $n = 2$ so that $v = x_1 - x_2$. Before looking at the variances, let us examine some of the representations possible.

$$u = a_1x_1 + \cdots + a_mx_m = a'X = X'a$$

where a prime denotes the transpose, $a$ and $X$ are column vectors, whose transposes are

$$a' = (a_1, \ldots, a_m), \quad X' = (x_1, \ldots, x_m).$$

Also, recall the square of a sum and its various representations

$$\begin{aligned}\left[\sum_{j=1}^{k} c_j\right]^2 &= (c_1 + \cdots + c_k)^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} c_ic_j \\ &= \sum_{j=1}^{k} c_j^2 + \sum_{i\neq j} c_ic_j = \sum_{j=1}^{k} c_j^2 + 2\sum_{i<j} c_ic_j \\ &= \sum_{j=1}^{k} c_j^2 + 2\sum_{i>j} c_ic_j \end{aligned} \tag{7.13}$$

where, for example, $\sum_{i>j}$, means the double sum over $i$ and $j$ subject to the condition $i > j$. Now let us see what are the variances of $u$ and $v$ and what is the covariance between $u$ and $v$. We use the basic definitions. By definition,

$$\mathrm{Var}(u) = E[u - E(u)]^2 = E[a'(X - E(X)]^2.$$

Since $a'[X - E(X)]$ is a scalar quantity, it is also equal to its transpose. Therefore, we may write

$$\begin{aligned} E[a'(X - E(X)]^2 &= E[\{a'(X - E(X))\}\{a'(X - E(X))\}'] \\ &= E[\{a'(X - E(X))\}\{X - E(X)\}'a] \\ &= a'E[(X - E(X))(X - E(X))']a = a'\Sigma a. \end{aligned} \tag{7.14}$$

That is, the variance of a linear function is a quadratic form:

$$\mathrm{Var}(a'X) = a'\Sigma a = [a_1, \ldots, a_m]\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}. \tag{7.15}$$

Writing it by using (7.13) or without using matrix notation, we have the following:

$$\mathrm{Var}(u) = \sum_{j=1}^{m} a_j^2 \,\mathrm{Var}(x_j) + 2\sum_{i<j} a_ia_j \,\mathrm{Cov}(x_i, x_j)$$

$$
\begin{aligned}
&= \sum_{j=1}^{m} a_j^2 \sigma_{jj} + 2\sum_{i<j} a_i a_j \sigma_{ij} \\
&= \sum_{j=1}^{m} a_j^2 \operatorname{Var}(x_j) \quad \text{when } x_1, \dots, x_m \text{ are non-correlated;}
\end{aligned}
$$

$$
\operatorname{Var}(\bar{x}) = \sum_{j=1}^{m} \frac{1}{m^2} \operatorname{Var}(x_j) = \frac{\sigma^2}{m} = \operatorname{Var}(\bar{x}) \quad \text{when } x_1, \dots, x_m \text{ are iid.} \tag{7.16}
$$

Those who are familiar with quadratic forms may use the most convenient form in (7.14), and others may use the form in (7.16). Note that the matrix $\Sigma$ in (7.15) is at least positive semi-definite, since it is coming from the form $BB'$ for some matrix $B$.

Let $y_1, \dots, y_n$ be another set of variables and let $v = b_1 y_1 + \cdots + b_n y_n = b'Y$ with $b' = (b_1, \dots, b_n)$ and $Y' = (y_1, \dots, y_n)$. Then from (7.14) it follows that

$$
\operatorname{Var}(v) = \operatorname{Var}(b'Y) = b'\Sigma_y b \tag{7.17}
$$

where $\Sigma_y$ is the covariance matrix in $Y$. The covariance between $u = a'X$ and $v = b'Y$ is then available as

$$
\begin{aligned}
\operatorname{Cov}(u,v) &= E[(u - E(u))((v - E(v))] = a'E[(X - E(X))(Y - E(Y))]b' \\
&= a'\Sigma_{x,y} b'
\end{aligned} \tag{7.18}
$$

where $\Sigma_{x,y}$ is the $m \times n$ matrix

$$
E\left\{ \begin{bmatrix} x_1 - E(x_1) \\ \vdots \\ x_m - E(x_m) \end{bmatrix} [y_1 - E(y_1), \dots, y_n - E(y_n)] \right\} \tag{7.19}
$$

and if $\operatorname{Cov}(v,u)$ is considered then we have

$$
\operatorname{Cov}(v,u) = b'\Sigma'_{x,y} a. \tag{7.20}
$$

When we have two linear forms in the same variables $x_1, \dots, x_k$, that is, $u = a_1 x_1 + \cdots + a_k x_k$ and $v = b_1 x_1 + \cdots + b_k x_k$ then the covariance between $u$ and $v$ is available from (7.17) by putting $X = Y$, $m = n = k$ or

$$
\operatorname{Cov}(u,v) = a'\Sigma b = b'\Sigma a. \tag{7.21}
$$

**Example 7.13.** Compute the (1) Variance of $u$; (2) Variance of $v$; (3) covariance between $u$ and $v$, where $u = 2x_1 - x_2$, $v = x_1 + x_2$, $\operatorname{Var}(x_1) = 1$, $\operatorname{Var}(x_2) = 2$, $\operatorname{Cov}(x_1, x_2) = 1$.

**Solution 7.13.** (1)

$$
\operatorname{Var}(u) = 4\operatorname{Var}(x_1) + \operatorname{Var}(x_2) - 4\operatorname{Cov}(x_1, x_2) = 4(1) + (2) - 4(1) = 2.
$$

(2)

$$\mathrm{Var}(v) = \mathrm{Var}(x_1) + \mathrm{Var}(x_2) + 2\,\mathrm{Cov}(x_1, x_2) = (1) + (2) + 2(1) = 5.$$

The covariance matrix in this case is

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

and hence the covariance between $u$ and $v$, by using the formula

$$a'\Sigma b = [2, -1] \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1.$$

This can also be computed as

$$\begin{aligned} \mathrm{Cov}(x_1, x_2) &= \mathrm{Cov}(2x_1 - x_2, x_1 + x_2) \\ &= 2\,\mathrm{Var}(x_1) + 2\,\mathrm{Cov}(x_1, x_2) - \mathrm{Cov}(x_1, x_2) - \mathrm{Var}(x_2) \\ &= 2(1) + 2(1) - (1) - (2) = 1. \end{aligned}$$

What are the answers if $\mathrm{Cov}(x_1, x_2) = 2$?

**Note 7.7.** If the students have tried to answer the question at the end of Solution 7.13, then they may have noticed that it is not possible to have a covariance between $x_1, x_2$ as 2 if the variances are 1 and 2. The reason is that the covariance matrix, or the variance-covariance matrix has to be at least positive semi-definite when real. If we put $\mathrm{Cov}(x_1, x_2)$ in the above example as 2, then the diagonal elements are positive but the determinant is negative and the matrix is indefinite and hence it cannot be a covariance matrix.

Sometimes the following notation is also used in the literature.

**Notation 7.5.** When $X$ is a $p \times 1$ vector the notation $\mathrm{Cov}(X)$ means the variance-covariance matrix or covariance matrix in $X$, which we already denoted by $\Sigma = (\sigma_{ij})$.

**Definition 7.11** (Correlation coefficient)**.** A scale-free measure of covariance between two real scalar random variables $x$ and $y$ is called *correlation coefficient* and it is denoted by $\rho$ or $\rho_{xy}$ and it is defined as

$$\rho = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}}, \quad \mathrm{Var}(x) \neq 0,\ \mathrm{Var}(y) \neq 0. \tag{7.22}$$

When the covariance is divided by the standard deviations, then the measure has become scale-free because covariance is available only in terms of the units of measurements of $x$ and $y$. Since covariance measures joint scatter in $x$ and $y$ or the scatter or dispersion in the point $(x, y)$, this correlation also can measure only the joint variation. Because of the phrase “correlation” people think that $\rho$ can measure relationship between $x$ and $y$. This is wrong. It cannot measure relationship between $x$ and $y$.

### 7.4.4 Some basic properties of the correlation coefficient

Note that the correlation coefficient is defined only for non-degenerate random variables.

**Result 7.7.** *Whatever be the non-degenerate random variables x and y,*

$$-1 \leq \rho \leq 1. \tag{7.23}$$

The proof is very trivial. Consider two new random variables $u = \frac{x}{\sigma_1} + \frac{y}{\sigma_2}$ and $v = \frac{x}{\sigma_1} - \frac{y}{\sigma_2}$ where $\sigma_1$ and $\sigma_2$ are the standard deviations of $x$ and $y$, respectively.

$$\begin{aligned}\mathrm{Var}(u) &= \mathrm{Var}\left(\frac{x}{\sigma_1}\right) + \mathrm{Var}\left(\frac{y}{\sigma_2}\right) + 2\,\mathrm{Cov}\left(\frac{x}{\sigma_1}, \frac{y}{\sigma_2}\right)\\ &= \frac{\mathrm{Var}(x)}{\sigma_1^2} + \frac{\mathrm{Var}(y)}{\sigma_2^2} + 2\frac{\mathrm{Cov}(x,y)}{\sigma_1\sigma_2} = 1 + 1 + 2\rho\\ &= 2(1+\rho)\end{aligned}$$

since $\mathrm{Cov}(x,y) = \rho\sigma_1\sigma_2$. But the variance of any real random variable is non-negative. Hence $2(1+\rho) \geq 0 \Rightarrow \rho \geq -1$. Similarly, $\mathrm{Var}(v) = 2(1-\rho) \geq 0 \Rightarrow \rho \leq 1$ or $-1 \leq \rho \leq 1$.

**Result 7.8.** *Let $u = ax + b$, $a \neq 0$ and $v = cy + d$, $c \neq 0$ where $a, b, c, d$ are constants. Let $\rho_{xy}$ and $\rho_{uv}$ denote the correlation coefficient between x and y, and u and v, respectively. Then*

$$\rho_{uv} = \frac{ac}{|ac|}\rho_{xy} = \pm\rho_{xy}.$$

**Result 7.9.**

$$\rho_{xy} = \pm 1 \quad \textit{if and only if } y = ax + b,\ a \neq 0,$$

*a linear function of x almost surely.*

Thus the only value of $\rho$, which can be given a physical interpretation, is for $\rho = \pm 1$ and no other point can be given a physical interpretation. Since

$$|\rho| \leq 1 \quad \Rightarrow \quad \mathrm{Cov}(x,y) \leq \sigma_1\sigma_2$$

where $\sigma_1$ and $\sigma_2$ are the standard deviations. [This is also Cauchy–Schwarz inequality.] Thus if the standard deviations are known then the covariance cannot be arbitrary, it has to be less than or equal to the product of the standard deviations. This is an important point to remember in practical applications.

## Exercises 7.4

**7.4.1.** Let $x \sim N(0,1)$, a standard normal variable. Let $y = a + bx + cx^2$, $c \neq 0$ be a quadratic function of $x$. Compute the correlation between $x$ and $y$ here and write it as a function of $b$ and $c$. By selecting $b$ and $c$ show that, while a perfect mathematical relationship existing between $x$ and $y$, as given above, $\rho$ can be made zero, very small $|\rho|$, very large $|\rho|$ (nearly $-1$ or 1, but not equal to $\pm1$). Thus it is meaningless to interpret relationship between $x$ and $y$ based on the magnitude or sign of $\rho$.

**7.4.2.** By using Exercise 7.4.1 show that the following statements are incorrect: "$\rho > 0$ means increasing values of $x$ go with increasing values of $y$ or decreasing values of $x$ go with decreasing values of $y$"; "$\rho < 0$ means the increasing values of $x$ go with decreasing values of $y$ or vice versa;" "$\rho$ near to 1 or $-1$ means near linearity between $x$ and $y$".

**7.4.3.** Compute (1) covariance between $x$ and $y$; (2) $E[xy^2]$; (3) $\rho$ for the following discrete probability function:

$$f(0,-1) = \frac{1}{5}, \quad f(0,1) = \frac{2}{5}, \quad f(1,-1) = \frac{1}{5}, \quad f(1,1) = \frac{1}{5}$$

and $f(x,y) = 0$ elsewhere.

**7.4.4.** Compute (1) $\mathrm{Cov}(x,y)$; (2) $E[x^3y^2]$; (3) $\rho$ for the following density function:

$$f(x,y) = 1, \quad 0 \le x \le 1,\ 0 \le y \le 1$$

and $f(x,y) = 0$ elsewhere.

**7.4.5.** Let $x_1,\dots,x_n$ be a simple random sample of size $n$ from a gamma population with parameters $(\alpha,\beta)$. Let $\bar{x} = (x_1 + \cdots + x_n)/n$. (1) Compute the moment generating function of $\bar{x}$; (2) Show that $\sum_{j=1}^{n} x_j$ as well as $\bar{x}$ are gamma distributed. (3) Compute the moment generating function of $u = \frac{\bar{x}-E[\bar{x}]}{\sqrt{\mathrm{Var}(\bar{x})}}$.

**7.4.6.** (1) Show that $u$ in Exercise 7.4.5 is a re-located, re-scaled gamma random variable for every $n$. (2) Show also that when $n \to \infty$, $u$ goes to a standard normal variable.

**7.4.7.** Going for an interview consists of $t_1$ = time taken for travel to the venue, $t_2$ = waiting to be called for interview and $t_3$ = the actual interview time, thus the total time spent for the interview is $t = t_1 + t_2 + t_3$. It is known from previous experience that $t_1, t_2, t_3$ are independently gamma distributed with scale parameter $\beta = 2$ and $E[t_1] = 6$, $E[t_2] = 8$, $E[t_3] = 6$. What is the distribution of $t$, work out its density.

**7.4.8.** Let $x_1, x_2$ be iid random variables from a uniform population over $[0,1]$. Compute the following probabilities without computing the density of $x_1 + x_2$; (1) $\Pr\{x_1 + x_2 \le 1\}$; (2) $\Pr\{\bar{x} \le \frac{2}{3}\}$; (3) $\Pr\{x_1^2 + x_2^2 \le 1\}$.

**7.4.9.** If the real scalar variables $x$ and $y$ are independently distributed, are the following variables independently distributed? (1) $u = ax$ and $v = by$ where $a$ and $b$ are constants; (2) $u = ax + b$ and $v = cy + d$ where $a, b, c, d$ are constants.

## 7.5 Conditional expectations

Conditional expectations are the expected values in the conditional distributions. In many of the applications such as model building, statistical prediction problems, Bayesian analysis, etc. conditional expectations play vital roles. Hence we will give a brief introduction to conditional expectations here. Two results which are basic will be listed first.

**Result 7.10.** *Whenever all the following expected values exist,*

$$E[y] = E_x[E(y|x)] \tag{7.24}$$

*and*

$$\mathrm{Var}(y) = \mathrm{Var}[E(y|x)] + E[\mathrm{Var}(y|x)]. \tag{7.25}$$

In (7.24), the first expectation, $E(y|x)$, is in the conditional space of $y$ for all given $x$, and then this is treated as a function of $x$ and then expectation is taken with respect to $x$. Thus the outside expectation is in the marginal space of $x$. But (7.25) says that the variance of any variable $y$ is the sum of the expected value of a conditional variance and the variance of a conditional expectation under the assumption that there is a joint distribution of $x$ and $y$ and the expected values exist.

Both (7.24) and (7.25) follow from the definition of expected values and conditional distributions. For the sake of illustration, (7.24) will be proved here for the continuous case. The proof of this for the discrete case and mixed cases and the proof of (7.25) are left as exercises to the students.

$$E(y|x) = \int_y y g(y|x)\mathrm{d}y = \int_y y \frac{f(x,y)}{f_1(x)}\mathrm{d}y$$

where $g(y|x)$ is the conditional density of $y$ given $x$, $f(x,y)$ is the joint density and $f_1(x)$ is the marginal density of $x$. Now, treating the above as a function of $x$ let us take the expected value in the marginal space of $x$. This is available by multiplying with the density of $x$ and integrating out. That is,

$$E[E(y|x)] = \int_x \left[\int_y y \frac{f(x,y)}{f_1(x)}\mathrm{d}y\right] f_1(x)\mathrm{d}x$$

$$= \int_y y\Big[\int_x f(x,y)\mathrm{d}x\Big]\mathrm{d}y$$

after canceling the non-zero part of $f_1(x)$

$$E[E(y|x)] = \int_y yf_2(y)\mathrm{d}y = E[y].$$

Note that when we integrate out $x$ from $f(x,y)$ we get the marginal density $f_2(y)$ of $y$. The proofs for the cases when both $x$ and $y$ discrete or one discrete and one continuous are parallel, and hence left to the students. In computing the above expected values, we assumed the existence of the joint and marginal densities and the existence of the expected values.

**Example 7.14.** For the following joint density,

$$f(x,y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-2-3x)^2}$$

for $-\infty < y < \infty$, $0 \le x \le 1$ and $f(x,y) = 0$ elsewhere, compute (1): $E[y]$; (2): $\mathrm{Var}(y)$.

**Solution 7.14.** From the given statement, it is clear that $x$ has a uniform distribution over $[0,1]$ (available by integrating out $y$ from the joint density) or with the density

$$f_1(x) = 1, \quad 0 \le x \le 1$$

and $f_1(x) = 0$ elsewhere. We can compute the mean value and variance of $x$ from this marginal density. That is,

$$E[x] = \int_0^1 x\mathrm{d}x = \frac{1}{2} \quad \text{and similarly } \mathrm{Var}(x) = \frac{1}{12}.$$

If we divide the given function by $f_1(x)$, we should get the conditional density of $y$ given $x$, denoted by $g(y|x)$. That is,

$$g(y|x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-2-3x)^2}, \quad -\infty < y < \infty$$

or $y|x \sim N(\mu = 2+3x, \sigma^2 = 1)$, that is, the conditional density of $y$ given $x$ is normal with mean value $2 + 3x$ and variance 1. Therefore, the conditional expectation of $y$ given $x$ or the conditional mean value, which is the expected value of $y$ in the conditional density of $y$, given $x$, is

$$E(y|x) = 2 + 3x, \tag{i}$$

and the conditional variance or the variance in the conditional density, in this case is $\mathrm{Var}(y|x) = 1$.

Then the expected value of this conditional expectation (expectation is taken in the marginal space $x$) and the variance of this conditional expectation are respectively

$$E[E(y|x)] = E[2+3x] = 2 + E[x] = 2 + \frac{3}{2} = \frac{7}{2} \tag{ii}$$

$$\mathrm{Var}[E(y|x)] = \mathrm{Var}[2+3x] = 3^2\,\mathrm{Var}(x) = \frac{9}{12} = \frac{3}{4}. \tag{iii}$$

In order to get the marginal density of $y$, from the joint density $f(x,y)$, one has to integrate out $x$. Here, the integral is over $[0,1]$ since $x$ is uniformly distributed. Hence no analytic form for the marginal density of $y$ is available from our joint density. But we can compute $E(y)$ and $\mathrm{Var}(y)$ by using Result 7.10. From (ii) above,

$$E[y] = E[E(y|x)] = \frac{7}{2}.$$

and

$$\mathrm{Var}(y) = E[\mathrm{Var}(y|x)] + \mathrm{Var}[E(y|x)] = E[1] + \frac{3}{4} = \frac{7}{4}.$$

**Example 7.15.** For the following joint density,

$$f(x,y) = \frac{1}{x^2\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-2-3x)^2}$$

for $-\infty < y < \infty$, $1 \le x < \infty$ and $f(x,y) = 0$ elsewhere, compute (1) $E(y)$; (2) $\mathrm{Var}(y)$.

**Solution 7.15.** The situation is similar to Example 7.14. Here, the marginal density is given by (available by integrating out $y$ from the joint density)

$$f_1(x) = \frac{1}{x^2}, \quad 1 \le x < \infty$$

and zero elsewhere. Therefore, the conditional density of $y$ given $x$, which is $f(x,y)/f_1(x)$, is given by the normal density part, excluding $\frac{1}{x^2}$. Therefore,

$$y|x \sim N(\mu = 2+3x, \sigma^2 = 1)$$

which gives

$$E(y|x) = 2+3x \quad \text{and} \quad \mathrm{Var}(y|x) = 1. \tag{i}$$

In order to compute $E(y)$ and $\mathrm{Var}(y)$ by using Result 7.10, we need to compute $E(x)$ and $\mathrm{Var}(x)$. Then

$$E(x) = 0 + \int_1^\infty \frac{x}{x^2}\mathrm{d}x = \int_1^\infty \frac{1}{x}\mathrm{d}x = [\ln x]_1^\infty = \infty.$$

Therefore, $E(x)$ does not exist, and hence Result 7.10 cannot be used to calculate $E(y)$ and $\mathrm{Var}(y)$.

**Note 7.8.** If $E(x)$ does not exist, then all higher moments $E(x^\alpha)$, $\alpha \ge 1$ will not exist. If $E(x^m)$ does not exist, but all lower moments up to $m$ exist then also all moments $E(x^\beta)$, $\beta \ge m$ will not exist.

### 7.5.1 Conditional expectation and prediction problem

An important use of the conditional expectation is in the area of prediction. An agriculturist may want to predict the yield of tapioca under a certain chemical fertilizer and would like to answer a question such as what will be the yield if 200 grams of that fertilizer is used? Here, let $y$ be the yield and $x$ be the amount of fertilizer used. Then the question is: what is the predicted yield $y$ if $x$ is fixed at 200, or given $x = 200$? As another example, someone may be trying to reduce weight by exercise. Here, $y$ is the reduction in weight and $x$ is the number of hours spent on daily exercise. What is likely to be the reduction $y$ if the exercise is 30 minutes daily, or $x = 30$, $x$ being measured in minutes? As another example, suppose we want to look at the cost of living index. Cost of living for a household in a certain township depends on many items such as per unit price of rice, say $x_1$, per unit price of vegetables, $x_2$, per unit price of milk, $x_3$, transportation cost $x_4$ etc. If cost of living is denoted by $y$, then it depends on many variables, $x_1, \dots, x_k$. What is the best function $g(x_1, \dots, x_k)$ to predict $y$, where $g$ is some function and we want the best function. We would like to use this function to predict $y$ at preassigned values of $x_1, \dots, x_k$, something like answering a question such as: What is the cost of living if price per kilogram of rice is Rs 20, price per kilogram of vegetables is Rs 15, etc. or at preassigned values of $x_1 = 20$, $x_2 = 15$, etc.

Theoretically, any function of $x_1, \dots, x_k$ can be used as a predictor but our prediction may be far off from the true value. If an arbitrary function $g(x_1, \dots, x_k)$, such as $g = 1 + x_1 + 2x_2$, is used to predict $y$ then the error $\epsilon$ in this prediction is $\epsilon = y - g(x_1, \dots, x_k)$ or $g(x_1, \dots, x_k) - y$. One criterion that one can use to come up with a good function as a predictor is to minimize the distance between $y$ and $g$. A mathematical distance between the random variable $y$ and $g(x_1 = a_1, \dots, x_k = a_k)$, where $a_1, \dots, a_k$ are the preassigned values of $x_1, \dots, x_k$, is

$$\{E[y - g(x_1 = a_1, \dots, x_k = a_k)]^2\}^{\frac{1}{2}}. \tag{i}$$

But minimization of (i), over all possible functions $g$ is equivalent to minimizing its square or minimizing

$$E[y - g]^2 \tag{7.26}$$

over all $g$. Since $g$ is evaluated at given points $x_1 = a_1, \dots, x_k = a_k$, it is equivalent to minimizing $E[y - a]^2$ for arbitrary $a$. This is already done in Chapter 3 and we have seen that the best value of $a$ is given by $a = E(y)$ at given values of $x_1, \dots, x_k$. Hence the best predictor is the conditional expectation of $y$ given $x_1, \dots, x_k$.

$$\min_g E[y - g(x_1 = a_1, \dots, x_k = a_k)]^2 \quad \Rightarrow \quad g = E[y|x_1, \dots, x_k]. \tag{7.27}$$

Therefore, the conditional expectation of $y$ given $x_1, \dots, x_k$ is the “best” predictor of $y$ at preassigned values of $x_1, \dots, x_k$, best in the sense of minimizing the mean (expected value) square error or *in the minimum mean square sense*. If other measures of "distance" are used, then we will come up with other rules (other functions). Since the mean squared error is a mathematically convenient form, (7.27) is taken as the "best" predictor.

Note that it may be possible to evaluate a conditional expectation if we have the joint distribution or at least the conditional distribution. In a practical situation, usually neither the joint distribution nor the conditional distribution may be available. In that case, we may try to estimate the prediction function by imposing desirable conditions.

**Example 7.16.** In villages across a state, it is found that the proportion $x_1$ of people having health problems and the proportion $x_2$ of people who are overweight (weight over the prescribed value by health standards) have a joint distribution given by the density

$$f(x_1,x_2) = x_1 + x_2, \quad 0 \le x_1 \le 1,\ 0 \le x_2 \le 1$$

and zero elsewhere. (1) Construct the best predictor of $x_1$ the proportion of people with health problems, by using $x_2$ the proportion of overweight people. (2) What is the predicted value of $x_1$ if a village selected at random has 30% of overweight people.

**Solution 7.16.** We have already evaluated the marginal and conditional densities in this case. The conditional density of $x_1$ given $x_2$, denoted by $g_1(x_1|x_2)$, is given by

$$g_1(x_1|x_2) = \frac{x_1 + x_2}{x_2 + \frac{1}{2}}, \quad 0 \le x_1 \le 1$$

and zero elsewhere. Hence the conditional expectation is

$$E[x_1|x_2] = \frac{1}{x_2 + \frac{1}{2}} \int_0^1 x_1(x_1 + x_2)\mathrm{d}x_1$$
$$= \frac{\frac{1}{3} + \frac{x_2}{2}}{x_2 + \frac{1}{2}}$$

is the best predictor of $x_1$ at preassigned values of $x_2$. (2) The best predicted value of $x_1$ at $x_2 = 0.3$ is then given by

$$\frac{\frac{1}{3} + \frac{x_2}{2}}{x_2 + \frac{1}{2}}\bigg|_{x_2=0.3} = \frac{29}{48}.$$

Here, the joint density was available. Now, we will consider a case where we have only a conditional density.

**Example 7.17.** The waiting time $t$ at a bus stop is known to be exponentially distributed but the expected waiting time is a function of the delay $t_1$ due to traffic congestion on the way. The conditional density of $t$ given $t_1$ is known to be of the form

$$g(t|t_1) = \frac{1}{3+2t_1}\mathrm{e}^{-\frac{t}{3+2t_1}}, \quad 0 \le t < \infty$$

and zero elsewhere. (1) Construct the best predictor function of $t_1$ for predicting $t$; (2) What is the best predicted $t$ if the traffic congestion delay is 5 minutes, time being measured in minutes.

**Solution 7.17.** For an exponential density, it was seen that the expected value is the parameter itself. Hence

$$E(t|t_1) = 3 + 2t_1$$

is the best predictor of $t$ at given values of $t_1$. For $t_1 = 5$, the best predicted value is $3 + 2(5) = 13$ minutes.

### 7.5.2 Regression

The word “regression” means to regress or to go back. This name came in because the original problem considered was to say something about ancestors by studying the offsprings. But now a days, “regression” means the area of predicting one variable by using one or more other variables. We have already seen that if we use the criterion of minimum mean square error then the conditional expectation is the best predictor function. Hence regression is defined as the conditional expectation.

**Definition 7.12** (Regression of $y$ on $x_1, \ldots, x_k$). The regression of $y$ on $x_1, \ldots, x_k$ is the best predictor function for $y$, best in the minimum mean square sense, at preassigned values of $x_1, \ldots, x_k$ and it is defined as the conditional expectation of $y$ given $x_1, \ldots, x_k$ or $E[y|x_1, \ldots, x_k]$, which is a function of $x_1, \ldots, x_k$.

Regression analysis is an area which is often misused and misinterpreted in statistical analysis. Regression is often misinterpreted as model building by using the method of least squares. It is not a model building problem but it is a search for the best predictor. Since regression is defined as a conditional expectation, regression analysis is done in the conditional space, the whole joint space of all the variables is not necessary to do regression analysis. But for correlation analysis we need the whole space of joint distributions and thus correlation analysis and regression analysis are not one and the same. As seen above, the best predictor or regression function can be constructed if either the joint distribution is available or the conditional distribution is available. We have done examples of both. If we do not have either the joint distribution or the conditional distribution, then the regression function cannot be evaluated. But sometime we may have some idea about the conditional expectation that at given values of $x_1,\dots,x_k$ may have a certain functional form for the expected value of $y$, such as a linear function of $x_1,\dots,x_k$ or a polynomial type function, etc. In that case, we may try to estimate that regression function by the help of the method of least squares. This aspect will be considered in later chapters, and hence further discussion will not be attempted here.

## Exercises 7.5

**7.5.1.** Prove (7.24) for the discrete case and (7.25) for both discrete and continuous cases.

**7.5.2.** For the joint density of $x$ and $y$,

$$f(x,y)=\frac{1}{1+x}\mathrm{e}^{-\frac{y}{1+x}},\quad 0\le y<\infty,\ 0\le x\le 1$$

and $f(x,y)=0$ elsewhere, compute (1) $E(y)$; (2) $\mathrm{Var}(y|x)$; (3) $\mathrm{Var}(y)$; (4) the marginal density of $y$.

**7.5.3.** For the joint density

$$f(x,y)=\frac{2}{(1+x)x^3}\mathrm{e}^{-\frac{y}{1+x}}$$

for $0\le y<\infty$, $1\le x<\infty$ and zero elsewhere, compute (1) $E(y)$; (2) $\mathrm{Var}(y|x)$; (3) $\mathrm{Var}(y)$; (4) the marginal density of $y$.

**7.5.4.** Construct an example of a joint density of $x$ and $y$ where $E(y|x)=1+x+2x^2$ and (a) $E(y)$ exists but $E(y^2)$ does not exist; (b) $E(y)$ does not exist.

**7.5.5.** Construct the regression function of $x_1$ on $x_2,x_3$ and show that it is free of the regressed variables $x_2$ and $x_3$ in the following joint density, why is it free of $x_2$ and $x_3$?

$$f(x_1,x_2,x_3)=c\mathrm{e}^{-2x_1-5x_2-3x_3}$$

for $0\le x_1,x_2,x_3<\infty$ and zero elsewhere.

## 7.6 Bayesian procedure

Another area which is based on the conditional distribution is Bayesian procedures and Bayesian inference. The name suggests that it has something to do with Bayes' theorem, which was considered in Chapter 2, which had to do with inverse reasoning or going from the effect to cause. After observing an event, we are asking about the cause for the occurrence of that event. Here, we look at the same Bayes' rule in terms of random variables. Let $f(x,y)$ be the joint density/probability function of two random variables. Let $f_1(x)$ and $f_2(y)$ be the marginal density/probability functions and $g_1(x|y)$ and $g_2(y|x)$ be the conditional density/probability functions. Then we have the following relations:

$$\begin{aligned} g_1(x|y) &= \frac{f(x,y)}{f_2(y)} = \frac{g_2(y|x)f_1(x)}{f_2(y)} \quad \Rightarrow \\ g_2(y|x) &= \frac{g_1(x|y)f_2(y)}{f_1(x)}. \end{aligned} \tag{7.28}$$

Let us interpret (7.28) in terms of one variable and one parameter. Let $y$ be a parameter in a probability/density function of a real scalar random variable $x$, something like $x$ at a fixed $\theta$ may be an exponential density

$$g_1(x|\theta) = \theta e^{-\theta x}, \quad 0 \le x < \infty,\ \theta > 0$$

and $g_1(x|\theta) = 0$ elsewhere. But $\theta$ may have its own distribution. Suppose that $\theta$ has a gamma density with known shape parameter $\alpha$ and scale parameter $\beta$. Then

$$f_2(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}e^{-\frac{\theta}{\beta}}.$$

Then (7.28) in this context becomes

$$g_2(\theta|x) = \frac{g_1(x|\theta)f_2(\theta)}{f_1(x)} = \frac{f(x,\theta)}{f_1(x)}. \tag{7.29}$$

How do we get the marginal density of $x$, namely $f_1(x)$, from the joint density $f(x,\theta)$ of $x$ and $\theta$, simply integrate out or sum up the other variable, namely $\theta$. That is,

$$\begin{aligned} f_1(x) &= \int_\theta f(x,\theta)d\theta = \int_\theta g_1(x|\theta)f_2(\theta)d\theta \quad \text{(continuous case)} \\ &= \sum_\theta f(x,\theta) = \sum_\theta g_1(x|\theta)f_2(\theta) \quad \text{(discrete case).} \end{aligned}$$

Here, $g_1(x|\theta)$ is the conditional probability/density function of $x$ given $\theta$ and $f_1(x)$ is the unconditional probability/density function of $x$. Thus (7.29) can be looked upon as a connection between conditional and unconditional probability/density functions of $x$. As far as $\theta$ is concerned, $f_2(\theta)$ is the prior probability/density of $\theta$ whereas $g_2(\theta|x)$ is the posterior, in the sense of after observing $x$, probability/density function of $\theta$. Then what is the best predictor, best in the minimum mean square sense, of $\theta$ in the light of the given observation on $x$? We have seen from Section 7.5 that it is the conditional expectation. That is,

$$\begin{aligned} E(\theta|x) &= \text{best predictor of } \theta, \text{ given } x \\ &= \text{Bayes' predictor of } \theta. \end{aligned}$$

In the example of $x|\theta$ being an exponential variable and $\theta$ being a gamma variable, we have the following computations:

$$\begin{aligned}
f(x,\theta) &= g_1(x|\theta)f_2(\theta)\\
&= \theta \mathrm{e}^{-\theta x} \times \frac{1}{\beta^\alpha \Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\frac{\theta}{\beta}} = \frac{1}{\beta^\alpha \Gamma(\alpha)}\theta^{\alpha}\mathrm{e}^{-\theta(x+\frac{1}{\beta})}\\
f_1(x) &= \int_\theta f(x,\theta)\mathrm{d}\theta = \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty \theta^\alpha \mathrm{e}^{-\theta(x+\frac{1}{\beta})}\mathrm{d}\theta\\
&= \frac{\Gamma(\alpha+1)(x+\frac{1}{\beta})^{-(\alpha+1)}}{\beta^\alpha\Gamma(\alpha)}. \qquad (7.30)
\end{aligned}$$

This is the unconditional density of $x$ in this case. The posterior density of $\theta$ is given by

$$g_2(\theta|x) = \frac{f(x,\theta)}{f_1(x)} = \frac{(x+\frac{1}{\beta})^{\alpha+1}}{\Gamma(\alpha+1)}\theta^\alpha \mathrm{e}^{-\theta(x+\frac{1}{\beta})}$$

for $0 < \theta < \infty$ and $g_2(\theta|x) = 0$ elsewhere.

What is the best predictor of $\theta$ in the presence of an observation on $x$ or at given $x$? It is the conditional expectation of $\theta$, given $x$. In the above example,

$$\begin{aligned}
E(\theta|x) &= \int_\theta \theta g_2(\theta|x)\mathrm{d}\theta = \frac{(x+\frac{1}{\beta})^{\alpha+1}}{\Gamma(\alpha+1)}\int_{\theta=0}^\infty \theta^{\alpha+1}\mathrm{e}^{-\theta(x+\frac{1}{\beta})}\mathrm{d}\theta\\
&= \frac{(x+\frac{1}{\beta})^{\alpha+1}}{\Gamma(\alpha+1)}\left(x+\frac{1}{\beta}\right)^{-(\alpha+2)}\Gamma(\alpha+2) = \frac{\alpha\beta+\beta}{1+\beta x}. \qquad (7.31)
\end{aligned}$$

What is the mean value of $\theta$ before observing $x$? It is $E(\theta)$ from the prior density of $\theta$, which is $E(\theta) = \alpha\beta$. Thus the mean value $\alpha\beta$ is changed to $\frac{\alpha\beta+\beta}{1+\beta x}$ in the presence of an observation on $x$.

**Example 7.18.** If a student is selected at random from a last year high school class or from the community of such last year high school classes, then the probability $p$ that he/she will answer a question correctly depends upon the background preparation, exposure to the topic, basic intelligence, etc. For one student, this probability $p$ may be 0.8, for another it may be 0.3, for another it may be 0.9, etc. This $p$ is a varying quantity. $p$ may have its own distribution. If a student is selected at random, then $p$ for this student is a fixed quantity. If 10 questions of similar types are asked what is the chance that there will be $x$ correct answers, something like 8 correct answers? Assume that $p$ has a prior type-1 beta distribution with known parameters $\alpha$ and $\beta$.

**Solution 7.18.** Using the standard notation, the probability function of $x$ for a given $p$ is binomial:

$$g_1(x|p) = \binom{n}{x}p^x(1-p)^{n-x}, \quad x = 0,1,\dots,n;\ 0 < p < 1$$

and zero elsewhere. In our example, there are 10 questions, then $n = 10$, with the probability of the correct answer is $p$ and the number of correct answers is $x = 8$. We assumed that $p$ has a prior type-1 beta density. Then the joint probability function of $x$ and $p$, denoted by $f(x,p)$ is given by

$$\begin{aligned} f(x,p) &= g_1(x|p)f_2(p) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\binom{n}{x} p^x(1-p)^{n-x} \end{aligned}$$

for $\alpha > 0$, $\beta > 0$, $x = 0,1,\dots,n$; $0 < p < 1$ and zero elsewhere. Then the unconditional probability function of $x$, denoted by $f_1(x)$, is available as

$$\begin{aligned} f_1(x) &= \int_p f(x,p)\mathrm{d}p \\ &= \frac{\binom{n}{x}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\mathrm{d}p \\ &= \frac{\binom{n}{x}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}, \quad x = 0,1,\dots,n \end{aligned}$$

Then what is the density of $p$ for given $x$ or the conditional density of $p$ for a given $x$? Let it be $g_2(p|x)$. Then

$$\begin{aligned} g_2(p|x) &= \frac{f(x,p)}{f_1(x)} \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1}, \quad 0 < p < 1 \end{aligned}$$

What is the expected value of $p$ given $x$?

$$\begin{aligned} E(p|x) &= \int_0^1 p g_2(p|x)\mathrm{d}p \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \frac{\Gamma(\alpha+1+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+1+n)} \\ &= \frac{\alpha+x}{\alpha+\beta+n}. \end{aligned}$$

This is the Bayes' estimate of $p$ or the best predictor of $p$ at given $x$. If $p$ was fixed, then we would have estimated $p$ by the sample proportion, namely $\frac{x}{n}$. In the light of a prior type-1 beta distribution for $p$, the estimate has changed to $\frac{\alpha+x}{\alpha+\beta+n}$.

In the above example, what is the distinction between the two estimates for $p$. $\frac{x}{n}$ is the estimate for the probability of success for a given student. If one student is selected at random and she gave 7 correct answers out of 10 questions of similar difficulties then $\frac{7}{10} = 0.7$ is the estimate for her probability of success. When $p$ has its own distribution, then we are considering the probability of success in the population of such final year students across the spectrum. What is the estimate of this $p$ across the spectrum, given the information that one girl from this spectrum gave correct answers for 7 out of 10 questions. Then the estimated value of $p$ is

$$\frac{\alpha+7}{\alpha+\beta+10} = \frac{1.5+7}{1.5+3.7+10} = \frac{85}{152}$$

if $\alpha = 1.5$ and $\beta = 3.7$. Note that the Bayes' estimate for $p$ here, $\frac{\alpha+x}{\alpha+\beta+n}$, can be smaller or bigger than $\frac{x}{n}$ depending upon the values of $\alpha$ and $\beta$.

## Exercises 7.6

**7.6.1.** Let $x$ given $\lambda > 0$ be a Poisson random variable with parameter $\lambda$. Let $\lambda$ have a prior gamma distribution. Compute (1) the unconditional probability function of $x$; (2) the posterior density of $\lambda$ given $x = 3$; (3) Bayes' estimate of $\lambda$.

**7.6.2.** Let $x$ given $b$ be generalized gamma with density of the form

$$g_1(x|b) = cx^{\alpha-1}\mathrm{e}^{-bx^{\delta}}, \quad x \geq 0,\ \delta > 0,\ \alpha > 0$$

and $c$ is the normalizing constant. Let $b$ have a gamma distribution. Then answer (1), (2), (3) of Exercise 7.6.1.

**7.6.3.** Let $x|\mu \sim N(\mu, 1)$ and let $\mu \sim N(0, 1)$. Answer (1), (2), (3) of Exercise 7.6.1.

**7.6.4.** Let $x|a$ be uniformly distributed over $[0, a]$. Let $a$ have a prior Pareto density $\frac{c}{a^5}$, $2 < a < 4$ where $c$ is the normalizing constant. Answer (1), (2), (3) of Exercise 7.6.1.

**7.6.5.** Let $x|p$ be binomial with parameters $(n = 10, p)$. Let $p$ have a prior power function density $f_2(p) = cp^5$, $0 < p < 0.7$ where $c$ is the normalizing constant. Answer (1), (2), (3) in Exercise 7.6.1.

## 7.7 Transformation of variables

In Chapter 6, Section 6.8, we considered transformation of variable involving one variable. Given the density $f(x)$ of a real random variable $x$ how to compute the density of $y = \phi(x)$ where $x$ to $y$ is a one to one function of $x$ or $x$ can be uniquely written as $x = \phi^{-1}(y)$. If $g(y)$ is the density of $y$, then we have seen that the relation is $f(x)\mathrm{d}x = g(y)\mathrm{d}y$ if $y$ is an increasing function of $x$ and $f(x)\mathrm{d}x = -g(y)\mathrm{d}y$ if $y$ is a decreasing function of $x$. In the discrete case, there is no Jacobian of transformation and $g(y)$, the probability function of $y$, can be computed by looking at the possible values $y$ can take and then computing the corresponding probabilities by using $f(x)$, the probability function of $x$.

Here, we will consider transformations when the real scalar random variables $x_1,\dots,x_k$ have a joint distribution. Consider the transformation

$$\begin{aligned}
y_1 &= \phi_1(x_1,\dots,x_k) = \phi_1(X)\\
y_2 &= \phi_2(x_1,\dots,x_k) = \phi_2(X)\\
\vdots &= \vdots\\
y_k &= \phi_k(x_1,\dots,x_k) = \phi_k(X)
\end{aligned}$$

Let $Y' = (y_1,\dots,y_k)$ and let the transformation be written as $Y = \phi_*(X)$. Then if the transformation $X$ to $Y$ is one to one then we can write $X$ uniquely in terms of $Y$ as $X = \phi_*^{-1}(Y)$. When all variables $x_1,\dots,x_k$ are discrete and if there is a transformation (need not be one to one) then the probability function $g(Y)$ of $Y$ can be computed parallel to the one variable case. Look at all possible values $Y$ can take then compute the corresponding probabilities by using the probability function $f(X)$ of $X$. This will give $g(Y)$. When all variables $x_1,\dots,x_k$ or $X$ are continuous and if $X$ to $Y$ is a one to one transformation, then there is a Jacobian of the transformation and the connection between the density $f(X)$ of $X$ and $g(Y)$ of $Y$ is that

$$f(X)\mathrm{d}X = g(Y)\mathrm{d}Y,$$

where

$$\begin{aligned}
&\mathrm{d}X = \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_k, \quad \mathrm{d}Y = \mathrm{d}y_1 \wedge \cdots \wedge \mathrm{d}y_k\\
&\mathrm{d}Y = J\mathrm{d}X, \quad J = \left|\left(\frac{\partial\phi_i}{\partial x_j}\right)\right|
\end{aligned}$$

is the determinant of the matrix of partial derivatives $\frac{\partial\phi_i}{\partial x_j}$. Then

$$\begin{aligned}
g(Y)\mathrm{d}Y = f(X)\mathrm{d}X \quad &\Rightarrow \quad g(Y)J\mathrm{d}X = f(X)\mathrm{d}X\\
g(Y) = \frac{1}{J}f(\phi^{-1}(Y)).&
\end{aligned} \tag{7.32}$$

Let us do some simple examples to see the significance of the relationship in (7.32). Before taking up continuous cases, let us do one discrete case for the sake of illustration.

**Example 7.19.** Consider the transformation $(y_1 = x_1^2, y_2 = 2x_1 + x_2)$ and compute the joint probability function $g(y_1,y_2)$ when $x_1,x_2$ have the joint probability function

$$f(x_1,x_2) = \begin{cases} 1/10, & \text{for } (x_1 = 0, x_2 = 1)\\ 2/10, & \text{for } (x_1 = 0, x_2 = 2)\\ 2/10, & \text{for } (x_1 = -1, x_2 = 1)\\ 5/10, & \text{for } (x_1 = -1, x_2 = 2)\\ 0, & \text{elsewhere}\end{cases}$$

**Solution 7.19.** The possible values $y_1 = x_1^2$ can take are $y_1 = 0, 1$. The possible values $y_2$ can take are $y_2 = 2(0) + 1 = 1$; $2(0) + 2 = 2$; $2(-1) + 1 = -1$; $2(-1) + 2 = 0$. Hence we have $(y_1, y_2) = (0, 1)$ with probability $\frac{1}{10}$; $(y_1, y_2) = (0, 2)$ with probability $\frac{2}{10}$, and so on. No points here coincide, and hence the points are all distinct. [If some points coincided, then we should add up the corresponding probabilities.] Hence the joint probability function $g(y_1, y_2)$ is given as

$$g(y_1,y_2) = \begin{cases} 1/10, & \text{for } (y_1,y_2) = (0,1) \\ 2/10, & \text{for } (y_1,y_2) = (0,2) \\ 2/10, & \text{for } (y_1,y_2) = (1,-1) \\ 5/10, & \text{for } (y_1,y_2) = (1,0) \\ 0, & \text{elsewhere.} \end{cases}$$

**Example 7.20.** Let $x_1$ and $x_2$ be independently distributed as uniform random variables over $[0, a]$ and $[0, b]$, respectively, that is, the densities of $x_1$ and $x_2$, denoted by $f_1(x_1)$ and $f_2(x_2)$, respectively, are $f_1(x_1) = \frac{1}{a}$, $0 \le x_1 \le a$ and $f_2(x_2) = \frac{1}{b}$, $0 \le x_2 \le b$ and zero elsewhere. Compute the densities of (1) $u = x_1 + x_2$; (2) $v = x_1 x_2$; (3) $w = \frac{x_1}{x_2}$.

**Solution 7.20.** Due to product probability property or statistical independence, the joint density, denoted by $f(x, y)$, is the product, that is,

$$f(x,y) = \frac{1}{ab}, \quad 0 \le x_1 \le a, \quad 0 \le x_2 \le b$$

and zero elsewhere. Let us make the transformation $y_1 = x_1 + x_2, y_2 = x_2$ so that it is a one to one transformation $x_2 = y_2, x_1 = y_1 - y_2$. It is a linear transformation with the matrix of the transformation is

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \Rightarrow \quad \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1.$$

Hence if $g(y_1, y_2)$ is the joint density of $y_1$ and $y_2$, then it is the same as $\frac{1}{ab}$ but the region in the $(y_1, y_2)$-plane will be different. The region will be the region bounded by the lines $x_1 = 0 \Rightarrow y_1 - y_2 = 0$, $x_1 = a \Rightarrow y_1 - y_2 = a$, $x_2 = 0 \Rightarrow y_2 = 0$, $x_2 = b \Rightarrow y_2 = b$. Thus the rectangle in the $(x_1, x_2)$-plane transforms to a parallelogram in the $(y_1, y_2)$-plane as shown in Figure 7.5.

The marginal density of $y_1$ is available by integrating out $y_2$. From Figure 7.5, note that when $0 \le y_1 \le a$ the integration of $y_2$ is from 0 to $y_1$. That is,

$$\int_{y_2=0}^{y_1} \frac{1}{ab} \mathrm{d}y_2 = \frac{y_1}{ab}, \quad 0 \le y_1 \le a.$$

When $b$ is greater than $a$, then in the interval $a \le y_1 \le b$, the integration is from $y_2 = y_1 - a$ to $y_1$. When $b \le y_1 \le a + b$, then the integration is from $y_1 - a$ to $b$. That is,

$$\int_{y_2=y_1-a}^{y_1} \frac{1}{ab} \mathrm{d}y_2 = \frac{a}{ab}, \quad a \le y_1 \le b$$

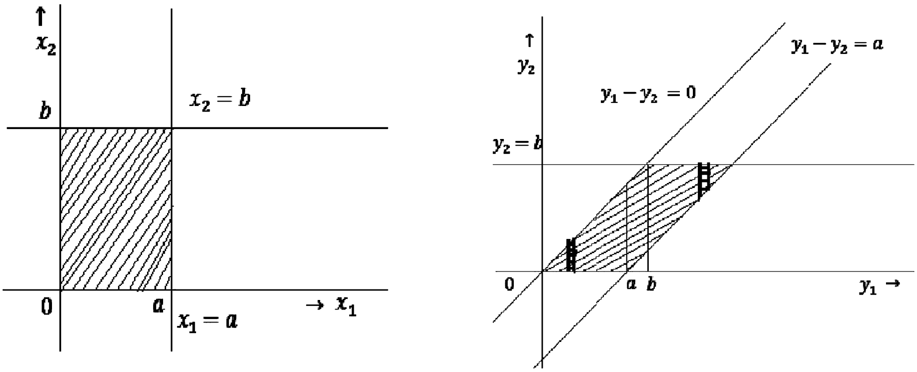


Figure 7.5: Left: Region in $(x_1,x_2)$-plane; Right: Region in $(y_1,y_2)$-plane.

and

$$\int_{y_2=y_1-a}^{b} \frac{1}{ab}\mathrm{d}y_2 = \frac{b+a-y_1}{ab}.$$

Thus the density of $y_1$, denoted by $g_1(y_1)$, for $b > a$, is given by

$$g_1(y_1) = \begin{cases} \frac{y_1}{ab}, & 0 \le y_1 \le a \\ \frac{a}{ab}, & a \le y_1 \le b \\ \frac{a+b-y_1}{ab}, & b \le y_1 \le a+b \\ 0, & \text{elsewhere.} \end{cases}$$

We can verify that it is a density by integrating out and showing that it gives 1. That is,

$$\begin{aligned}\int_0^{a+b} g_1(y_1)\mathrm{d}y_1 &= \frac{1}{ab}[\int_0^a y_1\mathrm{d}y_1 + a\int_a^b \mathrm{d}y_1 + \int_b^{a+b}[(a+b)-y_1]\mathrm{d}y_1 \\ &= \frac{1}{ab}\left\{\frac{a^2}{2} + a(b-a) + (a+b)a - \frac{1}{2}[(a+b)^2 - b^2]\right\} \\ &= \frac{ab}{ab} = 1.\end{aligned}$$

Now, let us look at the density of $v = x_1x_2$. Again, let us use the same notations. Let $y_1 = x_1x_2, y_2 = x_2$, which means $x_2 = y_2, x_1 = \frac{y_1}{y_2}$. Then the Jacobian of the transformation is $\frac{1}{x_2} = \frac{1}{y_2}$.

From Figure 7.6, observe that in the $(y_1,y_2)$-plane the integration of $y_2$ to be done from $y_2 = \frac{y_1}{a}$ to $b$. The joint density of $x_1$ and $x_2$ is $f(x_1,x_2) = \frac{1}{ab}$ for $0 \le x_1 \le a$, $0 \le x_2 \le b$. The joint density of $y_1$ and $y_2$, denoted by $g(y_1,y_2)$, is then $g(y_1,y_2) = \frac{1}{aby_2}$, including the Jacobian and then the marginal density of $y_1$ is given by

$$g_1(y_1) = \frac{1}{ab}\int_{y_1/a}^{b} \frac{1}{y_2}\mathrm{d}y_2$$

$$= \frac{1}{ab}[\ln y_2]_{y_1/a}^{b} = \frac{1}{ab}[\ln ab - \ln y_1], \quad 0 \le y_1 \le ab$$

and zero elsewhere. Let us see whether it is a density function by checking to see whether the total integral is 1 since it is already non-negative.

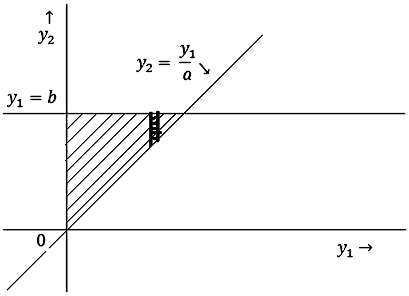


**Figure 7.6:** Region in the $(y_1,y_2)$-plane.

$$\begin{aligned}\int_0^{ab} g_1(y_1)\mathrm{d}y_1 &= \frac{1}{ab}\int_0^{ab}[\ln ab - \ln y_1]\mathrm{d}y_1 \\ &= \frac{1}{ab}\{(\ln ab)(ab) - [y_1 \ln y_1 - y_1]_0^{ab}\} \\ &= \frac{ab}{ab} = 1.\end{aligned}$$

Hence it is a density.

Now we look at the density of $w = \frac{x_1}{x_2}$. Again, we use the same notations. Let $y_1 = \frac{x_1}{x_2}$, $y_2 = x_2$, which means, $x_2 = y_2$, $x_1 = y_1 y_2$. Then the Jacobian is $y_2$. Therefore, the joint density of $y_1$ and $y_2$ is

$$g(y_1,y_2) = \frac{y_2}{ab}.$$

The region in the $(y_1,y_2)$-plane is given in Figure 7.7.

Then $x_2 = 0 \Rightarrow y_2 = 0$; $x_2 = b \Rightarrow y_2 = b$; $x_1 = 0 \Rightarrow y_1 = 0$; $x_1 = a \Rightarrow y_1 y_2 = a$, which is a part of a hyperbola. Hence the integration of $y_2$ in the range $0 \le y_1 \le \frac{a}{b}$ is from 0 to $b$ and the integration in the range $\frac{a}{b} \le y_1 < \infty$ is from 0 to $\frac{a}{y_1}$ and the Jacobian is $y_2$.

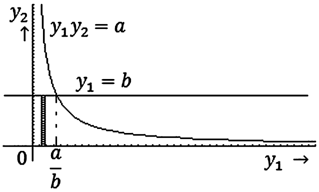


**Figure 7.7:** Region in the $(y_1,y_2)$-plane.

Therefore, the marginal density of $y_1$, denoted by $g_1(y_1)$, is given by

$$g_1(y_1) = \begin{cases} \frac{1}{ab}\int_0^b y_2 \mathrm{d}y_2, & 0 \le y_1 \le \frac{a}{b} \\ \frac{1}{ab}\int_0^{a/y_1} y_2 \mathrm{d}y_2 & \\ 0, & \text{elsewhere.} \end{cases}$$

$$= \begin{cases} \frac{1}{ab}\frac{b^2}{2}, & 0 \le y_1 \le \frac{a}{b} \\ \frac{1}{ab}\frac{a^2}{2y_1^2}, & \frac{a}{b} \le y_1 < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

Let us see whether it is a density. The total integral is given by

$$\begin{aligned} \int_0^\infty f_1(y_1)\mathrm{d}y_1 &= \frac{1}{ab}\left[\int_0^{a/b}\frac{b^2}{2}\mathrm{d}y_1 + \int_{a/b}^\infty \frac{a^2}{2y_1^2}\mathrm{d}y_1\right] \\ &= \frac{1}{ab}\left[\left(\frac{b^2}{2}\right)\left(\frac{a}{b}\right) - \left[\frac{a^2}{2y_1}\right]_{a/b}^\infty\right] \\ &= \frac{ab}{ab} = 1. \end{aligned}$$

Hence the result.

In the example above, we have done three forms, namely the sum, product and ratio. The students are advised to go through the geometry of the transformation from Figures 7.5, 7.6 and 7.7 so that the limits of integration are taken properly. Now there is only one more basic structure left, which is the density of the difference between two random variables. This will be illustrated by taking a simple example of an exponential distribution.

**Example 7.21.** Suppose that $x_1$ and $x_2$ are real scalar positive random variables, independently distributed as exponential with different parameters. Let the marginal densities be

$$f_i(x_1) = \frac{1}{\theta_i}\mathrm{e}^{-\frac{x_i}{\theta_i}}, \quad x_i \ge 0,\ \theta_i > 0,\ i = 1,2$$

and zero elsewhere. Compute the densities of (1) $u = x_1 + x_2$; (2) $v = x_1 - x_2$.

**Solution 7.21.** Transformation of variable technique for a sum is already illustrated in Example 7.19. Now, we shall try to find the density of $y_1$ by using moment generating function. Let the mgf of $x_i$ be denoted by $M_{x_i}(t_i)$, $i = 1,2$. Since the variables are assumed to be independently distributed the mgf of the sum is the product of the mgf's. From straight integration $M_{x_i}(t_i) = (1 - \theta_i t_1)^{-1}$. [This was evaluated for the gamma density already and exponential density is a special case of the gamma density.] Hence the mgf of the sum $x_1 + x_2$ is given by

$$\begin{aligned} E[\mathrm{e}^{t(x_1+x_2)}] &= E[\mathrm{e}^{tx_1}]E[\mathrm{e}^{tx_2}] \\ &= (1-\theta_1 t)^{-1}(1-\theta_2 t)^{-1}. \end{aligned} \tag{7.33}$$

But

$$\frac{1}{(1-\theta_1 t)(1-\theta_2 t)} = \frac{\theta_1}{\theta_1-\theta_2}\frac{1}{1-\theta_1 t} + \frac{\theta_2}{\theta_2-\theta_1}\frac{1}{1-\theta_2 t} \tag{7.34}$$

by using the partial fraction technique, when $\theta_1 \neq \theta_2$. If $\theta_1 = \theta_2 = \theta$, then (7.33) reduces to the mgf of a gamma random variable, and hence $y_1$ has a gamma density with the parameters $(\alpha = 2, \beta = \theta)$. When $\theta_1 \neq \theta_2$, then the sum on the right side in (7.34) can be inverted because each term is a constant multiple of the mgf of an exponential variable. Hence the density of $u$, denoted by $g_1(u)$, is given by

$$g_1(u) = \frac{1}{(\theta_1-\theta_2)}\mathrm{e}^{-\frac{u}{\theta_1}} + \frac{1}{(\theta_2-\theta_1)}\mathrm{e}^{-\frac{u}{\theta_2}},$$

for $u \geq 0$, $\theta_i > 0$, $i = 1, 2$, $\theta_1 \neq \theta_2$ and zero elsewhere. [The student may verify this result by using transformation of variables as done in Example 7.19.]

Now, we shall look at the density of $v = x_1 - x_2$. In the $(x_1, x_2)$-plane the non-zero part of the density is defined in the first quadrant, $\{(x_1, x_2) \mid 0 \leq x_1 < \infty, 0 \leq x_2 < \infty\}$. Let us use transformation of variables. Let $y_1 = x_1 - x_2$, $y_2 = x_2$, the Jacobian is 1 the joint density of $y_1$ and $y_2$, denoted by $g(y_1, y_2)$, is given by

$$g(y_1, y_2) = \frac{1}{\theta_1\theta_2}\mathrm{e}^{-\frac{1}{\theta_1}(y_1+y_2)-\frac{1}{\theta_2}(y_2)}.$$

Now let us look at the region in the $(y_1, y_2)$-plane where the first quadrant in $(x_1, x_2)$-plane is mapped into. $x_2 = 0 \Rightarrow y_2 = 0$; $x_2 \to \infty \Rightarrow y_2 \to \infty$; which is the region above the $y_1$-axis. $x_1 = 0 \Rightarrow y_2 = -y_1$; $x_1 \to \infty \Rightarrow y_1 + y_2 \to \infty$, and hence the region of integration is what is shown in Figure 7.8.

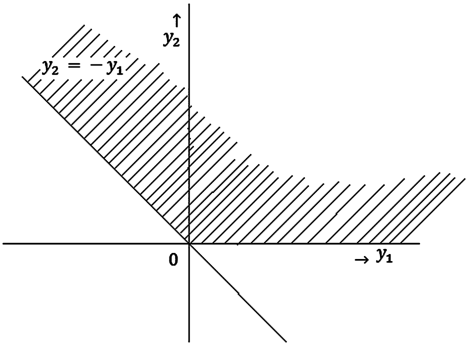


**Figure 7.8:** Region of integration.

Hence when $y_1 > 0$ then $y_2$ to be integrated out from zero to infinity and when $y_1 < 0$ then $y_2$ to be integrated out from $-y_1$ to infinity. If the marginal density of $y_1$ is denoted by $g_1(y_1)$, then

$$
\begin{aligned}
g_1(y_1) &= \frac{1}{\theta_1\theta_2}\mathrm{e}^{-\frac{y_1}{\theta_1}}\int \mathrm{e}^{-(\frac{1}{\theta_1}+\frac{1}{\theta_2})y_2}\mathrm{d}y_2 \\
&= \begin{cases} \frac{\mathrm{e}^{-\frac{y_1}{\theta_1}}}{(\theta_1+\theta_2)}[-\mathrm{e}^{-\frac{(\theta_1+\theta_2)}{(\theta_1\theta_2)}y_2}]_0^{\infty}, & 0 \le y_1 < \infty \\ \frac{\mathrm{e}^{-\frac{y_1}{\theta_1}}}{(\theta_1+\theta_2)}[-\mathrm{e}^{-\frac{\theta_1+\theta_2}{\theta_1\theta_2}y_2}]_{-y_1}^{\infty}, & -\infty < y_1 \le 0 \end{cases} = \begin{cases} \frac{\mathrm{e}^{-\frac{y_1}{\theta_1}}}{(\theta_1+\theta_2)}, & 0 \le y_1 < \infty \\ \frac{\mathrm{e}^{\frac{y_1}{\theta_2}}}{(\theta_1+\theta_2)}, & -\infty < y_1 < 0 \end{cases}
\end{aligned}
\tag{7.35}
$$

and zero elsewhere. It is easily verified that (7.35) is a density.

## Exercises 7.7

**7.7.1.** Use transformation of variable technique to show that the density of $u = x_1 + x_2$ is the same as the one obtained by partial fraction technique in Example 7.20.

**7.7.2.** Verify that (7.35) is a density.

**7.7.3.** If $x_1$ and $x_2$ are independently distributed type-1 beta random variables with different parameters, then evaluate the densities of (1): $u = x_1x_2$; (2): $v = \frac{x_1}{x_2}$.

**7.7.4.** Evaluate the densities of $u$ and $v$ in Exercise 7.7.3 by using the following technique: Take the Mellin transform and then take the inverse Mellin transform to get the result. For example, the Mellin transform of the unknown density $g(u)$ of $u$ is available from $E[u^{s-1}] = E[x_1^{s-1}]E[x_2^{s-1}]$ due to statistical independence and these individual expected values are available from the corresponding type-1 beta densities. Then take the inverse Mellin transform.

**7.7.5.** Let $x_1$ and $x_2$ be independently distributed gamma random variables with the parameters $(\alpha_1,\beta)$ and $(\alpha_2,\beta)$ with the same beta. By using transformation of variables, show that $u = \frac{x_1}{x_1+x_2}$ is type-1 beta distributed, $v = \frac{x_1}{x_2}$ is type-2 beta distributed, $w = x_1 + x_2$ is gamma distributed. [Hint: Use the transformation $x_1 = r\cos^2\theta$, $x_2 = r\sin^2\theta$. Then $J = 2r\cos\theta\sin\theta$.]

**7.7.6.** Prove that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Hint: Consider

$$
\begin{aligned}
\left[\Gamma\left(\frac{1}{2}\right)\right]^2 &= \Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) \\
&= \left[\int_0^{\infty} x^{\frac{1}{2}-1}\mathrm{e}^{-x}\mathrm{d}x\right]\left[\int_0^{\infty} y^{\frac{1}{2}-1}\mathrm{e}^{-y}\mathrm{d}y\right]
\end{aligned}
$$

and make the transformation $x = r\cos^2\theta$, $y = r\sin^2\theta$.

**7.7.7.** Show that

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

for $\Re(\alpha) > 0, \Re(\beta) > 0$. Hint: Start with $\Gamma(\alpha)\Gamma(\beta)$ and use integral representations.

**7.7.8.** Let $u = \frac{x_1}{x_2}$ where $x_1$ and $x_2$ are independently distributed with $x_1 = \chi_m^2/m$ and $x_2 = \chi_n^2/n$. Here, $\chi_\nu^2$ means a chi-square variable with $\nu$ degrees of freedom. Show that $u$ is F-distributed or $u$ has an F-density of the form

$$f(F) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{F^{\frac{m}{2}-1}}{(1+\frac{m}{n}F)^{\frac{m+n}{2}}}$$

for $0 \le F < \infty$, $m, n = 1, 2, \dots$ and zero elsewhere.

**7.7.9.** In Exercise 7.7.8, show that $x = \frac{m}{n}F$ has a type-2 beta distribution with the parameters $\frac{m}{2}$ and $\frac{n}{2}$.

**7.7.10.** Let $u = \frac{x_1}{x_2}$ where $x_1$ and $x_2$ are independently distributed with $x_1 \sim N(0,1)$ and $x_2 = \chi_\nu^2/\nu$. That is, $x_1$ is standard normal and $x_2$ is a chi-square with $\nu$ degrees of freedom, divided by its degrees of freedom. Show that $u$ is Student-t distributed with the density

$$f(u) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-(\frac{\nu+1}{2})}$$

for $-\infty < u < \infty$, $\nu = 1, 2, \dots$.

### A note on degrees of freedom

In general, $k$ “degrees of freedom” means free to vary in $k$ different directions. The phrase “degrees of freedom” appears in different disciplines under different contexts, each having its own interpretation. We will indicate how it is interpreted in statistical literature.

The moment generating function (mgf) of a real gamma variable $x$ with the parameters $(\alpha, \beta)$ is

$$M_x(t) = (1-\beta t)^{-\alpha}.$$

A chi-square variable with $m$ degrees of freedom, $\chi_m^2$, being a real gamma variable with $\beta = 2$, $\alpha = \frac{m}{2}$, has the mgf

$$M_{\chi_m^2}(t) = (1-2t)^{-\frac{m}{2}}.$$

Hence, if $\chi_m^2$ and $\chi_n^2$ are independently distributed then $u = \chi_m^2 + \chi_n^2$ has the mgf

$$M_u(t) = (1-2t)^{-\frac{m}{2}}(1-2t)^{-\frac{n}{2}} = (1-2t)^{-\frac{m+n}{2}}$$

which is the mgf of a chi-square with $m+n$ degrees of freedom. Hence when $\chi^2_m$ and $\chi^2_n$ are independently distributed then

$$\chi^2_m + \chi^2_n \equiv \chi^2_{m+n}.$$

Extending this result, we have

$$\chi^2_m \equiv u_1 + \cdots + u_m$$

where $u_i = \chi^2_1$ or a chi-square with one degree of freedom, where $u_1, \dots, u_m$ are independently distributed. But we have seen that when $x_i \sim N(0,1)$ then $x_i^2 \sim \chi^2_1$, a chi-square with one degree of freedom. Hence

$$\chi^2_m = x_1^2 + \cdots + x_m^2 \tag{7.36}$$

where $x_i^2$ is the square of a standard normal variable and $x_1, \dots, x_m$ are independently distributed. Hence "*m* degrees of freedom" here means that the $\chi^2_m$ can be written as the sum of squares of $m$ independently distributed standard normal variables.

# 8 Some multivariate distributions

## 8.1 Introduction

There are several multivariate (involving more than one random variable) densities, where all the variables are continuous, as well as probability functions where all variables are discrete. There are also mixed cases where some variables are continuous and others are discrete.

## 8.2 Some multivariate discrete distributions

Two such examples, where all the variables are discrete, are the multinomial probability law and the multivariate hypergeometric probability law. These will be considered first.

### 8.2.1 Multinomial probability law

In Bernoulli trials, each trial could result in only one of two events $A_1$ and $A_2$, $A_1 \cup A_2 = S$, $A_1 \cap A_2 = \phi$ where $S$ is the sure event and $\phi$ is the impossible event. We called one of them success and the other failure. We could have also called both "successes" with probabilities $p_1$ and $p_2$ with $p_1 + p_2 = 1$. Now, we look at multinomial trials. Each trial can result in one of $k$ events $A_1, \dots, A_k$ with $A_i \cap A_j = \phi$ for all $i \neq j$, $A_1 \cup \cdots \cup A_k = S$, the sure event. Let the probability of occurrence of $A_i$ be $p_i$. That is, $P(A_i) = p_i$, $i = 1, \dots, k, p_1 + \cdots + p_k = 1$. Suppose that persons in a township are categorized into various age groups, less than or equal to 20 years old (group 1), more than 20 and less than or equal to 30 years old (group 2), 30 to 50 (group 3), over 50 (group 4). If a person is selected at random from this township, then she will belong only to one of these four groups, that is, each trial can result in one of $A_1, A_2, A_3, A_4$ with $A_i \cap A_j = \phi$, $i \neq j$ and $A_1 \cup \cdots \cup A_4 = S =$ sure event. Each such selection is a multinomial trial. If the selection is done independently, then we have independent multinomial trials.

As another example, suppose that the persons are categorized according to their monthly incomes into 10 distinct groups. Then a selected person will belong to one of these 10 groups. Here, $k = 10$ and in the first example $k = 4$.

As another example, consider taking a hand of five cards with replacement. There are four suits of 13 cards each (clubs, diamonds, hearts and spades). If cards are selected at random with replacement, then $p_1 = \frac{13}{52} = \frac{1}{4} = p_2 = p_3 = p_4$.

A general multinomial situation can be described as follows: Each trial results in one of $k$ mutually exclusive and totally exhaustive events $A_1, \dots, A_k$ with the probabilities $p_i = P(A_i)$, $i = 1, \dots, k$, $p_1 + \cdots + p_k = 1$, and we consider $n$ such independent trials. What is the probability that $x_1$ times the event $A_1$ occurs, $x_2$ times the event $A_2$

https://doi.org/10.1515/9783110562545-008

occurs, ..., $x_k$ times the event $A_k$ occurs, so that $x_1 + \cdots + x_k = n$ = the total number of trials. We assume that the probabilities $(p_1, \ldots, p_k)$ remain the same from trial to trial and the trials are independent. Let us denote the joint probability function of $x_1, \ldots, x_k$ by $f(x_1, \ldots, x_k)$. For any given sequence of $x_1$ times $A_1, \ldots, x_k$ times $A_k$, the probability is $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$. Hence the required probability is this probability times the total number of such sequences. Note that $n$ can be permuted among themselves in $n!$ ways, $x_1$ in $x_1!$ ways and so on. Since repetitions are there, the total number of distinct sequences possible is

$$\frac{n!}{x_1! x_2! \cdots x_k!}.$$

Therefore,

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \tag{8.1}$$

for $x_i = 0, 1, \ldots, n$; $0 \le p_i \le 1$, $i = 1, \ldots, k$; $x_1 + x_2 + \cdots + x_k = n$; $p_1 + \cdots + p_k = 1$; and zero otherwise. Since there is a condition $x_1 + \cdots + x_k = n$, one of the variables can be written in terms of others, and hence $f(x_1, \ldots, x_k)$ is a $(k-1)$-variate probability law, not $k$-variate. For example, for $k = 2$ we have

$$\begin{aligned} f(x_1, x_2) &= \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2} \\ &= \frac{n!}{x_1!(n - x_1)!} p_1^{x_1} (1 - p_1)^{n - x_1} \end{aligned}$$

which is the binomial law. Note that the multinomial expansion gives

$$(p_1 + \cdots + p_k)^n = \sum_{x_1 + \cdots + x_k = n} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}. \tag{8.2}$$

What is the joint moment generating function for the multinomial probability law?

$$M(t_1, \ldots, t_k) = E[\mathrm{e}^{t_1 x_1 + \cdots + t_{k-1} x_{k-1}}]$$

since there are only $k - 1$ variables, and it is

$$\begin{aligned} &= \sum_{x_1 + \cdots + x_k = n} \frac{n!}{x_1! \cdots x_k!} (p_1 \mathrm{e}^{t_1})^{x_1} \cdots (p_{k-1} \mathrm{e}^{t_{k-1}})^{x_{k-1}} p_k^{x_k} \\ &= (p_1 \mathrm{e}^{t_1} + \cdots + p_{k-1} \mathrm{e}^{t_{k-1}} + p_k)^n, \end{aligned} \tag{8.3}$$

available from (8.2) by replacing $p_i$ by $p_i \mathrm{e}^{t_i}$, $i = 1, \ldots, k - 1$ and $p_k$ remaining the same. This mgf is differentiable as well as expansible. Hence we should get the integer moments by differentiation.

$$\begin{aligned} E(x_i) &= \frac{\partial}{\partial t_i} M(t_1, \ldots, t_k) \Big|_{t_1 = 0, \ldots, t_{k-1} = 0} \\ &= n p_i \mathrm{e}^{t_i} (p_1 \mathrm{e}^{t_1} + \cdots + p_{k-1} \mathrm{e}^{t_{k-1}} + p_k)^{n-1} |_{t_1 = 0, \ldots, t_{k-1} = 0} \\ &= n p_i (p_1 + \cdots + p_k)^{n-1} = n p_i, \quad i = 1, \ldots, k - 1. \end{aligned}$$

But

$$E(x_k) = E[n - x_1 - \cdots - x_{k-1}] = n - np_1 - \cdots - np_{k-1} = np_k.$$

Hence the formula holds for all $i = 1, \dots, k$ or

$$E(x_i) = np_i, \quad i = 1, \dots, k. \tag{8.4}$$

For $i \neq j$,

$$\begin{aligned} E(x_i x_j) &= \frac{\partial}{\partial t_i}\frac{\partial}{\partial t_j} M(t_1, \dots, t_k)\Big|_{t_1=0,\dots,t_{k-1}=0} \\ &= np_i e^{t_i} \frac{\partial}{\partial t_j}(p_1 e^{t_1} + \cdots + p_{k-1} e^{t_{k-1}} + p_k)^{n-1}\Big|_{t_1=0=\cdots=t_{k-1}} \\ &= np_i(n-1)p_j. \end{aligned}$$

Hence the covariance between $x_i$ and $x_j$ for $i \neq j$,

$$\begin{aligned} \mathrm{Cov}(x_i, x_j) &= E(x_i x_j) - E(x_i)E(x_j) \\ &= n(n-1)p_i p_j - (np_i)(np_j) = -np_i p_j, \quad i \neq j = 1, \dots, k. \end{aligned} \tag{8.5}$$

$$\begin{aligned} E(x_i^2) &= \frac{\partial^2}{\partial t_i^2} M(t_1, \dots, t_k)\Big|_{t_1=0=\cdots=t_{k-1}} \\ &= \frac{\partial}{\partial t_i} np_i e^{t_i}(p_1 e^{t_1} + \cdots + p_{k-1} e^{t_{k-1}} + p_k)^{n-1}\Big|_{t_1=0=\cdots=t_{k-1}} \\ &= n(n-1)p_i^2 + np_i. \end{aligned}$$

Hence the variance of $x_i$ is given by

$$\mathrm{Var}(x_i) = n(n-1)p_i^2 + np_i - (np_i)^2 = np_i(1-p_i), \quad i = 1, 2, \dots, k. \tag{8.6}$$

For $i = 1, \dots, k-1$, they come from differentiation and for $i = k$ by substitution. But $\mathrm{Cov}(x_i, x_j) = -np_i p_j$, $i \neq j = 1, \dots, k$. Hence the covariance matrix for $x_1, \dots, x_{k-1}$ will be non-singular and positive definite but for $x_1, \dots, x_k$ it will be positive semi-definite and singular. The singular covariance matrix, denoted by $\Sigma$, is then given by

$$\begin{aligned} \Sigma = \mathrm{Cov}(X) &= \mathrm{Cov}\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \\ &= \begin{bmatrix} np_1(1-p_1) & -np_1 p_2 & \dots & -np_1 p_k \\ -np_2 p_1 & np_2(1-p_2) & \dots & -np_2 p_k \\ \vdots & \vdots & \dots & np_k(1-p_k) \end{bmatrix}. \end{aligned} \tag{8.7}$$

This matrix $\Sigma$ is a singular matrix of rank $k-1$.

**Example 8.1.** A balanced die is rolled 10 times. What is the probability of getting 5 ones, 3 twos, 2 sixes?

**Solution 8.1.** Since it is told that the die is balanced, we have a multinomial law with $k = 6$, $p_1 = \cdots = p_6 = \frac{1}{6}$. Now, we have a multinomial law with $n = 10$, $x_1 = 5$, $x_2 = 3$, $x_3 = 0 = x_4 = x_5$, $x_6 = 2$. Hence the required probability $p$ is given by

$$\begin{aligned} p &= \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \\ &= \frac{10!}{5!3!0!0!0!2!} \left(\frac{1}{6}\right)^5 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \\ &= 2520\left(\frac{1}{6}\right)^{10} = \frac{70}{6^8} \approx 0.00004. \end{aligned}$$

**Example 8.2.** At Thekkady wild-life reserve, suppose that on any given day the probability of finding a tourist from Kerala is 0.4, from Tamilnadu is 0.3, from other states in India is 0.2 and foreigners is 0.1. On a particular day, there are 20 tourists. What is the probability that out of these 20, 10 are from Tamilnadu and 10 are from other states in India?

**Solution 8.2.** We can take this as a multinomial situation with $k = 4$, $n = 20$, $p_1 = 0.4$, $p_2 = 0.3$, $p_3 = 0.2$, $p_4 = 0.1$, $x_2 = 10$, $x_3 = 10$, $x_1 = 0 = x_4$. The required probability, $p$, is then given by

$$\begin{aligned} p &= \frac{n!}{x_1! \cdots x_4!} p_1^{x_1} \cdots p_4^{x_4} \\ &= \frac{20!}{10!10!0!0!} (0.4)^0 (0.3)^{10} (0.2)^{10} (0.1)^0 \\ &= (11)(13)(17)(19)(4)(0.3)^{10}(0.2)^{10}. \\ &\approx 0.0000001 \end{aligned}$$

### 8.2.2 The multivariate hypergeometric probability law

This law is applicable when sampling is done without replacement. A given trial may result in one of $k$ possible events but the trials are not independent. Suppose that there are $a_1$ objects of one type, $a_2$ objects of a second type etc and $a_k$ objects of the $k$-th type. Suppose that these $a_1 + \cdots + a_k$ objects are well shuffled and a subset of $n$ objects is taken at random. At random, here means that every such subset of $n$ is given and equal chance of being included. This experiment can also be done by picking one at a time at random and without replacement. Both will lead to the same answer. The total number of sample points possible is $\binom{a_1+\cdots+a_k}{n}$. If we obtain $x_1$ of $a_1$ type, $x_2$ of $a_2$ type, etc. and $x_k$ of $a_k$ type so that $x_1 + \cdots + x_k = n$ then the total number of sample points favorable to the event of getting $x_1,\dots,x_k$ is $\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}$. Hence the probability of getting $x_1$ of $a_1$ type,..., $x_k$ of $a_k$ type, denoted by $f(x_1,\dots,x_k)$, is given by

$$f(x_1,\dots,x_k)=\frac{\binom{a_1}{x_1}\cdots\binom{a_k}{x_k}}{\binom{a_1+\cdots+a_k}{n}} \tag{8.8}$$

for $x_i=0,1,\dots,n$ or $a_i$; $x_1+\cdots+x_k=n$; $n=1,2,\dots$; $i=1,\dots,k$ and zero elsewhere.

Note that it is a $(k-1)$-variate probability function because there is one condition that $x_1+\cdots+x_k=n$ so that one variable can be written in terms of others. From Chapter 3, we may note that

$$\sum_{x_1+\cdots+x_k=n} f(x_1,\dots,x_k)=1$$

since

$$\sum_{x_1+\cdots+x_k=n}\binom{a_1}{x_1}\cdots\binom{a_k}{x_k}=\binom{a_1+\cdots+a_k}{x_1+\cdots+x_k}=\binom{a_1+\cdots+a_k}{n}.$$

In this probability function, since factorials are appearing in the denominators, factorial moments can be easily computed.

$$\sum_{x_i=0}^{a_i,n} x_i\binom{a_i}{x_i}=\sum_{x_i=1}^{a_i,n} x_i\binom{a_i}{x_i}=a_i\sum_{x_i=1}^{a_i-1}\binom{a_i-1}{x_i-1}.$$

Therefore,

$$\begin{aligned}E(x_1)&=\sum_{x_1+\cdots+x_k=n}\frac{x_1\binom{a_1}{x_1}\cdots\binom{a_k}{x_k}}{\binom{a_1+\cdots+a_k}{n}}\\&=\frac{a_1\binom{a_1+\cdots+a_k-1}{n-1}}{\binom{a_1+\cdots+a_k}{n}}=\frac{na_1}{a_1+\cdots+a_k}.\end{aligned}$$

Similarly,

$$\begin{aligned}E[x_1(x_1-1)]&=\frac{n(n-1)a_1(a_1-1)}{(a_1+\cdots+a_k)(a_1+\cdots+a_k-1)}\quad\Rightarrow\\ \mathrm{Var}(x_i)&=E[x_i(x_i-1)]+E[x_i]-[E(x_i)]^2\\&=\frac{n(n-1)a_i(a_i-1)}{(a_1+\cdots+a_k)(a_1+\cdots+a_k-1)}+\frac{na_i}{a_1+\cdots+a_k}-\frac{n^2a_i^2}{(a_1+\cdots+a_k)^2}\qquad(8.9)\\ E[x_1x_2]&=\frac{n(n-1)a_1a_2}{(a_1+\cdots+a_k)(a_1+\cdots+a_k-1)}\quad\Rightarrow\\ \mathrm{Cov}(x_i,x_j)&=\frac{n(n-1)a_ia_j}{(a_1+\cdots+a_k)(a_1+\cdots+a_k-1)}-\frac{n^2a_ia_j}{(a_1+\cdots+a_k)^2}.\qquad(8.10)\end{aligned}$$

The joint moment generating function goes into multiple series, and hence we will not discuss here. Also note that the variance does not have a nice representation compared to the covariance expression in (8.10).

**Example 8.3.** From a well-shuffled deck of 52 playing cards, a hand of 8 cards is selected at random. What is the probability that this hand contains 3 clubs, 3 spades and 2 hearts?

**Solution 8.3.** This is a multivariate hypergeometric situation with $k = 4$, $a_1 = 13 = a_2 = a_3 = a_4$, $x_1 = 3$, $x_2 = 3$, $x_3 = 2$, $x_4 = 0$. Hence the required probability is given by

$$\begin{aligned}\frac{\binom{a_1}{x_1}\cdots\binom{a_k}{x_k}}{\binom{a_1+\cdots+a_k}{n}} &= \frac{\binom{13}{3}\binom{13}{3}\binom{13}{2}\binom{13}{0}}{\binom{52}{8}}\\ &= \frac{(1)(2)(3)(4)(5)(6)(7)(8)}{(52)(51)(50)(49)(48)(47)(46)(45)}\left[\frac{(13)(12)(11)}{(1)(2)(3)}\right]^2\frac{(13)(12)}{(1)(2)}\\ &= \frac{(13)(13)(11)(11)(4)}{(47)(23)(5)(15)(17)(7)}.\\ &\approx 0.008478.\end{aligned}$$

## Exercises 8.2

**8.2.1.** In a factory, three machines are producing nuts of a certain diameter. These machines also sometimes produce defective nuts (nuts which do not satisfy quality specifications). Machine 1 is known to produce 40% of the defective nuts, machine 2, 30%, machine 3, 20% and machine 4, 10%. From a day's production, 5 nuts are selected at random and 3 are defective. What is the probability that one defective came from machine 1, and the other 2 from machine 2?

**8.2.2.** Cars on the roads in Kerala are known to be such that 40% are of Indian make, 30% of Indo-Japanese make and 30% others. Out of the 10 cars which came to a toll booth at a particular time, what is the probability that $s$ are Indo-Japanese and 4 are Indian make?

**8.2.3.** A small township has households belonging to the following income groups, based on monthly incomes. (Income group, Number) = (<10 000, 100), (10 000 to 20 000, 50), (over 30 000, 50). Four families are selected from this township, at random. What is the probability that two are in the group (10 000 to 20 000) and two are in the group (<10 000)?

**8.2.4.** A class consists of students in the following age groups: (Age group, Number) = (below 20, 10), (20 to 21, 15), (21 to 22, 20), (above 22, 5). A set of four students is selected ar random. What is the probability that there are one each from each age group?

**8.2.5.** In Exercise 8.2.4, what is the probability that at least one group has none in the selected set?

## 8.3 Some multivariate densities

There are many types of multivariate densities in current use in statistical literature. The most commonly used ones are multivariate normal, Dirichlet type-1, Dirichlet type-2 and multivariate and matrix variate gamma.

Corresponding to a univariate (one variable case) density, do we have something called the unique multivariate density? For example, if $x \sim N(0,1)$, standard normal, and if we have a bivariate density $f(x,y)$ such that $f(x,y) \geq 0$ for all $x$ and $y$, $\int_x \int_y f(x,y)\mathrm{d}x \wedge \mathrm{d}y = 1$, and $\int_x f(x,y)\mathrm{d}x = f_2(y) \sim N(0,1)$, $\int_y f(x,y)\mathrm{d}y = f_1(x) \sim N(0,1)$. Is $f(x,y)$ a unique function or can we have many such $f(x,y)$ satisfying the above conditions having the marginal densities as standard normal? The answer is in the affirmative and we can have many functions satisfying all the above conditions. Hence there is nothing called the unique multivariate analogue of a given univariate density. As two examples, we can give

$$f_1(x,y) = \frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(x^2+y^2)}, \quad f_2(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{1}{2}(x^2-2\rho xy+y^2)},$$

for $-1 < \rho < 1$ are two functions which are both multivariate analogues of a standard normal density. Since there is nothing called a unique multivariate analogue to any given univariate density, some functions are taken as multivariate analogues due to some desirable properties. But the student must keep in mind that when we take a particular multivariate density as an analogue of a given univariate density this function is not taken as the unique multivariate analogue. It is only one multivariate analogue.

### 8.3.1 Type-1 Dirichlet density

This is one generalization of a type-1 beta density. The following is the form of the density:

$$\begin{aligned} f(x_1,\dots,x_k) = {} & c x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_k^{\alpha_k-1} \\ & \times (1-x_1-\cdots-x_k)^{\alpha_{k+1}-1} \end{aligned}$$

for $\Re(\alpha_j) > 0$, $j = 1,\dots,k+1$; $(x_1,\dots,x_k) \in \Omega$, $\Omega = \{(x_1,\dots,x_k) \mid 0 \leq x_i \leq 1, i = 1,\dots,k,$ $x_1 + \cdots + x_k \leq 1\}$, and $f(x_1,\dots,x_k) = 0$ otherwise. In statistical problems, usually the parameters are real and then the conditions will be $\alpha_j > 0$, $j = 1,\dots,k+1$. Note that for $k = 1$ we have type-1 beta density, and hence the above density can be taken as a generalization of a type-1 beta density. By integrating out variables one at a time, we can evaluate the normalizing constant $c$. For example, let

$$\begin{aligned} I_{x_1} = {} & \int_{x_1=0}^{1-x_2-\cdots-x_k} x_1^{\alpha_1-1} \\ & \times [1-x_1-\cdots-x_k]^{\alpha_{k+1}-1}\mathrm{d}x_1 \end{aligned}$$

$$
\begin{aligned}
&= (1-x_2-\cdots-x_k)^{\alpha_{k+1}-1}\int_0^{1-x_2-\cdots-x_k} x_1^{\alpha_1-1} \\
&\quad\times\left[1-\frac{x_1}{1-x_2-\cdots-x_k}\right]^{\alpha_{k+1}-1}\mathrm{d}x_1;
\end{aligned}
$$

Put $y_1 = \frac{x_1}{1-x_2-\cdots-x_k}$ then

$$
\begin{aligned}
I_{x_1} &= (1-x_2-\cdots-x_k)^{\alpha_1+\alpha_{k+1}-1} \\
&\quad\times\int_0^1 y_1^{\alpha_1-1}(1-y_1)^{\alpha_{k+1}-1}\mathrm{d}y_1 \\
&= (1-x_2-\cdots-x_k)^{\alpha_1+\alpha_{k+1}-1} \\
&\quad\times\frac{\Gamma(\alpha_1)\Gamma(\alpha_{k+1})}{\Gamma(\alpha_1+\alpha_{k+1})}
\end{aligned}
$$

for $\Re(\alpha_1) > 0$, $\Re(\alpha_{k+1}) > 0$ or $\alpha_1 > 0$, $\alpha_{k+1} > 0$ if real. Proceeding like this, the final result is the following:

$$
\int_\Omega f(x_1,\dots,x_k)\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_k = cD_k
$$

where

$$
\begin{aligned}
D_k &= D(\alpha_1,\dots,\alpha_k;\alpha_{k+1}) \\
&= \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}.
\end{aligned}
\tag{8.11}
$$

Therefore,

$$
D_k = \frac{1}{c}.
$$

The product moment $E[x_1^{h_1}\cdots x_k^{h_k}]$ is available from (8.11) by replacing $\alpha_j$ by $\alpha_j + h_j$, $j = 1,\dots,k$. That is,

$$
\begin{aligned}
E[x_1^{h_1}\cdots x_k^{h_k}] &= \frac{D(\alpha_1+h_1,\dots,\alpha_k+h_k;\alpha_{k+1})}{D(\alpha_1,\dots,\alpha_k;\alpha_{k+1})} \\
&= \frac{\Gamma(\alpha_1+h_1)}{\Gamma(\alpha_1)}\cdots\frac{\Gamma(\alpha_k+h_k)}{\Gamma(\alpha_k)} \\
&\quad\times\frac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1+h_1+\cdots+\alpha_k+h_k+\alpha_{k+1})}
\end{aligned}
\tag{8.12}
$$

for $\Re(\alpha_j + h_j) > 0$, $j = 1,\dots,k$. This means that if the $\alpha_j$'s are real then the moments will exist for some negative values of $h_j$ also. Some basic properties of type-1 Dirichlet are the following.

**Result 8.1.** *If $(x_1,\dots,x_k)$ has a type-1 Dirichlet distribution, then every subset of $(x_1,\dots,x_k)$ is also type-1 Dirichlet distributed and the individual variables are type-1 beta distributed.*

The proof follows by using the property that arbitrary product moments (8.12) will uniquely determine the corresponding distributions. Retain $h_j$ for the variables in the selected subset and put the remaining $h_j$'s zeros and then identify the corresponding distribution to show that all subsets have the same structure of the density or all subsets are type-1 Dirichlet distributed.

**Result 8.2.** *If $(x_1, \dots, x_k)$ has a type-1 Dirichlet distribution, then $y_1 = 1 - x_1 - \dots - x_k$ and $y_2 = x_1 + \dots + x_k$ are both type-1 beta distributed.*

For proving this, let us consider the $h$-th moment of $1 - x_1 - \dots - x_k$ for an arbitrary $h$.

$$\begin{aligned} E[1 - x_1 - \dots - x_k]^h &= \frac{1}{D_k}\int_\Omega x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} \\ &\quad \times (1 - x_1 - \dots - x_k)^{\alpha_{k+1}+h-1} \mathrm{d}x_1 \wedge \dots \wedge \mathrm{d}x_k \\ &= \frac{\Gamma(\alpha_{k+1} + h)}{\Gamma(\alpha_{k+1})} \\ &\quad \times \frac{\Gamma(\alpha_1 + \dots + \alpha_{k+1})}{\Gamma(\alpha_1 + \dots + \alpha_{k+1} + h)} \end{aligned} \tag{8.13}$$

for $\mathbb{R}(\alpha_{k+1} + h) > 0$. But (8.13) is the $h$-th moment of a type-1 beta random variable with the parameters $(\alpha_{k+1}, \alpha_1 + \dots + \alpha_k)$. Hence $y_1$ is type-1 beta distributed. For any type-1 variable $z$, $1 - z$ is also type-1 beta distributed with the parameters interchanged. Hence $1 - y_1 = y_2$ is type-1 beta distributed with the parameters $(\alpha_1 + \dots + \alpha_k, \alpha_{k+1})$.

**Example 8.4.** If the multinomial probabilities have a prior type-1 Dirichlet distribution, then derive (1) the unconditional distribution of the multinomial variables; (2) the posterior distribution of the multinomial parameters.

**Solution 8.4.** The multinomial probability law for given values of the parameters $p_1, \dots, p_{k-1}$ is given by

$$\begin{aligned} g_1(x_1, \dots, x_{k-1} | p_1, \dots, p_{k-1}) &= \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} \\ &\quad \times (1 - p_1 - \dots - p_{k-1})^{x_k} \end{aligned}$$

for $x_1 + \dots + x_k = n$, and zero otherwise. Let the prior density for $p_1, \dots, p_{k-1}$ be a type-1 Dirichlet. Let

$$\begin{aligned} f_2(p_1, \dots, p_{k-1}) &= \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_{k-1}^{\alpha_{k-1} - 1} \\ &\quad \times (1 - p_1 - \dots - p_{k-1})^{\alpha_k - 1} \end{aligned}$$

for $0 \le p_i \le 1$, $p_1 + \dots + p_{k-1} \le 1$, $\alpha_j > 0$, $j = 1, \dots, k$ and all known, and $f_2(p_1, \dots, p_{k-1}) = 0$ elsewhere. Then the joint probability function of $x_1, \dots, x_{k-1}, p_1, \dots, p_{k-1}$ is given by

$$g_1(x_1, \dots, x_{k-1} | p_1, \dots, p_{k-1}) f_2(p_1, \dots, p_{k-1}).$$

(1) The unconditional probability function of $x_1,\dots,x_{k-1}$, denoted by $f_1(x_1,\dots,x_{k-1})$, is available by integrating out $p_1,\dots,p_{k-1}$.

$$
\begin{aligned}
&f_1(x_1,\dots,x_{k-1})\\
&\quad=\frac{n!}{x_1!\cdots x_k!}\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\int_\Omega p_1^{\alpha_1+x_1-1}\cdots\\
&\qquad\times p_{k-1}^{\alpha_{k-1}+x_{k-1}}(1-p_1-\cdots-p_{k-1})^{n-x_1-\cdots-x_{k-1}+\alpha_k-1}\mathrm{d}p_1\wedge\cdots\wedge\mathrm{d}p_{k-1}\\
&\quad=\frac{n!}{x_1!\cdots x_k!}\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\\
&\qquad\times\frac{\Gamma(\alpha_1+x_1)\cdots\Gamma(\alpha_{k-1}+x_{k-1})\Gamma(n-x_1-\cdots-x_{k-1}+\alpha_k)}{\Gamma(\alpha_1+\cdots+\alpha_k+n)}
\end{aligned}
$$

(2) The posterior density of $p_1,\dots,p_{k-1}$ is available by dividing the joint probability function by $f_1(x_1,\dots,x_{k-1})$. Denoting the posterior density of $p_1,\dots,p_{k-1}$, given $x_1,\dots,x_{k-1}$, by $g_2(p_1,\dots,p_{k-1}|x_1,\dots,x_{k-1})$ we have

$$
\begin{aligned}
&g_2(p_1,\dots,p_{k-1}|x_1,\dots,x_{k-1})\\
&\quad=\frac{\Gamma(\alpha_1+\cdots+\alpha_k+n)}{\Gamma(\alpha_1+x_1)\cdots\Gamma(\alpha_{k-1}+x_{k-1})\Gamma(n-x_1-\cdots-x_{k-1}+\alpha_k)}\\
&\qquad\times p_1^{\alpha_1+x_1-1}\cdots p_{k-1}^{\alpha_{k-1}+x_{k-1}-1}\\
&\qquad\times(1-p_1-\cdots-p_{k-1})^{n-x_1-\cdots-x_{k-1}+\alpha_k-1},
\end{aligned}
$$

for $(p_1,\dots,p_k)\in\Omega$, $\Re(\alpha_j)>0, j=1,\dots,k$, $x_j=0,1,\dots,n, j=1,\dots,k-1$ and $g_2(p_1,\dots,p_{k-1}|x_1,\dots,x_{k-1})=0$ elsewhere. These density functions (1) and (2) are very important in Bayesian analysis and Bayesian statistical inference.

### 8.3.2 Type-2 Dirichlet density

As an extension of type-2 beta density, we have the type-2 Dirichlet density.

$$
f(x_1,\dots,x_k)=\frac{1}{D_k}x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1}(1+x_1+\cdots+x_k)^{-(\alpha_1+\cdots+\alpha_{k+1})}
$$

for $\Re(\alpha_j)>0$, $j=1,\dots,k+1$, $x_j\geq 0$, $j=1,\dots,k$, and $f(x_1,\dots,x_k)=0$ elsewhere. Going through the same steps as in type-1 Dirichlet case, we can show that

$$
\begin{aligned}
&\int_0^\infty\cdots\int_0^\infty x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1}(1+x_1+\cdots+x_k)^{-(\alpha_1+\cdots+\alpha_{k+1})}\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_k\\
&\quad=D(\alpha_1,\dots,\alpha_k;\alpha_{k+1})=\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}
\end{aligned}
\tag{8.14}
$$

This integral is known as type-2 Dirichlet integral.

Arbitrary product moment, $E[x_1^{h_1}\cdots x_k^{h_k}]$, is available from the type-2 Dirichlet integral by replacing $\alpha_j$ by $\alpha_j+h_j$, $j=1,\dots,k$ and $\alpha_{k+1}$ by $\alpha_{k+1}-h_1-\cdots-h_k$. That is,

$$E[x_1^{h_1}\cdots x_k^{h_k}] = \frac{\Gamma(\alpha_1+h_1)}{\Gamma(\alpha_1)}\cdots\frac{\Gamma(\alpha_k+h_k)}{\Gamma(\alpha_k)}\frac{\Gamma(\alpha_{k+1}-h_1-\cdots-h_k)}{\Gamma(\alpha_{k+1})} \tag{8.15}$$

for $\Re(\alpha_j+h_j)>0, j=1,\ldots,k$, $\Re(\alpha_{k+1}-h_1-\cdots-h_k)>0$. This means that the product moment can exist only for some values of $h_1,\ldots,h_k$. Since arbitrary moments uniquely determine the distribution in this case, from (8.14) and (8.15) it is clear that if $x_1,\ldots,x_k$ are jointly type-2 Dirichlet distributed then any subset therein will also be type-2 Dirichlet distributed.

**Result 8.3.** *If $x_1,\ldots,x_k$ are jointly type-2 Dirichlet distributed, then all subsets of $x_1,\ldots,x_k$ are type-2 Dirichlet distributed and individual variables are type-2 beta distributed.*

**Result 8.4.** *If $x_1,\ldots,x_k$ are jointly type-2 Dirichlet distributed, then $y_1=\frac{1}{1+x_1+\cdots+x_k}$ and $y_2=\frac{x_1+\cdots+x_k}{1+x_1+\cdots+x_k}$ are type-1 beta distributed.*

For proving this result, let us take the $h$-th moment of $\frac{1}{1+x_1+\cdots+x_k}$ for arbitrary $h$. That is,

$$\begin{aligned}E\Big[\frac{1}{1+x_1+\cdots+x_k}\Big]^h &= E[1+x_1+\cdots+x_k]^{-h}\\&=[D(\alpha_1,\ldots,\alpha_k;\alpha_{k+1})]^{-1}\int_0^\infty\cdots\int_0^\infty x_1^{\alpha_1-1}\cdots\\&\quad\times x_k^{\alpha_k-1}(1+x_1+\cdots+x_k)^{-(\alpha_1+\cdots+\alpha_{k+1}+h)}\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_k\\&=[D(\alpha_1,\ldots,\alpha_k;\alpha_{k+1})]^{-1}\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1}+h)}{\Gamma(\alpha_1+\cdots+\alpha_{k+1}+h)}\\&=\frac{\Gamma(\alpha_{k+1}+h)}{\Gamma(\alpha_{k+1})}\frac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1+\cdots+\alpha_{k+1}+h)}\end{aligned}$$

which is the $h$-th moment of a type-1 beta random variable with the parameters $(\alpha_{k+1},\alpha_1+\cdots+\alpha_k)$. Hence the result. The second part goes by observing that the second part is $y_2=1-y_1$, and hence the result. Here, the parameters are $(\alpha_1+\cdots+\alpha_k,\alpha_{k+1})$.

There are various generalizations of type-1 and type-2 Dirichlet densities. Two forms which appear in reliability analysis and life-testing models are the following:

$$\begin{aligned}f_1(x_1,\ldots,x_k)&=c_1x_1^{\alpha_1-1}(1-x_1)^{\beta_1}x_2^{\alpha_2-1}\\&\quad\times(1-x_1-x_2)^{\beta_2}\cdots x_k^{\alpha_k-1}\\&\quad\times(1-x_1-\cdots-x_k)^{\alpha_{k+1}+\beta_k-1},\quad(x_1,\ldots,x_k)\in\Omega\end{aligned}\tag{8.16}$$

$$\begin{aligned}f_2(x_1,\ldots,x_k)&=c_2x_1^{\alpha_1-1}x_2^{\alpha_2-1}(x_1+x_2)^{\beta_2}x_3^{\alpha_3-1}\\&\quad\times(x_1+x_2+x_3)^{\beta_3}\cdots x_k^{\alpha_k-1}\\&\quad\times(1-x_1-\cdots-x_k)^{\alpha_{k+1}-1},\quad(x_1,\ldots,x_k)\in\Omega\end{aligned}\tag{8.17}$$

where $c_1$ and $c_2$ are the normalizing constants. For evaluating the normalizing constants in (8.16) and (8.17), start with integrating from $x_k$, $x_{k-1}$ to $x_1$. In (8.17), make the substitution $u_1 = x_1$, $u_2 = x_1 + x_2$ etc. One generalization of type-2 Dirichlet is the following:

$$\begin{aligned} f_3(x_1,\dots,x_k) = c_3 x_1^{\alpha_1-1}(1+x_1)^{-\beta_1}x_2^{\alpha_2-1}\cdots \\ \times (1+x_1+\cdots+x_{k-1})^{-\beta_{k-1}}x_k^{\alpha_k-1} \\ \times (1+x_1+\cdots+x_k)^{-(\alpha_1+\cdots+\alpha_{k+1}+\beta_k)} \end{aligned} \tag{8.18}$$

for $0 \le x_j < \infty$, $i = 1,\dots,k$. For evaluating the normalizing constant $c_3$ start integrating from $x_k$, $x_{k-1},\dots,x_1$. These computations and evaluations of the corresponding marginal densities are left as exercises to the students.

Before concluding this section, let us look into the meaning of largest and smallest random variables.

**Example 8.5.** Let $x_1,x_2,x_3$ be independently distributed exponential random variables with mean values $\lambda_1^{-1},\lambda_2^{-1},\lambda_3^{-1}$ , respectively. Let $y_1 = \max\{x_1,x_2,x_3\}$ and $y_2 = \min\{x_1,x_2,x_3\}$. Evaluate the densities of $y_1$ and $y_2$.

**Solution 8.5.** The student may be confused about the meaning of largest of a set of random variables when $x_1,x_2,x_3$ are all defined on $[0,\infty)$. Let one set of observations on $\{x_1,x_2,x_3\}$ be $\{2,8,5\}$, another set be $\{10,7.2,1\}$, yet another set be $\{2,4.2,8.5\}$. The largest of these observations from each set are $\{8,10,8.5\}$ and the smallest are $\{2,1,2\}$. If $\{8,10,8.5,\dots\}$ are observations on some random variable $y_1$, then $y_1$ is called largest of $x_1,x_2,x_3$ or $y_1 = \max\{x_1,x_2,x_3\}$. Similarly, if $\{2,1,2,\dots\}$ are observations on some random variable $y_2$ then $y_2 = \min\{x_1,x_2,x_3\}$. Let the densities and distribution functions of these be denoted by $f_{y_1}(y_1), f_{y_2}(y_2), F_{y_1}(y_1), F_{y_2}(y_2)$. If the largest $y_1$ is less than a number $u$, then all $x_1,x_2,x_3$ must be less than $u$. Similarly, if the smallest one $y_2$ is greater than $v$, then all must be greater than $v$. But $F_{y_1}(u) = \Pr\{y_1 \le u\}$ and $f_{y_1}(u) = \frac{\mathrm{d}}{\mathrm{d}u}F_{y_1}(u)$. Similarly, $1 - F_{y_2}(v) = \Pr\{y_2 > v\}$. But due to independence,

$$\begin{aligned} F_{y_1}(u) &= \Pr\{x_1 \le u\}\Pr\{x_2 \le u\}\Pr\{x_3 \le u\} \\ &= \prod_{j=1}^{3}\left[\int_0^u \lambda_j \mathrm{e}^{-\lambda_j x_j}\mathrm{d}x_j\right] = \prod_{j=1}^{3}[1-\mathrm{e}^{-\lambda_j u}] \\ f_{y_1}(u) &= \frac{\mathrm{d}}{\mathrm{d}u}F_{y_1}(u) = \sum_{j=1}^{3}\lambda_j \mathrm{e}^{-\lambda_j u} - (\lambda_1+\lambda_2)\mathrm{e}^{-(\lambda_1+\lambda_2)u} \\ &\quad - (\lambda_1+\lambda_3)\mathrm{e}^{-(\lambda_1+\lambda_3)u} - (\lambda_2+\lambda_3)\mathrm{e}^{-(\lambda_2+\lambda_3)u} \\ &\quad + (\lambda_1+\lambda_2+\lambda_3)\mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3)u} \end{aligned}$$

for $0 \le u < \infty$ and zero otherwise. Similarly,

$$
\begin{aligned}
f_{y_2}(v) &= -\frac{\mathrm{d}}{\mathrm{d}v}\Pr\{y_2 > v\} = -\frac{\mathrm{d}}{\mathrm{d}v}\prod_{j=1}^{3}\left(\int_v^{\infty}\lambda_j \mathrm{e}^{-\lambda_j x_j}\mathrm{d}x_j\right) \\
&= -\frac{\mathrm{d}}{\mathrm{d}v}(\mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3)v}) = (\lambda_1+\lambda_2+\lambda_3)\mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3)v}, \quad 0 \le v < \infty
\end{aligned}
$$

and zero elsewhere.

## Exercises 8.3

**8.3.1.** Evaluate the normalizing constant $c_1$ in (8.16). Then evaluate the joint marginal densities of $(x_1,\dots,x_{k-1})$, $(x_1,\dots,x_{k-2})$, ..., $x_1$.

**8.3.2.** For the model in (8.17) evaluate $E[x_1^{h_1}\cdots x_k^{h_k}]$.

**8.3.3.** By using Exercise 8.3.2, or otherwise, show that $x_1$ in the model (8.17) can be written equivalently as a product of independently distributed type-1 beta random variables. (Hint: Take $E(x_1^h)$ and look at the decomposition of this gamma product.)

**8.3.4.** Evaluate the normalizing constant $c_2$ in (8.17).

**8.3.5.** Evaluate the normalizing constant $c_3$ in (8.18).

**8.3.6.** Take the sum $u = x_1 + \cdots + x_k$, the sum of type-1 Dirichlet variables. In Result 8.2, it is shown that $u$ is type-1 beta variable. By using the fact that if $u$ is type-1 beta, then $\frac{u}{1-u}$ and $\frac{1}{1-u}$ are type-2 beta variables write down the results on $(x_1,\dots,x_k)$.

**8.3.7.** It is shown in Result 8.4 that $u = \frac{1}{1+x_1+\cdots+x_k}$ is a type-1 beta if $x_1,\dots,x_k$ have a type-2 Dirichlet distribution. Using the fact that if $u$ is type-1 beta, then $\frac{u}{1-u}$ and $\frac{1}{1-u}$ are type-2 beta distributed, write down the corresponding results for $(x_1,\dots,x_k)$.

**8.3.8.** Using Exercises 8.3.6 and 8.3.7 and by using the properties that if $w$ is type-2 beta, then $\frac{1}{w}$ is type-2 beta, $\frac{1}{1+w}$ is type-1 beta, $\frac{w}{1+w}$ is type-1 beta write down the corresponding results on $(x_1,\dots,x_k)$ when $x_1,\dots,x_k$ have a type-2 Dirichlet distribution.

**8.3.9.** If $(x_1,\dots,x_k)$ is type-1 Dirichlet, then evaluate the conditional density of $x_1$ given $x_2,\dots,x_k$.

**8.3.10.** For $k = 2$, consider type-1 and type-2 Dirichlet densities. By using Maple or Mathematica, draw the 3-dimensional surfaces for (1) fixed $\alpha_1,\alpha_2$ and varying $\alpha_3$; (2) fixed $\alpha_2,\alpha_3$ and varying $\alpha_1$.

## 8.4 Multivariate normal or Gaussian density

As discussed earlier, there is nothing called a unique multivariate analogue of a univariate normal density. But the following form is used as a multivariate analogue due to many parallel characterization results and also due to mathematical convenience.

Let $X' = (x_1, \dots, x_p)$ be the transpose of the column vector with elements $x_1, \dots, x_p$. Consider the following real-valued scalar function ($1 \times 1$ matrix) of $X$, denoted by $f(X)$:

$$f(X) = c\mathrm{e}^{-\frac{1}{2}(X-\tilde{\mu})'\Sigma^{-1}(X-\tilde{\mu})} \tag{8.19}$$

for $-\infty < x_j < \infty$, $-\infty < \mu_j < \infty$, $j = 1, \dots, p$, $\tilde{\mu}' = (\mu_1, \dots, \mu_p)$, $\Sigma = \Sigma' > O$ (positive definite $p \times p$ matrix), where $\tilde{\mu}$ is a parameter vector, $\Sigma > O$ is a parameter matrix. Parallel to the notation used for the scalar case, we will use the following notation.

**Notation 8.1** ($p$-variate Gaussian).

$$X \sim N_p(\tilde{\mu}, \Sigma), \quad \Sigma > O \tag{8.20}$$

meaning that the $p \times 1$ vector $X$ is normal or Gaussian distributed as $p$-variate normal with parameters $\tilde{\mu}$ and $\Sigma$ with $\Sigma > O$ (positive definite).

In order to study properties of $p$-variate Gaussian as well as generalizations of $p$-variate Gaussian and other generalized densities, a few results on Jacobians will be useful. These will be listed here as a note. Those who are familiar with these may skip the note and go to the text.

**Note 8.1** (A note on Jacobians). Before starting the discussion of Jacobians, some basic notations from differential calculus will be recalled here.

**Notation 8.2.** $\wedge$: wedge product.

**Definition 8.1** (Wedge product or skew symmetric product). The wedge product or skew symmetric product of differentials is defined as follows:

$$\mathrm{d}x \wedge \mathrm{d}y = -\mathrm{d}y \wedge \mathrm{d}x \quad \Rightarrow \quad \mathrm{d}x \wedge \mathrm{d}x = 0, \quad \mathrm{d}y \wedge \mathrm{d}y = 0.$$

Now let $y_1 = f_1(x_1, x_2)$, $y_2 = f_2(x_1, x_2)$ be two functions of the real scalar variables $x_1$ and $x_2$. From differential calculus,

$$\mathrm{d}y_1 = \frac{\partial f_1}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f_1}{\partial x_2}\mathrm{d}x_2 \tag{i}$$

$$\mathrm{d}y_2 = \frac{\partial f_2}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f_2}{\partial x_2}\mathrm{d}x_2 \tag{ii}$$

where $\frac{\partial f_i}{\partial x_j}$ denotes the partial derivative of $f_i$ with respect to $x_j$. Then

$$\begin{aligned}\mathrm{d}y_1 \wedge \mathrm{d}y_2 &= \left[\frac{\partial f_1}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f_1}{\partial x_2}\mathrm{d}x_2\right] \wedge \left[\frac{\partial f_2}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f_2}{\partial x_2}\mathrm{d}x_2\right] \\ &= \frac{\partial f_1}{\partial x_1}\frac{\partial f_2}{\partial x_1}\mathrm{d}x_1 \wedge \mathrm{d}x_1 + \frac{\partial f_1}{\partial x_1}\frac{\partial f_2}{\partial x_2}\mathrm{d}x_1 \wedge \mathrm{d}x_2\end{aligned}$$

$$
\begin{aligned}
&+\frac{\partial f_1}{\partial x_2}\frac{\partial f_2}{\partial x_1}\mathrm{d}x_2\wedge\mathrm{d}x_1+\frac{\partial f_1}{\partial x_2}\frac{\partial f_2}{\partial x_2}\mathrm{d}x_2\wedge\mathrm{d}x_2\\
&=\left[\frac{\partial f_1}{\partial x_1}\frac{\partial f_2}{\partial x_2}-\frac{\partial f_1}{\partial x_2}\frac{\partial f_2}{\partial x_1}\right]\mathrm{d}x_1\wedge\mathrm{d}x_2+0+0\\
&=\begin{vmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2}\\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2}\end{vmatrix}\mathrm{d}x_1\wedge\mathrm{d}x_2=J\mathrm{d}x_1\wedge\mathrm{d}x_2
\end{aligned}
$$

where $J$ is the Jacobian of the matrix of partial derivatives $(\frac{\partial f_i}{\partial x_j})$. In general, if $y_j = f_j(x_1,\dots,x_p)$, $j=1,\dots,p$ and the matrix of partial derivatives is $(\frac{\partial f_i}{\partial x_j})$ then the Jacobian is the determinant

$$J=\left|\left(\frac{\partial f_i}{\partial x_j}\right)\right|. \tag{8.21}$$

When $J\neq 0$, then we have

$$
\begin{aligned}
\mathrm{d}y_1\wedge\cdots\wedge\mathrm{d}y_p&=J\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_p.\\
\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_p&=\frac{1}{J}\mathrm{d}y_1\wedge\cdots\wedge\mathrm{d}y_p.
\end{aligned}
\tag{8.22}
$$

As an application of (8.22) we will evaluate a few Jacobians.

**Result 8.5.** *Let $x_1,\dots,x_p$ be distinct real scalar variables and $a_{ij}$'s be constants. Consider the linear forms:*

$$
\begin{aligned}
y_1&=a_{11}x_1+a_{12}x_2+\cdots+a_{1p}x_p\\
y_2&=a_{21}x_1+a_{22}x_2+\cdots+a_{2p}x_p\\
\vdots&=\vdots\\
y_p&=a_{p1}x_1+a_{p2}x_2+\cdots+a_{pp}x_p.
\end{aligned}
$$

*We may write this as $Y=AX$, $Y'=(y_1,\dots,y_p)$, $X'=(x_1,\dots,x_p)$, $A=(a_{ij})$. Then*

$$Y=AX,\quad |A|\neq 0\quad\Rightarrow\quad \mathrm{d}Y=|A|\mathrm{d}X$$

*where $\mathrm{d}Y=\mathrm{d}y_1\wedge\mathrm{d}y_2\wedge\cdots\wedge\mathrm{d}y_p$ and $\mathrm{d}X=\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_p$.*

The proof is trivial. $\frac{\partial y_i}{\partial x_j}=a_{ij}$ and the Jacobian is the determinant or $J=|A|$. When $|A|\neq 0$, then the transformation is one to one.

$$Y=AX,\quad |A|\neq 0\quad\Rightarrow\quad X=A^{-1}Y.$$

We may generalize Result 8.5 for more general linear transformations. Consider a $m\times n$ matrix $X$ of distinct or functionally independent real scalar variables. Here, the wedge product in $X$ will be of the form:

$$\mathrm{d}X=\mathrm{d}x_{11}\wedge\cdots\wedge\mathrm{d}x_{1n}\wedge\mathrm{d}x_{21}\wedge\cdots\wedge\mathrm{d}x_{mn}.$$

Let $A$ be $m\times m$ non-singular matrix of constants. Then we have the following result.

**Result 8.6.** *For the $m \times n$ matrix of distinct real scalar variables and $A$ a $m \times m$ non-singular matrix*

$$Y = AX, \quad |A| \neq 0 \quad \Rightarrow \quad \mathrm{d}Y = |A|^n \mathrm{d}X.$$

**Proof.** Let $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ be the columns of $Y$, $X_{(1)}, \dots, X_{(n)}$ be the columns of $X$. Then

$$Y = AX \quad \Rightarrow \quad (Y_{(1)}, \dots, Y_{(n)}) = (AX_{(1)}, \dots, AX_{(n)}).$$

Now if we look at the string of variables in $Y_{(1)}$, then in $Y_{(2)}, \dots$, in $Y_{(n)}$ and the corresponding variables in $X_{(1)}, \dots, X_{(n)}$. Then the matrix of partial derivatives will be a diagonal block matrix of the form:

$$\begin{bmatrix} A & O & \dots & O \\ O & A & \dots & O \\ \vdots & \vdots & \dots & A \end{bmatrix}$$

where the diagonal blocks are $A$ each and there are $n$ such $A$'s, and hence the determinant is $|A|^n$. This establishes the result.

Now let us see what happens if we post-multiply $X$ with a non-singular $n \times n$ matrix $B$. This will be stated as the next result.

**Result 8.7.** *Let $X$ be a $m \times n$ matrix of distinct real scalar variables and let $B$ be $n \times n$ non-singular matrix of constants. Then*

$$Y = XB, \quad |B| \neq 0 \quad \Rightarrow \quad \mathrm{d}Y = |B|^m \mathrm{d}X.$$

For proving this result, consider the rows of $X$ and $Y$ and proceed as in the case of Result 8.6. Now, combining Results 8.6 and 8.7 we have the following result.

**Result 8.8.** *Let $Y$ and $X$ be $m \times n$ matrices of real distinct variables. Let $A$ be $m \times m$ and $B$ be $n \times n$ non-singular matrices of constants. Then*

$$Y = AXB, \quad |A| \neq 0, \ |B| \neq 0 \quad \Rightarrow \quad \mathrm{d}Y = |A|^n |B|^m \mathrm{d}X.$$

For proving this result, use Results 8.6 and 8.7. Put $Z = BX$ and $Y = AZ$, compute $\mathrm{d}Y$ in terms of $\mathrm{d}Z$ and $\mathrm{d}Z$ in terms of $\mathrm{d}X$ to prove the result.

Now we shall consider a linear transformation on a symmetric matrix. When a $p \times p$ matrix $X = (x_{ij})$ of real scalar variables is symmetric then we have only $p(p+1)/2$ distinct real scalar variables. Then the product of differentials will be of the form:

$$\mathrm{d}X = \mathrm{d}x_{11} \wedge \dots \wedge \mathrm{d}x_{1p} \wedge \mathrm{d}x_{22} \wedge \dots \wedge \mathrm{d}x_{2p} \wedge \dots \wedge \mathrm{d}x_{pp}.$$

**Result 8.9.** *Let $X$, $A$ be $p \times p$, $X = X' = (x_{ij})$ be a matrix of $p(p+1)/2$ distinct real scalar variables and let $A$ be a non-singular matrix of constants. Then*

$$Y = AXA', \quad X = X', \ |A| \neq 0 \quad \Rightarrow \quad \mathrm{d}Y = |A|^{p+1}\mathrm{d}X.$$

We can prove the result by using the fact that any non-singular matrix can be represented as a product of elementary matrices. For the proof of this result as well as those of many other results, the students may see the book [3]. We will list two results on non-linear transformations, without proofs, before closing this note.

**Result 8.10.** *Let the $p \times p$ matrix $X$ be non-singular so that its regular inverse $X^{-1}$ exists. Then*

$$Y = X^{-1} \quad \Rightarrow \quad \mathrm{d}Y = \begin{cases} |X|^{-2p}\mathrm{d}X & \text{for a general } X \\ |X|^{-(p+1)} & \text{for } X = X' \\ |X|^{-(p-1)} & \text{for } X' = -X. \end{cases}$$

**Result 8.11.** *Let the $p \times p$ matrix $X$ be symmetric and let it be positive definite with $\frac{p(p+1)}{2}$ distinct real variables. Let $T = (t_{ij})$ be a lower triangular matrix with positive diagonal elements, $t_{jj} > 0, j = 1, \dots, p$ and $t_{ij}$'s, $i \geq j$ being distinct real variables. Then the transformation $X = TT'$ is one to one and*

$$X = TT', \quad t_{jj} > 0, \ j = 1, \dots, p \quad \Rightarrow \quad \mathrm{d}X = 2^p \left\{ \prod_{j=1}^{p} t_{jj}^{p+1-j} \right\} \mathrm{d}T.$$

With the help of the above Jacobians, a number of results can be established. Some applications to statistical distribution theory will be considered next.

We shall evaluate the normalizing constant in the $p$-variate normal density. Let

$$Y = \Sigma^{-\frac{1}{2}}(X - \tilde{\mu}) \quad \Rightarrow \quad \mathrm{d}Y = |\Sigma|^{-\frac{1}{2}}\mathrm{d}X; \mathrm{d}(X - \tilde{\mu}) = \mathrm{d}X$$

since $\tilde{\mu}$ is a constant vector, where $\Sigma^{-\frac{1}{2}}$ is the positive definite square root of $\Sigma^{-1} > O$. Now, we shall evaluate the normalizing constant $c$. We use the standard notation $\int_X$ which means the integral over $X$. Then

$$\begin{aligned} 1 &= \int_X f(X)\mathrm{d}X = c \int_X \mathrm{e}^{-\frac{1}{2}(X-\tilde{\mu})'\Sigma^{-1}(X-\tilde{\mu})}\mathrm{d}X \\ &= c|\Sigma|^{\frac{1}{2}} \int_Y \mathrm{e}^{-\frac{1}{2}Y'Y}\mathrm{d}Y, \quad Y = \Sigma^{-\frac{1}{2}}(X - \tilde{\mu}) \end{aligned}$$

because, under the substitution $Y = \Sigma^{-\frac{1}{2}}(X - \tilde{\mu})$,

$$\begin{aligned} (X - \tilde{\mu})'\Sigma^{-1}(X - \tilde{\mu}) &= (X - \tilde{\mu})'\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(X - \tilde{\mu}) = Y'Y \\ &= y_1^2 + \cdots + y_p^2; \quad Y' = (y_1, \dots, y_p). \end{aligned}$$

Here, $\int_Y$ means the multiple integral $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}$. But

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z = 2\int_{0}^{\infty} \mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z$$

due to even and convergent functions, and it is

$$\begin{aligned} &= \sqrt{2}\int_0^{\infty} u^{\frac{1}{2}-1}\mathrm{e}^{-u}\mathrm{d}u \quad \left(\text{Put } u = \frac{1}{2}z^2\right) \\ &= \sqrt{2}\Gamma\left(\frac{1}{2}\right) = \sqrt{2\pi} \quad \Rightarrow \\ \int_Y \mathrm{e}^{-\frac{1}{2}Y'Y}\mathrm{d}Y &= (2\pi)^{\frac{p}{2}}. \end{aligned}$$

Hence

$$c = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}.$$

Therefore, the $p$-variate normal density is given by

$$f(X) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}\mathrm{e}^{-\frac{1}{2}(X-\tilde{\mu})'\Sigma^{-1}(X-\tilde{\mu})}, \quad \Sigma > O.$$

What is the mean value of $X$, that is, $E(X)$, and what is the covariance matrix of $X$, that is, $\mathrm{Cov}(X)$? By definition,

$$\begin{aligned} E(X) &= \int_X Xf(X)\mathrm{d}X = \int_X (X - \tilde{\mu} + \tilde{\mu})f(X)\mathrm{d}X \\ &= \tilde{\mu}\int_X f(X)\mathrm{d}X + \int_X (X-\tilde{\mu})f(X)\mathrm{d}X \\ &= \tilde{\mu} + \int_X (X-\tilde{\mu})f(X)\mathrm{d}X \end{aligned}$$

since the total integral is 1. Make the same substitution,

$$Y = \Sigma^{-\frac{1}{2}}(X-\tilde{\mu}) \quad \Rightarrow \quad \mathrm{d}Y = |\Sigma|^{-\frac{1}{2}}\mathrm{d}(X-\tilde{\mu}) = |\Sigma|^{-\frac{1}{2}}\mathrm{d}X.$$

Then

$$\int_X (X-\tilde{\mu})f(X)\mathrm{d}X = \frac{1}{(2\pi)^{p/2}}\int_Y Y\mathrm{e}^{-\frac{1}{2}Y'Y}\mathrm{d}Y$$

where the integrand is an odd function in the elements of $Y$ and each piece is convergent, and hence the integral is zero. Therefore,

$$E(X) = \tilde{\mu}$$

or the first parameter vector is the mean value of $X$ itself. The covariance matrix of $X$, by definition,

$$
\begin{aligned}
\operatorname{Cov}(X) &= E[X - E(X)][X - E(X)]' = E[X - \tilde{\mu}][X - \tilde{\mu}]' \\
&= \Sigma^{\frac{1}{2}} E(YY')\Sigma^{\frac{1}{2}},
\end{aligned}
$$

under the substitution $Y = \Sigma^{-\frac{1}{2}}(X - \tilde{\mu})$

$$
\operatorname{Cov}(X) = \Sigma^{\frac{1}{2}}\left\{\int_Y YY' \mathrm{e}^{-\frac{1}{2}Y'Y}\mathrm{d}Y\right\}\Sigma^{\frac{1}{2}}.
$$

But

$$
YY' = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} [y_1, \dots, y_p] = \begin{bmatrix} y_1^2 & y_1y_2 & \dots & y_1y_p \\ y_2y_1 & y_2^2 & \dots & y_2y_p \\ \vdots & \vdots & \dots & \vdots \\ y_py_1 & y_py_2 & \dots & y_p^2 \end{bmatrix}.
$$

The non-diagonal elements are all odd functions, and hence the integrals over all non-diagonal elements will be zeros. A diagonal element is of the form:

$$
\begin{aligned}
&\frac{1}{(2\pi)^{p/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_j^2 \mathrm{e}^{-\frac{1}{2}(y_1^2+\cdots+y_p^2)} \mathrm{d}y_1 \wedge \cdots \wedge \mathrm{d}y_p \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_j^2 \mathrm{e}^{-\frac{1}{2}y_j^2} \mathrm{d}y_j \left[\prod_{i \neq j=1}^{p} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}y_i^2} \mathrm{d}y_i\right] \\
&= \left[\frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} y_j^2 \mathrm{e}^{-\frac{1}{2}y_j^2} \mathrm{d}y_j\right]\left[\prod_{i \neq j=1}^{p} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}y_i^2} \mathrm{d}y_i\right].
\end{aligned}
$$

But

$$
\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_j^2 \mathrm{e}^{-\frac{1}{2}y_j^2} \mathrm{d}y_j &= \frac{\sqrt{2}}{\sqrt{2\pi}} \int_{0}^{\infty} t^{\frac{1}{2}-1} \mathrm{e}^{-t} \mathrm{d}t \quad \text{under } t = \frac{1}{2}y_j^2 \\
&= \frac{\sqrt{2}}{\sqrt{2\pi}} \Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = 1.
\end{aligned}
$$

Thus each diagonal element gives 1, and thus the integral over $Y$ gives an identity matrix and, therefore,

$$
\operatorname{Cov}(X) = \Sigma
$$

which is the parameter matrix in the density. Thus the two parameters in the density are the mean value and the covariance matrix.

Let us evaluate the moment generating function. Let $T' = (t_1, \dots, t_p)$ be the vector of parameters. Then by definition the moment generating function

$$
\begin{aligned}
M(t_1, \dots, t_p) = M_X(T) &= E[\mathrm{e}^{t_1x_1+\cdots+t_px_p}] = E[\mathrm{e}^{T'X}] \\
&= E[\mathrm{e}^{T'(X-\tilde{\mu}+\tilde{\mu})}] = \mathrm{e}^{T'\tilde{\mu}} E[\mathrm{e}^{T'(X-\tilde{\mu})}]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{e}^{T'\tilde{\mu}} E[\mathrm{e}^{T'\Sigma^{\frac{1}{2}}Y}] \quad \text{for } Y = \Sigma^{-\frac{1}{2}}(X-\tilde{\mu}) \\
&= \mathrm{e}^{T'\tilde{\mu}} \frac{1}{(2\pi)^{\frac{p}{2}}} \int_Y \mathrm{e}^{T'\Sigma^{\frac{1}{2}}Y - \frac{1}{2}Y'Y} \mathrm{d}Y.
\end{aligned}
$$

But

$$
\begin{aligned}
(T'\Sigma^{\frac{1}{2}})Y - \frac{1}{2}Y'Y &= -\frac{1}{2}\{Y'Y - 2T'\Sigma^{\frac{1}{2}}Y\} \\
&= -\frac{1}{2}\{(Y-\Sigma^{\frac{1}{2}}T)'(Y-\Sigma^{\frac{1}{2}}T) - T'\Sigma T\}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
M_X(T) &= \mathrm{e}^{T'\tilde{\mu}+\frac{1}{2}T'\Sigma T} \int_Y \frac{1}{(2\pi)^{\frac{p}{2}}} \mathrm{e}^{-\frac{1}{2}(Y-\Sigma^{\frac{1}{2}}T)'(Y-\Sigma^{\frac{1}{2}}T)} \mathrm{d}Y \\
&= \mathrm{e}^{T'\tilde{\mu}+\frac{1}{2}T'\Sigma T} \qquad (8.23)
\end{aligned}
$$

since the integral is 1. It can be looked upon as the total integral coming from a $p$-variate normal with the parameters $(\Sigma^{\frac{1}{2}}T, I)$. Thus, for example, for $p = 1$, $\Sigma = \sigma_{11} = \sigma_1^2$, $T' = t_1$ and, therefore,

$$
M_X(T) = \exp\left\{t_1\mu_1 + \frac{1}{2}t_1^2\sigma_1^2\right\}. \qquad (8.24)
$$

This equation (8.23) can be taken as the definition for a $p$-variate normal and then in this case $\Sigma$ can be taken as positive semi-definite also because even for positive semi-definite matrix $\Sigma$ the right side in (8.23) will exist. Then in that case when $\Sigma$ is singular or when $|\Sigma| = 0$ we will call the corresponding $p$-variate normal as *singular normal*. For a singular normal case, there is no density because when $\Sigma$ on the right side of (8.23) is singular the inverse Laplace transform does not give a density function, and hence a singular normal has no density but all its properties can be studied by using the mgf in (8.23).

For $p = 2$, $T' = (t_1, t_2)$, $\tilde{\mu}' = (\mu_1, \mu_2)$, $T'\tilde{\mu} = t_1\mu_1 + t_2\mu_2$;

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}; \\
T'\Sigma T &= [t_1, t_2] \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \\
&= \sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2; \\
M_X(T) &= \exp\left\{(t_1\mu_1 + t_2\mu_2) + \frac{1}{2}(t_1^2\sigma_1^2 + 2t_1t_2\rho\sigma_1\sigma_2 + t_2^2\sigma_2^2)\right\} \qquad (8.25)
\end{aligned}
$$

where $\rho$ is the correlation between $x_1$ and $x_2$ in this case.

One important result in this connection is that if $X \sim N_p(\tilde{\mu}, \Sigma)$, then any linear function of $X$, say $u = a_1x_1 + \cdots + a_px_p = a'X = X'a$, $a' = (a_1, \dots, a_p)$, is a univariate normal.

**Result 8.12.** *If $X \sim N_p(\tilde{\mu}, \Sigma)$, then $u = a_1 x_1 + \cdots + a_p x_p = a'X = X'a$ is univariate normal with the parameters $\mu = E(u) = a'\tilde{\mu}$ and $\sigma^2 = \mathrm{Var}(u) = a'\Sigma a$.*

This can be proved by looking at the moment generating function. Since $u$ is a function of $X \sim N_p(\tilde{\mu}, \Sigma)$ we can compute the moment generating function of $u$ from the moment generating function of $X$.

$$\begin{aligned} M_u(t) &= E[\mathrm{e}^{tu}] = E[\mathrm{e}^{(ta')X}] = E[\mathrm{e}^{T'X}], \quad T' = ta' \\ &= \mathrm{e}^{T'\tilde{\mu} + \frac{1}{2}T'\Sigma T} = \mathrm{e}^{t(a'\tilde{\mu}) + \frac{1}{2}t^2(a'\Sigma a)}. \end{aligned}$$

But this is the mgf of a univariate normal with parameters $a'\tilde{\mu}$ and $a'\Sigma a$. Hence $u \sim N_1(a'\tilde{\mu}, a'\Sigma a)$. This shows that every linear function of a multivariate normal vector $X$ is univariate normal. One may also note that we have not used the non-singularity of $\Sigma$ in the proof here. Hence the result holds for singular normal case also.

Naturally, one may ask: if $a'X$ is univariate normal, for a given $a$, will $X$ be a multivariate normal? Obviously, this need not hold. But if $a$ is arbitrary or for all $a$ if $u = a'X$ is univariate normal, will $X$ be multivariate normal? The answer is in the affirmative. This, in fact, also provides a definition for a multivariate normal law as the law satisfied by $X$ when $a'X$ is univariate normal for all constant vector $a$.

**Result 8.13.** *For any vector random variable $X$ and for a constant vector $a$, if $u = a'X = X'a$ is univariate normal for all $a$, then $X$ is multivariate normal $X \sim N_p(\tilde{\mu}, \Sigma)$, $\Sigma \geq 0$.*

The proof is available by retracing the steps in the proof of the previous result. If $u = a'X$ is univariate normal, then its parameters are $E[u] = a'\tilde{\mu}$ and $\mathrm{Var}(u) = a'\Sigma a$. Therefore, the mgf of $u$ is available as

$$\begin{aligned} M_u(t) &= \mathrm{e}^{t(a'\tilde{\mu}) + \frac{t^2}{2}(a'\Sigma a)} \\ &= \mathrm{e}^{T'\tilde{\mu} + \frac{1}{2}T'\Sigma T}, \quad T' = ta' = (ta_1, \dots, ta_p) \end{aligned} \tag{8.26}$$

where $a_1, \dots, a_p$ are arbitrary, and hence $ta_1, \dots, ta_p$ are also arbitrary. There are $p$ arbitrary parameters here in (8.26), and hence it is the mgf of the vector $X$. In other words, $X$ is multivariate normal. Note that the proof holds whether $\Sigma$ is non-singular or singular, and hence the result holds for singular as well as non-singular normal cases.

**Definition 8.2** (Singular normal distribution). Any vector random variable $X$ having the mgf in (8.23), where $\Sigma$ is singular, is called the singular normal vector $X$ and it is denoted as $X \sim N_p(\tilde{\mu}, \Sigma)$, $\Sigma \geq O$.

As mentioned earlier, there is no density for singular normal or when $|\Sigma| = 0$ but all properties can be studied by using (8.26).

Since further properties of a multivariate normal distribution involve a lot of matrix algebra, we will not discuss them here. We will conclude this chapter by looking at a matrix-variate normal.

### 8.4.1 Matrix-variate normal or Gaussian density

Consider a $m\times n$ matrix $X$ of distinct $mn$ real scalar random variables and consider the following non-negative function:

$$f(X)=c\mathrm{e}^{-\frac{1}{2}\operatorname{tr}[A(X-M)B(X-M)']} \tag{8.27}$$

where $X$ and $M$ are $m\times n$, $M$ is a constant matrix, $A$ is $m\times m$ and $B$ is $n\times n$ constant matrices where $A$ and $B$ are positive definite, $X=(x_{ij})$, $M=(m_{ij})$, $-\infty<x_{ij}<\infty$, $-\infty<m_{ij}<\infty$, $i=1,\dots,m$; $j=1,\dots,n$ and $c$ is the normalizing constant.

We can evaluate the normalizing constant by using the Jacobians of linear transformations that we discussed in Note 8.1. Observe that any positive definite matrix can be represented as $CC'$ for some matrix $C$, where $C$ could be rectangular also. Also unique square root is defined when a matrix is positive definite. Let $A^{\frac{1}{2}}$ and $B^{\frac{1}{2}}$ denote the unique square roots of $A$ and $B$, respectively. For the following steps to hold, only a representation in the form $A=A_1A_1'$ and $B=B_1B_1'$, with $A_1$ and $B_1$ being nonsingular, is sufficient but we will use the symmetric positive definite square roots for convenience. Consider the general linear transformation:

$$Y=A^{\frac{1}{2}}(X-M)B^{\frac{1}{2}}\quad\Rightarrow\quad \mathrm{d}Y=|A|^{\frac{n}{2}}|B|^{\frac{m}{2}}\mathrm{d}(X-M)=|A|^{\frac{n}{2}}|B|^{\frac{m}{2}}\mathrm{d}X$$

since $M$ is a constant matrix. Observe that by using the property $\operatorname{tr}(PQ)=\operatorname{tr}(QP)$ when $PQ$ and $QP$ are defined, $PQ$ need not be equal to $QP$, we have

$$\begin{aligned}\operatorname{tr}[A(X-M)B(X-M)']&=\operatorname{tr}[A^{\frac{1}{2}}(X-M)B^{\frac{1}{2}}B^{\frac{1}{2}}(X-M)'A^{\frac{1}{2}}]\\&=\operatorname{tr}(YY')=\sum_{i=1}^{m}\sum_{j=1}^{n}y_{ij}^2,Y=(y_{ij}).\end{aligned}$$

But

$$\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z=\sqrt{2\pi}\quad\Rightarrow\quad\int_Y\mathrm{e}^{-\frac{1}{2}\operatorname{tr}(YY')}\mathrm{d}Y=(2\pi)^{\frac{mn}{2}}.$$

Since the total integral is 1, the normalizing constant

$$c=\frac{|A|^{\frac{n}{2}}|B|^{\frac{m}{2}}}{(2\pi)^{\frac{mn}{2}}}.$$

That is,

$$f(X)=\frac{|A|^{\frac{n}{2}}|B|^{\frac{m}{2}}}{(2\pi)^{\frac{mn}{2}}}\mathrm{e}^{-\frac{1}{2}[A(X-M)B(X-M)']} \tag{8.28}$$

for $A = A' > O, B = B' > O, X = (x_{ij}), M = (m_{ij}), -\infty < x_{ij} < \infty, -\infty < m_{ij} < \infty, i = 1, \dots, m;$ $j = 1, \dots, n$. The density in (8.28) is known as the matrix-variate Gaussian density. Note that when $m = 1$ we have the usual multivariate density or $n$-variate Gaussian density in this case.

## Exercises 8.4

**8.4.1.** If $X \sim N_p(\tilde{\mu}, \Sigma)$, $\Sigma > 0$ and if $X$ is partitioned as $X' = (X'_{(1)}, X'_{(2)})$ where $X_{(1)}$ is $r \times 1$, $r < p$ and if $\Sigma$ is partitioned accordingly as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Sigma_{11} = \mathrm{Cov}(X_{(1)}) > O,$$

$\Sigma_{22} = \mathrm{Cov}(X_{(2)}) > O$, $\Sigma_{12} = \mathrm{Cov}(X_{(1)}, X_{(2)})$, $\Sigma_{12} = \Sigma'_{21}$. Then show that

$$X_{(1)} \sim N_r(\mu_{(1)}, \Sigma_{11}), \quad \Sigma_{11} > O \tag{i}$$

$$X_{(2)} \sim N_{p-r}(\mu_{(2)}, \Sigma_{22}), \quad \Sigma_{22} > O \tag{ii}$$

where $\tilde{\mu}' = (\mu'_{(1)}, \mu'_{(2)})$, $\mu_{(1)}$ is $r \times 1$ and $\mu_{(2)}$ is $(p - r) \times 1$.

**8.4.2.** In Exercise 8.4.1, evaluate the conditional density of $X_{(1)}$ given $X_{(2)}$ and show that the it is also a $r$-variate Gaussian. Evaluate (1) $E[X_{(1)}|X_{(2)}]$, (2) covariance matrix of $X_{(1)}$ given $X_{(2)}$.

**8.4.3.** Answer the questions in Exercise 8.4.2 if $r = 1$, $p - r = p - 1$.

**8.4.4.** Show that when $m = 1$ the matrix-variate Gaussian becomes $n$-variate normal. What are the mean value and covariance matrix in this case?

**8.4.5.** Write the explicit form of a $p$-variate normal density for $p = 2$. Compute (1) the mean value vector; (2) the covariance matrix; (3: correlation $\rho$ between the two components and show that $-1 < \rho < 1$ for the Gaussian to be non-singular and that when $\rho = \pm 1$ the Gaussian is singular.

## 8.5 Matrix-variate gamma density

The integral representation of a scalar variable gamma function is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \mathrm{d}x, \quad \Re(\alpha) > 0.$$

Suppose that we have a $p \times p$ matrix $X = X'$ of $p(p + 1)/2$ distinct real scalar variables. Further, let us assume that $X$ is positive definite. Consider the integral of the form:

$$\Gamma_p(\alpha) = \int_{X>0} |X|^{\alpha - \frac{p+1}{2}} \mathrm{e}^{-\mathrm{tr}(X)} \mathrm{d}X, \quad \Re(\alpha) > \frac{p-1}{2}, \tag{8.29}$$

where $|X|$ means the determinant of $X$, $\mathrm{tr}(X)$ is the trace of $X$, $\int_{X>O}$ means the integration over the positive definite matrix $X$ and $\mathrm{d}X$ stands for the wedge product of the $p(p+1)/2$ differential elements $\mathrm{d}x_{ij}$'s. Observe that (8.29) reduces to the scalar case of the gamma function for $p=1$. Let us try to evaluate the integral in (8.29). Direct integration over the individual variables in $X$ is not possible. Even for a simple case of $p=2$, note that the determinant

$$|X| = \begin{vmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{vmatrix} = x_{11}x_{22} - x_{12}^2, \quad \mathrm{tr}(X) = x_{11} + x_{22}$$

and the integration is over the positive definite matrix $X$, which means a triple integral over $x_{11}, x_{22}, x_{12}$ subject to the conditions $x_{11} > 0$, $x_{22} > 0$, $x_{11}x_{22} - x_{12}^2 > 0$. [Observe that due to symmetry $x_{21} = x_{12}$.] Evaluation of this triple integral, that is, the evaluation of (8.29) for $p=2$ is left as an exercise to the students.

The integral in (8.29) can be evaluated by using Result 8.10 in Note 8.1 for the nonlinear transformation $X = TT'$. Let us make the transformation $X = TT'$ where $T = (t_{ij})$ is a lower triangular matrix with positive diagonal elements, that is, $t_{jj} > 0$, $j = 1, \ldots, p$, $t_{ij} = 0$, for all $i < j$. Under this transformation,

$$\begin{aligned} |X|^{\alpha - \frac{p+1}{2}} &= \prod_{j=1}^{p} (t_{jj}^2)^{\alpha - \frac{p+1}{2}} \\ |X|^{\alpha - \frac{p+1}{2}} \mathrm{d}X &= \left\{ \prod_{j=1}^{p} (t_{jj}^2)^{\alpha - \frac{j}{2}} \right\} \mathrm{d}T \quad \text{and} \\ \Gamma_p(\alpha) &= \int_X |X|^{\alpha - \frac{p+1}{2}} \mathrm{e}^{-\mathrm{tr}(X)} \mathrm{d}X \\ &= \int_T \left\{ \prod_{j=1}^{p} 2 \int_0^\infty (t_{jj}^2)^{\alpha - \frac{j}{2}} \mathrm{e}^{-t_{jj}^2} \mathrm{d}t_{jj} \right\} \\ &\quad \times \prod_{i<j} \left\{ \int_{-\infty}^{\infty} \mathrm{e}^{-t_{ij}^2} \mathrm{d}t_{ij} \right\}. \end{aligned}$$

We need to evaluate only two types of integrals here, one type on $t_{jj}$ and the other type on $t_{ij}$. That is,

$$2 \int_0^\infty (t_{jj}^2)^{\alpha - \frac{j}{2}} \mathrm{e}^{-t_{jj}^2} \mathrm{d}t_{jj} = \int_0^\infty u^{\alpha - \frac{j-1}{2}} \mathrm{e}^{-u} \mathrm{d}u$$

under the substitution $u = t_{jj}^2$, and then the integral

$$= \Gamma\left(\alpha - \frac{j-1}{2}\right), \quad \Re(\alpha) > \frac{j-1}{2}$$

for $j = 1, \ldots, p$, and hence the final condition will be $\Re(\alpha) > \frac{p-1}{2}$ and the gamma product is then

$$\prod_{j=1}^{p} \Gamma\left(\alpha - \frac{j-1}{2}\right) = \Gamma(\alpha)\Gamma\left(\alpha - \frac{1}{2}\right) \cdots \Gamma\left(\alpha - \frac{p-1}{2}\right).$$

In $\prod_{i<j}$ there are $\frac{p(p-1)}{2}$ factors and each factor is the integral

$$\int_{-\infty}^{\infty} \mathrm{e}^{-t_{tij}^2} \mathrm{d}t_{ij} = \sqrt{\pi}$$

and thus this product gives $(\sqrt{\pi})^{\frac{p(p-1)}{2}} = (\pi)^{\frac{p(p-1)}{4}}$. Therefore, the integral reduces to the following:

$$\Gamma_p(\alpha) = \pi^{\frac{p(p-1)}{4}} \Gamma(\alpha)\Gamma\left(\alpha - \frac{1}{2}\right) \cdots \Gamma\left(\alpha - \frac{p-1}{2}\right) \tag{8.30}$$

for $\Re(\alpha) > \frac{p-1}{2}$.

**Notation 8.3.** $\Gamma_p(\alpha)$: Real matrix-variate gamma function.

**Definition 8.3** (The real matrix-variate gamma)**.** The following are the definition of $\Gamma_p(\alpha)$ and its integral representation:

$$\begin{aligned}\Gamma_p(\alpha) &= \pi^{\frac{p(p-1)}{4}} \Gamma(\alpha)\Gamma\left(\alpha - \frac{1}{2}\right) \cdots \Gamma\left(\alpha - \frac{p-1}{2}\right), \quad \Re(\alpha) > \frac{p-1}{2} \\ &= \int_{X>0} |X|^{\alpha - \frac{p+1}{2}} \mathrm{e}^{-\operatorname{tr}(X)} \mathrm{d}X, \quad \Re(\alpha) > \frac{p-1}{2}.\end{aligned}$$

We can define a matrix variate gamma, corresponding to $\Gamma_p(\alpha)$, when the elements in the $p \times p$ matrix $X$ are in the complex domain. This will be called *complex matrix-variate gamma* as opposed to real matrix-variate gamma. This will not be discussed here. For those students, who are interested in or curious to know about random variables on the complex domain, may see the book [3].

For example, for $p = 2$, we can obtain the integral from the formula (8.30). That is,

$$\Gamma_2(\alpha) = \pi^{\frac{1}{2}} \Gamma(\alpha)\Gamma\left(\alpha - \frac{1}{2}\right), \quad \Re(\alpha) > \frac{1}{2}.$$

For $p = 3$,

$$\Gamma_3(\alpha) = \pi^{\frac{3}{2}} \Gamma(\alpha)\Gamma\left(\alpha - \frac{1}{2}\right)\Gamma(\alpha - 1), \quad \Re(\alpha) > 1.$$

By using the integral representation of a real matrix-variate gamma, one can define a real matrix-variate gamma density. Let us consider the following function:

$$f(X) = c|X|^{\alpha - \frac{p+1}{2}} \mathrm{e}^{-\operatorname{tr}(BX)}$$

for $X = X' > O$, $B = B' > O$ where the matrices are $p \times p$ positive definite and $B$ is a constant matrix and $X$ is a matrix of $p(p+1)/2$ distinct real scalar variables. If $f(X)$ is a density, then let us evaluate the normalizing constant $c$. Write $\operatorname{tr}(BX) = \operatorname{tr}(B^{\frac{1}{2}}XB^{\frac{1}{2}})$ by using the property that for two matrices $A$ and $B$, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ as long as $AB$ and $BA$ are defined. Make the transformation

$$Y = B^{\frac{1}{2}}XB^{\frac{1}{2}} \quad \Rightarrow \quad \mathrm{d}X = |B|^{-\frac{p+1}{2}}\mathrm{d}Y$$

by using Result 8.9. Also note that

$$|X|^{\alpha-\frac{p+1}{2}}\mathrm{d}X = |B|^{-\alpha}|Y|^{\alpha-\frac{p+1}{2}}\mathrm{d}Y.$$

Now integrating out, we have

$$\begin{aligned}&\int_{X>O}|X|^{\alpha-\frac{p+1}{2}}\mathrm{e}^{-\operatorname{tr}(BX)}\mathrm{d}X\\ &\quad= |B|^{-\alpha}\int_{Y>O}|Y|^{\alpha-\frac{p+1}{2}}\mathrm{e}^{-\operatorname{tr}(Y)}\mathrm{d}Y\\ &\quad= |B|^{-\alpha}\Gamma_p(\alpha).\end{aligned}$$

Hence the normalizing constant is $|B|^{\alpha}/\Gamma_p(\alpha)$ and, therefore, the density is given by

$$f(X) = \frac{|B|^{\alpha}}{\Gamma_p(\alpha)}|X|^{\alpha-\frac{p+1}{2}}\mathrm{e}^{-\operatorname{tr}(BX)} \tag{8.31}$$

for $X = X' > O, B = B' > O, \Re(\alpha) > \frac{p-1}{2}$ and zero otherwise. This density is known as the real matrix-variate gamma density.

A particular case of this density for $B = \frac{1}{2}\Sigma^{-1}$, $\Sigma > O$ and $\alpha = \frac{n}{2}$, $n = p, p+1, \ldots$ is the most important density in multivariate statistical analysis, known as the Wishart density. By partitioning the matrix $X$, it is not difficult to show that all the leading submatrices in $X$ also have matrix-variate gamma densities when $X$ has a matrix-variate gamma density. This density enjoys many properties, parallel to the properties enjoyed by the gamma density in the scalar case. Also there are other matrix-variate densities such as matrix-variate type 1 and type 2 beta densities and matrix-variate versions of almost all other densities in current use. We will not elaborate on these here. This density is introduced here for the sake of information. Those who are interested to read more on the matrix-variate densities in the real as well as in the complex cases may see the above mentioned book on Jacobians.

## Exercises 8.5

**8.5.1.** Evaluate the integral $\int_{X>O}\mathrm{e}^{-\operatorname{tr}(X)}\mathrm{d}X$ and write down the conditions needed for the convergence of the integral, where the matrix is $p \times p$.

**8.5.2.** Starting with the integral representation of $\Gamma_p(\alpha)$ and then taking the product $\Gamma_p(\alpha)\Gamma_p(\beta)$ and treating it as a double integral, show that

$$\int_{O<X<I}|X|^{\alpha-\frac{p+1}{2}}|I-X|^{\beta-\frac{p+1}{2}}\mathrm{d}X = \frac{\Gamma_p(\alpha)\Gamma(\beta)}{\Gamma_p(\alpha+\beta)}$$

and write down the existence conditions of the integral, where $O < X < I$ means that the $p \times p$ matrix $X$ is positive definite and $I - X$ is also positive definite where $I$ is the identity matrix. [Observe that definiteness is defined only for symmetric matrices when real and Hermitian matrices when in the complex domain.]

**8.5.3.** Evaluate the integral $\int_{O<X<I} \mathrm{d}X$ by using Exercise 8.5.2 or otherwise, where $X$ is $p \times p$ real and verify your result for (1) $p = 1$; (2) $p = 2$; (3) $p = 3$ by direct integration as multiple integrals.

**8.5.4.** Evaluate the integral $\int_{O<X<I} |X|\mathrm{d}X$ where $X$ is $p \times p$, and verify your result for (1): $p = 1$, (2): $p = 2$ by direct integration as multiple integrals.

**8.5.5.** Evaluate the integral $\int_{O<X<I} |I - X|\mathrm{d}X$ where $X$ is $p \times p$, and verify the result by direct integration for (1) $p = 1$; (2) $p = 2$.

# 9 Collection of random variables

## 9.1 Introduction

We had come across one collection of random variables called a simple random sample, where the variables were independently and identically distributed, iid variables. First, we look at some properties of scalar variables, which will be used in the discussions to follow. Hence these will be listed here as results.

**Result 9.1.** *For a real scalar variable $x$, let $E(x) = \mu$,* $\mathrm{Var}(x) = \sigma^2 < \infty$*. Then*

$$\Pr\{|x-\mu| \geq k\sigma\} \leq \frac{1}{k^2}, \quad k > 0. \tag{9.1}$$

This result says that if we are $k\sigma$ away from the mean value $\mu$ then the total probability in the tails is less than or equal to $\frac{1}{k^2}$. This result is also known as *Chebyshev's inequality*. The variables could be discrete or continuous. The probability in the tail is marked in Figure 9.1.

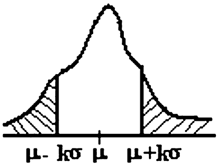


**Figure 9.1:** Probability in the tail: Chebyshev's inequality.

Then the probability in the middle portion is available from (9.1) as one minus the probability in the tails. That is,

$$\Pr\{|x-\mu| \leq k\sigma\} \geq 1 - \frac{1}{k^2}. \tag{9.2}$$

If we replace $k\sigma$ by some $k_1$ then (9.1) and (9.2) can be written in different forms:

$$\begin{aligned} \Pr\{|x-\mu| \geq k\sigma\} &\leq \frac{1}{k^2} \\ \Pr\{|x-\mu| \geq k\} &\leq \frac{\sigma^2}{k^2} \\ \Pr\{|x-\mu| \leq k\sigma\} &\geq 1 - \frac{1}{k^2} \\ \Pr\{|x-\mu| \leq k\} &\geq 1 - \frac{\sigma^2}{k^2}. \end{aligned} \tag{9.3}$$

All these forms in (9.3) are called Chebyshev's inequalities. The proof is quite simple. We will illustrate the proof for the continuous case. For the discrete case, it is parallel. Consider the variance $\sigma^2$.

https://doi.org/10.1515/9783110562545-009

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\mathrm{d}x\\ &= \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)\mathrm{d}x\\ &\quad + \int_{\mu-k\sigma}^{\mu+k\sigma}(x-\mu)^2 f(x)\mathrm{d}x.\end{aligned}$$

Let us delete the middle portion. Then the left side must be bigger than or equal to the balance on the right. That is,

$$\sigma^2 \ge \int_{-\infty}^{\mu-k\sigma}(x-\mu)^2 f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty}(x-\mu)^2 f(x)\mathrm{d}x.$$

In these integrals in the tails $|x-\mu| \ge k\sigma$, and hence if we replace $|x-\mu|$ by its lowest possible point in these two intervals, namely $k\sigma$ then the inequality must remain in the same direction. That is,

$$\begin{aligned}\sigma^2 &\ge \int_{-\infty}^{\mu-k\sigma}(k\sigma)^2 f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty}(k\sigma)^2 f(x)\mathrm{d}x\\ &\ge (k^2\sigma^2)\left[\int_{-\infty}^{\mu-k\sigma} f(x)\mathrm{d}x + \int_{\mu+k\sigma}^{\infty} f(x)\mathrm{d}x\right]\\ &= (k^2\sigma^2)\Pr\{|x-\mu| \ge k\sigma\}.\end{aligned}$$

Dividing by $\sigma^2$ on both sides we have the inequality

$$\Pr\{|x-\mu| \ge k\sigma\} \le \frac{1}{k^2}$$

which holds for all non-degenerate random variables with finite variance. Since we divided by $\sigma^2$ the step is valid only if the variable is non-degenerate. From this result, the above four results in (9.3) are now available.

**Result 9.2** (Chebyshev's inequality). *For real non-degenerate scalar random variable $x$ for which the variance $\sigma^2$ is finite, the inequalities in* (9.3) *hold.*

From the procedure above, it is clear that similar results will hold true if we take any other distance measure. Consider the distance

$$M_r = [E(|x-\mu|^r)]^{\frac{1}{r}} \quad \Rightarrow \quad M_r^r = E(|x-\mu|^r)$$

and let us look at the tail areas after $k$ times $M_r$ from the mean value, that is, $\Pr\{|x-\mu| \ge kM_r\}$. Then proceeding as above, we have

$$\begin{aligned}\Pr\{|x-\mu| \ge kM_r\} &\le \frac{1}{k^r}, \quad r=1,2,\dots, \quad \Rightarrow\\ \Pr\{|x-\mu| \ge k\} &\le 1-\frac{M_r^r}{k^r}, \quad r=1,2,\dots.\end{aligned} \tag{9.4}$$

Hence we have the following result.

**Result 9.3.** *For non-degenerate real scalar random variables for which the $r$-th absolute moment about the mean value exists then the inequality in* (9.4) *and the corresponding four inequalities, parallel to the ones in* (9.3) *hold.*

In the above procedure, we are deleting the middle portion of the probabilities and then replacing the distance measure by the lowest point in the interval. Then if the lowest point always remains positive, we can extend the idea to positive random variables and obtain an inequality in terms of the mean value.

Let $x$ be a real scalar positive random variable with mean value denoted by $\mu$ so that $\mu > 0$. Let $a$ be an arbitrary positive number. Then

$$\begin{aligned}\mu &= \int_0^\infty xf(x)\mathrm{d}x\\ &= \int_0^a xf(x)\mathrm{d}x + \int_a^\infty xf(x)\mathrm{d}x.\end{aligned}$$

Here, all quantities involved are non-negative. Hence if we delete the integral $\int_0^a xf(x)\mathrm{d}x$ then the balance should be less than or equal to $\mu$. If the deleted area is zero, then it will be equal. Then

$$\mu \geq \int_a^\infty xf(x)\mathrm{d}x.$$

But in the interval $[a,\infty)$ the value of $x$ is bigger than or equal to $a$. Hence if we replace $x$ by $a$ then we should get a quantity still less than or equal to the previous quantity. Therefore,

$$\begin{aligned}\mu \geq \int_a^\infty af(x)\mathrm{d}x &= a\int_a^\infty f(x)\mathrm{d}x\\ &= a\Pr\{x \geq a\} \quad \Rightarrow\\ \Pr\{x \geq a\} &\leq \frac{\mu}{a}.\end{aligned}$$

**Result 9.4.** *For non-degenerate positive real scalar random variables for which the mean value $\mu$ exists, then for any positive number $a$,*

$$\Pr\{x \geq a\} \leq \frac{\mu}{a}. \tag{9.5}$$

The inequality in (9.5) is often known as *Markov's inequality*. Thus, if we have iid variables with finite common variance $\sigma^2$ and mean value $\mu$, then we have shown in (7.16) of Chapter 7 that

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}, \quad \bar{x} = \frac{x_1 + \cdots + x_n}{n}, \quad E(\bar{x}) = \mu.$$

Then from Chebyshev's inequality in (9.3), it follows that

$$\begin{aligned}\Pr\{|\bar{x}-\mu| \le k\} &\ge 1 - \frac{\operatorname{Var}(\bar{x})}{k^2} = \frac{\sigma^2}{nk^2} \\ &\to 1 \quad \text{when } n \to \infty. \end{aligned} \tag{9.6}$$

Thus it follows that with probability 1, $\bar{x}$ goes to $\mu$. This is known as the weak law of large numbers.

## 9.2 Laws of large numbers

**Result 9.5** (The weak law of large numbers). *If $x_1, \dots, x_n$ are iid variables with a common finite variance $\sigma^2$ and common mean value $\mu$, and if $\bar{x} = \frac{x_1+\dots+x_n}{n}$ is the sample mean, then the sample mean goes to the population mean value $\mu$ with probability 1 or*

$$\lim_{n\to\infty} \Pr\{|\bar{x}-\mu| \le k\} = 1$$

*for $k > 0$.*

This is the limit of a probability. This can also be looked upon as the stochastic convergence of $\bar{x}$ to its mean value $\mu$ or convergence in the sense of probability. The phrase "weak law" suggests that there is a strong law, which will be a statement on the probability of a limit, which will not be discussed here because it needs more analysis for its proof. A simple illustration of the weak law and its consequence can be seen from Bernoulli trials.

Consider $n$ independent Bernoulli trials. In each trial, the result is either 1 (success) or 0 (failure). If $x_i$ denotes the outcome in the $i$-th trial, then $x_1, \dots, x_n$ are $n$ iid variables taking values 1 with probability $p$ and 0 with probability $q = 1 - p$, or it is a simple random sample of size $n$ from a Bernoulli population. Then sum $x_1 + \dots + x_n$ will give $x$ = the number of successes in $n$ Bernoulli trials, and this $x$ is the binomial random variable. Then $\bar{x} = \frac{x}{n}$ is the proportion of successes in $n$ independent Bernoulli trials. What the weak law of large numbers says is that this proportion of successes converges to the true probability of success $p$ when the number of trials $n$ goes to infinity. This is a very important observation. If we conduct Bernoulli trials, such as getting a head when a coin is tossed repeatedly under the same conditions, and if 40 successes are observed in 100 trials then $\frac{40}{100} = 0.4$ can be taken as an estimate of the true probability. When the number of trials becomes larger and larger, we get a better and better estimate of the true probability of getting a head, and finally when $n$ goes to infinity the sample proportion coincides with the true probability $p$. This is the basis for taking relative frequencies as estimates for the true probability of success in Bernoulli trials.

## 9.3 Central limit theorems

Another interesting result connected with a collection of random variables is a limiting property known as the central limit theorem. We will illustrate it for iid variables. Let $x_1,\dots,x_n$ be a simple random sample of size $n$ from some population with finite variance $\sigma^2$. Let the sample mean be denoted by $\bar{x}=\frac{x_1+\dots+x_n}{n}$. Then $E(\bar{x})=\mu$ the mean value in the population and $\mathrm{Var}(\bar{x})=\frac{\sigma^2}{n}$. Let us look at the standardized sample mean, observing that for any random variable $u$, $v=\frac{u-E(u)}{\sqrt{\mathrm{Var}(u)}}$ is the standardized $u$ so that $E(v)=0$ and $\mathrm{Var}(v)=1$, denoted by $z$.

$$\begin{aligned} z &= \frac{\bar{x}-E(\bar{x})}{\sqrt{\mathrm{Var}(\bar{x})}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \\ &= \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \end{aligned}$$

One central limiting property says that the standardized sample mean, whatever be the population with finite variance, continuous or discrete, will go to standard normal or Gaussian when the sample size $n$ goes to infinity. There are various versions of this limiting property depending upon the conditions that we impose. We will state a central limiting property under the existence of the second moment and then prove it by assuming that the mgf exists.

**Result 9.6** (The central limit theorem). *Consider a simple random sample of size n from a population with finite variance $\sigma^2<\infty$. Let $\bar{x}=\frac{x_1+\dots+x_n}{n}$. Then the standardized sample mean goes to standard normal when $n\to\infty$.*

**Proof.** Let $z$ be the standardized sample mean

$$\begin{aligned} z &= \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} \\ &= \frac{\sqrt{n}}{\sigma}\frac{\sum_{j=1}^{n}(x_i-\mu)}{n} \\ &= \frac{1}{\sigma\sqrt{n}}\left[\sum_{j=1}^{n}(x_j-\mu)\right] \end{aligned}$$

where $\mu$ is the population mean value and $\sigma^2$ is the population variance, which is assumed to be finite. Let $M_{x_i}(t)$ be the mgf of $x_i$. Since the $x_i$'s are identically distributed, we have

$$M_z(t)=\left[M_{x_i-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Taking logarithms and expanding we have

$$\ln M_z(t)=n\ln M_{x_i-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)$$

$$
\begin{aligned}
&= n\ln\left[1+\frac{t}{1!\sigma\sqrt{n}}E(x_i-\mu)+\left(\frac{t}{\sigma\sqrt{n}}\right)^2\frac{E[x_i-\mu]^2}{2!}+\cdots\right]\\
&= n\ln[1+\epsilon],\quad \epsilon=\frac{t^2}{2n}+O\left(\frac{1}{n^{3/2}}\right)\\
&= \frac{t^2}{2}+O\left(\frac{1}{n^{\frac{1}{2}}}\right)\to\frac{t^2}{2}
\end{aligned}
$$

when $n\to\infty$. Then $M_z(t)\to \mathrm{e}^{\frac{t^2}{2}}$ which is the mgf of a standard normal variable. Hence when $n\to\infty$ the standardized sample mean goes to a standard normal variable.

Let us examine the consequences of this result.

(1) If the population is normal, then the standardized sample mean is exactly a standard normal variable for all or for every $n$.

(2) When we have a Bernoulli population and if $x_1,\dots,x_n$ are iid Bernoulli variables then the sample sum $x=x_1+\cdots+x_n$ is the binomial variable because the sample sum gives the number of successes in $n$ independent Bernoulli trials. Then the sample mean $\bar{x}=\frac{x}{n}$ = the binomial proportion with expected value and variance given by

$$
E\left[\frac{x}{n}\right]=p,\quad \mathrm{Var}\left(\frac{x}{n}\right)=\frac{p(1-p)}{n}.
$$

This means that the standardized sample mean

$$
z=\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}=\frac{(\frac{x}{n}-p)}{\sqrt{pq/n}}=\frac{x-np}{\sqrt{npq}}
$$

which is nothing but the standardized binomial variable itself, $q=1-p$. Hence for the binomial variable its standardized form will go to a standard normal when the number of trials $n$ goes to infinity. This is due to the fact that the binomial proportion is nothing but the sample mean when the sample comes from a Bernoulli population.

(3) When the population is gamma with shape parameter $\alpha$ and scale parameter $\beta$, we know that the population mean value is $\alpha\beta$ and the population variance is $\alpha\beta^2$. Hence

$$
z=\frac{(\bar{x}-\alpha\beta)}{\beta\sqrt{\alpha}/\sqrt{n}}=\frac{\sqrt{n}(\bar{x}-\alpha\beta)}{\beta\sqrt{\alpha}}\to N(0,1)
$$

as $n\to\infty$. We had already seen that when the population is gamma then for every $n$, $z$ is a relocated and re-scaled gamma variable and this gamma variable goes to a standard normal, which is an interesting result.

The importance of this limit theorem is that whatever be the population, whether discrete or continuous, the standardized sample mean will go to a standard normal when $n\to\infty$ and when the population variance is finite. Thus the normal or Gaussian distribution becomes a very important distribution in statistical analysis. Misuses come from interpreting this limit theorem in terms of $\bar{x}-\mu$ or $\bar{x}$. This limit theorem does not imply that $\bar{x} - \mu \sim N(0, \frac{\sigma^2}{n})$ for large $n$. It does not imply that $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ for large $n$.

Before concluding this section, some more technical terms will be introduced but a detailed discussion of these will be done in later chapters.

**Definition 9.1** (A statistic). Let $x_1, \ldots, x_n$ be iid variables or a simple random sample of size $n$ from some population. Any observable function of $x_1, \ldots, x_n$ is called a statistic. Several such functions are called statistics, different from the subject of statistics.

For example, $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$ is a statistic. $T_1 = \sum_{j=1}^{n} x_j^2$ is another statistic. $\bar{x}$ and $T_1$ are two statistics. If the function contains some unknown parameters such as $\sum_{i=1}^{n}(x_i - \mu)^2$ it is not a statistic because $\mu$ here is not known. But if $\mu$ is known, such as $\mu = 2$ then $\sum_{i=1}^{n}(x_i - 2)^2$ is a statistic. One important statistic is the sample mean. Another important statistic is the sample variance.

**Definition 9.2** (Sample variance). Consider $x_1, \ldots, x_n$ iid variables. Then

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

is called the sample variance.

**Definition 9.3** (Sampling distributions). The distribution of a statistic is known as a sampling distribution.

If we consider the distributions of $\bar{x}$ and $S^2$, then these are two sampling distributions. Since when $n = 1$ the original population is described, the population distribution itself can be looked upon as a sampling distribution also. The most important sampling distributions in statistical literature are the chi-square distribution, Student-t distribution and the F-distribution. Out of these, chi-square was discussed as a special case of a gamma distribution but it is also associated with a sampling distribution. Discussion of sampling distributions will be postponed to Chapter 10. Before concluding this section, a small property will be examined. When we have a simple random sample from a population with mean value $\mu$ and variance $\sigma^2$ then we have seen that

$$E[\bar{x}] = \mu \quad \text{and} \quad \mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n} = \frac{\text{Population variance}}{\text{Sample size}}$$

What is the expected value of the sample variance? This can be computed by using a standard result. Note that $E[x_i - \mu]^2 = \mathrm{Var}(x_i) = \sigma^2$ for $i = 1, \ldots, n$, and hence $E[\sum_{i=1}^{n}(x_i - \mu)^2] = n\sigma^2$. Consider

$$\begin{aligned}\sum_{i=1}^{n}[x_i-\mu]^2 &= \sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-\mu)]^2\\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2+2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})\\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2 \quad \text{since } \sum_{i=1}^{n}(x_i-\bar{x})=0\end{aligned}$$

Now taking expectations on both sides, we have

$$n\sigma^2 = E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]+n\operatorname{Var}(\bar{x}) \quad \Rightarrow$$

$$E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=(n-1)\sigma^2.$$

In other words, if

$$S_1^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \quad \text{then } E[S_1^2]=\sigma^2$$

and this property is called unbiasedness.

**Definition 9.4** (Unbiasedness). If $T$ is a statistic and if $E[T]=\theta$ for all admissible values of $\theta$, then $T$ is called unbiased for $\theta$ or $T$ is an unbiased estimator of $\theta$.

This is a desirable property in many cases. We have seen that $S_1^2$ is unbiased for $\sigma^2$ but $S^2$ is not unbiased for the population variance $\sigma^2$. Because of this property some people define sample variance as $S_1^2$ instead of $S^2$. But $S_1^2$ should not be taken as sample variance because it is not consistent with the original definition of variance as $\operatorname{Var}(u)=E[u-E(u)]^2$. For example, take a discrete random variable $x$ taking values $x_1,\dots,x_n$ with probabilities $\frac{1}{n}$ each. Then $E[x]=\bar{x}$ and $\operatorname{Var}(x)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$. Further, $S^2$ is the square of per unit distance or dispersion of $x$ from the point of location $\bar{x}$ and consistent with the idea of dispersion or scatter. Thus the proper measure to take for sample variance is $S^2$ and not $S_1^2$. Besides, unbiasedness is not a desirable property in many situations.

### Some general properties on independence

Some general properties on product probability property or statistical independence will be mentioned here.

(a) If the real scalar random variables $x$ and $y$ are independently distributed, then $u=ax+b$, $a\neq 0$ and $v=cy+d$, $c\neq 0$ are also independently distributed.

(b) If the real scalar random variables $x$ and $y$ are independently distributed, then (i) $x^2$ and $y$; (ii) $x$ and $y^2$; (iii) $x^2$ and $y^2$ are independently distributed. Note that when

$x$ and $y$ are independently distributed, that is a property holding in all four quadrants. But (iii) is a property holding in the first quadrant only, (ii) is a property holding in the first and second quadrants only and (i) is a property holding in the first and fourth quadrants only. A property holding in a few quadrants need not hold in all quadrants unless the variables are restricted such as positive variables. If $x$ and $y$ are positive random variables then all properties in (i), (ii), (iii) will hold, otherwise (i) or (ii) or (iii) need not necessarily imply that $x$ and $y$ are independently distributed.

## Exercises 9.3

**9.3.1.** Use a computer and select random numbers between 0 and 1. This is equivalent to taking independent observations from a uniform population over $[0,1]$. For each point, starting from the number of points $n = 5$, calculate the standardized sample mean $z = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$, remembering that for a uniform random variable over $[0,1]$, $\mu = \frac{1}{2}$, $\sigma^2 = \frac{1}{12}$. Make many samples of size 5, form the frequency table of $z$ values and smooth to get the approximate curve. Repeat this for samples of sizes, $n = 5, 6, \dots$ and estimate $n$ so that the simulated curve approximates well with a standard normal curve.

**9.3.2.** Repeat Exercise 9.3.1 if the population is exponential with mean value $\mu = 5$. [Select a random number from the interval $[0,1]$. Convert that into an observation from the exponential population by the probability integral transformation of Section 6.8 in Chapter 6, and then proceed.]

**9.3.3.** Consider the standardized sample mean when the sample comes from a gamma population with the scale parameter $\beta = 1$ and shape parameter $\alpha = 5$. Show that the standardized sample mean is a relocated and re-scaled gamma variable.

**9.3.4.** By using a computer or with the help of MAPLE or MATHEMATICA, compute the upper 5% tail as a function of $n$, the sample size. Determine $n$ when the upper tail has good agreement with the upper 5% tail from a standard normal variable.

**9.3.5.** Repeat the same Exercise 9.3.4 when the population is Bernoulli with the probability of success (1) $p = \frac{1}{2}$, symmetric case; (2) $p = 0.2$ non-symmetric case.

## 9.4 Collection of dependent variables

So far we considered only collections of independent random variables. But practical examples of dependent variables are plenty. General stochastic processes and time series come under this category. Here, we give one example of a sequence of dependent variables.

If the dependent sequence of random variables is considered over time such as monitoring the price of staple food over time, stock market values of shares over time, water level in a dam over time, stock in a grain storage facility over time, etc. then such sequences of random variables over time, are called time series.

There are sequences of variables which are branching in nature. Consider the population size in a banana or pineapple plant. Let us consider one banana plant to start with. This plant produces one bunch of bananas and when that bunch is cut the mother plant dies. But there will be two to three shoots from the bottom. These are the next generation plants. If these shoots are planted, then each will produce new shoots which will be the next generation plants. If the first generation had three shoots and each of these three shoots produced $2, 3, 3$ shoots in the next generation, then the second generation size is $2 + 3 + 3 = 8$. Thus the population size is available from a branching process. Such sequences of random variables are called branching processes.

If we check the number of fish in a particular pool area in a river every morning, then the numbers are likely to be different on different days. By the next morning, some fish may have migrated out of the pool area and some others may have immigrated into the area. If we check the population size in a given community of people every 10 years, then the numbers during successive observation periods are likely to be different. Some may have died out and some new births may have taken place. Such processes are birth and death processes. Special cases are the pure death process and pure birth process.

An ideal hero portrayed in Malayalam movies may be walking home in the following fashion. He comes out of the liquor shop. At every minute, he takes either a step to the left or to the right. That step is followed by a random step at the next minute, and so on. Such processes are called random walk processes.

The above are a few examples of dependent sequences of random variables, generally known as stochastic processes.

# 10 Sampling distributions

## 10.1 Introduction

In Chapter 9, we have already defined a simple random sample from a given population. The population may be designated by a random variable, its probability/density function, its distribution function, its moment generating function (mgf) or its characteristic function. For ready reference, we will list the definition once again. In this chapter, we will deal only with real random variables (not variables defined in the complex domain).

**Definition 10.1** (A simple random sample)**.** Let $x_1, x_2, \dots, x_n$ be a set of independently and identically distributed (iid) random variables; for brevity, we write as iid random variables. Let the common probability/density function be denoted by $f(x)$. Then the collection of random variables $\{x_1, \dots, x_n\}$ is called a simple random sample of size $n$ from the *population* designated by $f(x)$.

**Example 10.1.** If $x_1, \dots, x_n$ are iid random variables following a Poisson distribution with probability function,

$$f_1(x) = \begin{cases} \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, & \lambda > 0,\ x = 0, 1, \dots \\ 0, & \text{elsewhere,} \end{cases} \tag{10.1}$$

then compute the joint probability function of the sample values.

**Solution 10.1.** Here, for example, $\{x_1, \dots, x_4\}$ is a simple random sample of size $n = 4$ from this Poisson *population* with parameter $\lambda$. Then the joint probability function, denoted by $f_1(x_1, \dots, x_n)$, is the product of the marginal probability functions, due to independence. That is,

$$\begin{aligned} f_1(x_1, \dots, x_n) &= \begin{cases} \prod_{j=1}^{n} \frac{\lambda^{x_j}}{x_j!}\mathrm{e}^{-\lambda}, & \lambda > 0,\ x_j = 0, 1, \dots;\ j = 1, \dots, n \\ 0, & \text{elsewhere} \end{cases} \\ &= \begin{cases} \frac{\lambda^{x_1 + \cdots + x_n}}{x_1! \cdots x_n!}, & \lambda > 0, \\ 0, & \text{elsewhere.} \end{cases} \end{aligned} \tag{10.2}$$

**Definition 10.2** (Likelihood function)**.** Let $\{x_1, \dots, x_n\}$ be a collection of random variables with the joint probability/density function $f(x_1, \dots, x_n)$. Then this $f(x_1, \dots, x_n)$ at an observed value of $(x_1, \dots, x_n)$ is called the likelihood function of the random variables $x_1, \dots, x_n$.

If $x_1, \dots, x_n$ are a simple random sample from the population with probability/density function $f(x)$, then the likelihood function, denoted by $L$, is given by the following

https://doi.org/10.1515/9783110562545-010

product due to iid nature:

$$L = \prod_{j=1}^{n} f(x_j) \tag{10.3}$$

when $\{x_1, \dots, x_n\}$ is a set of observed values.

**Note 10.1.** Here, we use the same small letters to denote mathematical variables, random variables and the values assumed by the random variables. The usage will be clear from the context. Many authors denote random variables by capital letters and the values assumed by them by small letters. This can create double notation for the same variable and logical inconsistencies when statements such as $\Pr\{X \leq x\}$ are made, where a big $X$ is smaller than a small $x$. Besides, small letters are used to denote mathematical variables also. Hence we denote mathematical variables as well as random variables by small letters so that degenerate random variables will be interpreted as mathematical variables.

In Example 10.1, suppose that the random variables $x_1, x_2, x_3, x_4$ represent the number of traffic accidents on a stretch of a highway on 4 different independent occasions. Suppose that on the first occasion, possibly the first day of January, $x_1$ is observed as 2, on the second occasion, possibly first day of February, $x_2$ is observed as 0, on the third occasion, possibly the first day of March, $x_3$ is observed as 1 and on the 4th occasion $x_4$ is observed as 5. Then the likelihood function in this case is available from (10.3) by substituting the observations, namely

$$f_1(x_1 = 2, x_2 = 0, x_3 = 1, x_4 = 5) = \frac{e^{-4\lambda}\lambda^{2+0+1+5}}{2!0!1!5!} = \frac{\lambda^8 e^{-4\lambda}}{240}. \tag{10.4}$$

**Example 10.2.** Let $x_1, x_2, x_3$ be independently distributed gamma random variables with parameters $(\alpha_1, \beta)$, $(\alpha_2, \beta)$, $(\alpha_3, \beta)$, respectively. Evaluate the likelihood function.

**Solution 10.2.** Let $L_1$ denote the likelihood function here. Then

$$\begin{aligned} L_1 &= \frac{x_1^{\alpha_1-1} e^{-\frac{x_1}{\beta}}}{\beta^{\alpha_1}\Gamma(\alpha_1)} \times \frac{x^{\alpha_2-1} e^{-\frac{x_2}{\beta}}}{\beta^{\alpha_2}\Gamma(\alpha_2)} \times \frac{x_3^{\alpha_3-1} e^{-\frac{x_3}{\beta}}}{\beta^{\alpha_3}\Gamma(\alpha_3)} \\ &= \frac{x_1^{\alpha_1-1} x_2^{\alpha_2-1} x_3^{\alpha_3-1} e^{-\frac{1}{\beta}(x_1+x_2+x_3)}}{\beta^{\alpha_1+\alpha_2+\alpha_3}\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \end{aligned}$$

at an observed point. Let the observations on the variables be the following: $x_1 = 2$, $x_2 = 1$, $x_3 = 4$. Then

$$L_1 = \frac{2^{\alpha_1-1}(1)^{\alpha_2-1}4^{\alpha_3-1}e^{-\frac{1}{\beta}(2+1+4)}}{\beta^{\alpha_1+\alpha_2+\alpha_3}\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} = \frac{2^{\alpha_1-1}4^{\alpha_3-1}e^{-\frac{7}{\beta}}}{\beta^{\alpha_1+\alpha_2+\alpha_3}\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}. \tag{10.5}$$

**Note 10.2.** Since the joint probability/density function at the observed sample point is defined as the likelihood function, once the point is substituted then the function becomes a function of the parameters only, which may be observed from (10.4) and (10.5), and not a function of the variables $x_1, \dots, x_n$.

**Note 10.3.** In most of the applications in this and succeeding chapters, we will be dealing with simple random samples or iid variables only, coming from a real univariate (scalar variable case) distribution. Hence, hereafter whenever we refer to a sample it will mean a simple random sample or iid variables.

**Definition 10.3** (Sample mean and the sample variance). Let $x_1, \dots, x_n$ be iid variables. Then the sample mean, denoted by $\bar{x}$ = sample mean, and the sample variance, denoted by $s^2$ = sample variance, are defined as follows:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}; \quad s^2 = \sum_{j=1}^{n} \frac{(x_j - \bar{x})^2}{n}. \tag{10.6}$$

Note that when $x_1, \dots, x_n$ are real scalar random variables then $\bar{x}$ and $s^2$ are random variables having their own distributions. If $\{x_1 = a_1, \dots, x_n = a_n\}$ is a given set of observations on $x_1, \dots, x_n$, then an observed value of $\bar{x}$ is $\bar{a} = \frac{a_1 + \cdots + a_n}{n}$ and that of $s^2$ is $\sum_{j=1}^{n} \frac{(a_j - \bar{a})^2}{n}$. For example, if $n = 2$, $a_1 = 1$, $a_2 = 4$ then the observed value of the sample mean is $\frac{4+1}{2} = 2.5$ and the observed value of $s^2$ is $\frac{1}{2}[(1 - \frac{5}{2})^2 + (4 - \frac{5}{2})^2] = 2.25$. In general, $\bar{x}$ and $s^2$ are random variables, and not numbers.

**Definition 10.4** (A statistic). Let the real scalar random variables $x_1, \dots, x_n$ be a sample of size $n$ coming from some population (need not be iid but usually we have iid variables). Any observable function $T = T(x_1, \dots, x_n)$ of the sample values, having its own distribution, is called a statistic. [Plural of the term “statistic” is also called “statistics”, different from the subject matter Statistics. This is yet another unfortunate technical term in Statistics.] For example, the following are statistics:

$$T_1 = x_1 + \cdots + x_n; \quad T_2 = a_1 x_1 + \cdots + a_n x_n,$$

where $a_1, \dots, a_k$ are known constants;

$$T_3 = x_1^2 + \cdots + x_n^2; \quad T_4 = (x_1 - 2)^2 + \cdots + (x_n - 2)^2$$

are statistics, and one can construct many such statistics on a given sample. Students usually have the following doubts: Suppose that we consider functions of the type

$$u_1 = (x_1 - \theta) + \cdots + (x_n - \theta);$$
$$u_2 = \frac{(x_1 - \theta_1)^2 + \cdots + (x_n - \theta_1)^2}{\theta_2^2}$$

where $\theta, \theta_1$ and $\theta_2$ are some unknown parameters. Are $u_1$ and $u_2$ statistics? If the distributions of $u_1$ and $u_2$ are free of $\theta, \theta_1, \theta_2$ will $u_1$ and $u_2$ be statistics? The answer is "no". As long as unknown parameters such as $\theta, \theta_1, \theta_2$ are present, then the functions are not observable, and hence not statistics. If functions of sample values and some unknown parameters are there such that their distributions are free of all parameters then such quantities are called "pivotal" quantities and their uses will be discussed in the chapter on confidence intervals. Hence the most important basic property for a statistic is its observability.

**Definition 10.5** (Sampling distributions). The distribution of a statistic is known as the sampling distribution of that statistic such as the sampling distribution of the sample mean $\bar{x}$, sampling distribution of the sample variance $s^2$, etc.

Observe that the phrase "distribution" is used here in the sense that we have identified a random variable by its probability/density function or its distribution function, etc. It is unfortunate that there are too many similar sounding technical terms which are used in statistical literature, such as "a distribution"(means that a variable is identified such as normal distribution, gamma distribution, etc.), "a distribution function" (means the cumulative probability/density function), "sampling distribution" (means a statistic is identified by its density or probability or distribution function). Also "probability function" is used for discrete and mixed cases only but some authors use it for all cases. Similarly, "density function" is used for continuous cases only but some authors use for all cases. Hence there is no unanimous convention in the use of the terms "probability function" or "density function". In this book, we will use "probability function" for discrete and mixed cases and "density function" for the continuous case.

## 10.2 Sampling distributions

A major part of statistical inference in this module is concerned with Gaussian or normal populations, and hence sampling distributions, when the sample comes from a normal population, are very important here. But we will also consider sampling distributions when the sample comes from other populations as well.

**Example 10.3.** Consider a real gamma population with the density function:

$$f_2(x) = \begin{cases} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha\Gamma(\alpha)}, & x \geq 0,\ \beta > 0,\ \alpha > 0 \\ 0, & \text{elsewhere.} \end{cases} \tag{10.7}$$

Evaluate the density functions of (1) $u_1 = x_1 + \cdots + x_n$; (2) $u_2 = \bar{x} = \frac{(x_1+\cdots+x_n)}{n}$; (3) $u_3 = \bar{x} - \alpha\beta$; (4) $u_4 = \frac{\sqrt{n}(\bar{x}-\alpha\beta)}{\beta\sqrt{\alpha}}$; (5) $u_5 = \lim_{n\to\infty} u_4$.

**Solution 10.3.** It is easier to solve the problems with the help of the mgf of a gamma random variable or with the Laplace transform of a gamma density. The mgf of the random variable $x$ or the density function $f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)$, where $F(x)$ is the distribution function, denoted by $M_x(t)$ is the following:

$$M_x(t) = E[\mathrm{e}^{tx}] = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} \mathrm{e}^{tx}\mathrm{d}F(x)$$

where $t$ is a parameter and $E$ denotes the expected value, is defined by the above integral when the integral exists. [Replace integrals by sums in the discrete case.] Hence the $M_x(t)$ for the gamma density in (10.7) is the following:

$$M_x(t) = \int_0^{\infty} \mathrm{e}^{tx} \frac{x^{\alpha-1}\mathrm{e}^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}\mathrm{d}x = (1-\beta t)^{-\alpha} \quad \text{for } 1-\beta t > 0. \tag{10.8}$$

Observe that the integral is not convergent if $1-\beta t \le 0$:

(1) If $x_1,\dots,x_n$ are iid variables with the mgf $M_x(t)$, then the sum has the mgf $[M_x(t)]^n$ due to independence and identical distribution. Hence

$$M_{u_1}(t) = (1-\beta t)^{-n\alpha}, \quad 1-\beta t > 0 \tag{10.9}$$

where $u_1 = x_1 + \dots + x_n$. Since the mgf is unique, by examining (10.3) we see that $u_1$ is a gamma variable with the parameters $(n\alpha,\beta)$.

(2) Since the mgf of $au_1$ is mgf of $u_1$ with $t$ replaced by $at$, then for $u_2 = \frac{u_1}{n}$ we have

$$M_{u_2}(t) = M_{u_1}\left(\frac{t}{n}\right) = \left(1 - \frac{\beta t}{n}\right)^{-n\alpha}, \quad 1 - \frac{\beta t}{n} > 0. \tag{10.10}$$

This shows that $u_2$ is gamma distributed with the parameters $(n\alpha, \frac{\beta}{n})$ for each $n = 1,2,\dots$. One interesting property is obvious from (10.10). When $n \to \infty$,

$$\lim_{n\to\infty} M_{u_2}(t) = \lim_{n\to\infty}\left(1 - \frac{\beta t}{n}\right)^{-n\alpha} = \mathrm{e}^{\alpha\beta t} \tag{10.11}$$

which is the mgf of a degenerate random variable, taking the value $\alpha\beta$ with probability 1. In other words, as $n$ becomes larger and larger the curve becomes more and more peaked around the line $x = \alpha\beta$, which is the mean value of the gamma variable with parameters $(\alpha,\beta)$, and then eventually the whole probability mass will be concentrated at the point $x = \alpha\beta$. The behavior of the graphs of the density of $\bar{x}$ for various values of the sample size $n$ is shown in Figure 10.1.

**Remark 10.1.** Some students may have the wrong notion, that since the standardized sample mean goes to a standard normal as the sample size goes to infinity, the sample mean itself has an approximate normal distribution for large $n$. This is incorrect, which

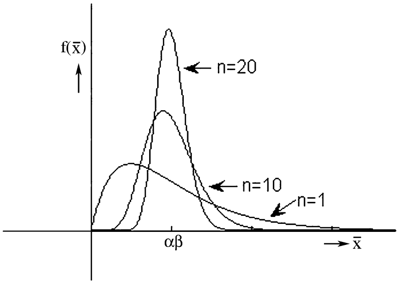


**Figure 10.1:** The density of $\bar{x}$ when the population is gamma.

may be seen from Figure 10.1. Even $\bar{x} - E[\bar{x}]$ does not approximate to a normal, only the standardized sample mean will approximate to a standard normal when $n$ is large. In other words,

$$\frac{\bar{x} - E[\bar{x}]}{\sqrt{\mathrm{Var}(\bar{x})}} \sim N(0,1) \quad \not\Rightarrow \quad \bar{x} \approx N(\mu_0, \sigma_0^2) \quad \text{or} \quad \bar{x} - E(\bar{x}) \approx N(0, \sigma_0^2)$$

for some $\mu_0$ and $\sigma_0^2$.

**Result 10.1.** *When the population is a gamma population with the parameters* $(\alpha, \beta)$, *the sample mean* $\bar{x}$ *goes to* $\alpha\beta$ *with probability* 1 *when the sample size n goes to infinity or* $\bar{x}$ *converges to* $E(x) = \alpha\beta$ *with probability* 1 *when n goes to infinity.*

(3) If a variable $x$ is relocated at the point $x = a$, then the mgf, by definition is the following:

$$M_{x-a}(t) = E[\mathrm{e}^{t(x-a)}] = \mathrm{e}^{-ta} M_x(t).$$

If the variable $x$ is relocated and re-scaled, that is, if $y = ax + b$ then

$$M_y(t) = \mathrm{e}^{tb} M_x(at). \tag{10.12}$$

Therefore,

$$M_{u_3}(t) = \mathrm{e}^{-\alpha\beta t} M_{\bar{x}}(t) = \mathrm{e}^{-\alpha\beta t}\left(1 - \frac{\beta t}{n}\right)^{-n\alpha}, \quad 1 - \frac{\beta t}{n} > 0 \tag{10.13}$$

which shows that $u_3$ is a relocated gamma random variable with parameters $(n\alpha, \frac{\beta}{n})$ and re-location parameter $\alpha\beta$ or with the density, denoted by $f_{u_3}(u_3)$,

$$f_{u_3}(u_3) = \frac{(u_3 - \alpha\beta)^{n\alpha-1}\mathrm{e}^{-\frac{n}{\beta}(u_3-\alpha\beta)}}{(\beta/n)^{n\alpha}\Gamma(n\alpha)} \tag{10.14}$$

for $u_3 \geq \alpha\beta$, $\alpha > 0$, $\beta > 0$, $n = 1, 2, \ldots$

(4)

$$\begin{aligned} u_4 &= \frac{\sqrt{n}}{\beta\sqrt{\alpha}}(\bar{x} - \alpha\beta) = \frac{\sqrt{n}}{\beta\sqrt{\alpha}}\bar{x} - \sqrt{n}\sqrt{\alpha} \\ &= \frac{1}{\beta\sqrt{\alpha}} \frac{\sum_{j=1}^{n} x_j}{\sqrt{n}} - \sqrt{n\alpha}. \end{aligned}$$

Therefore, the mgf of $u_4$ is given by

$$M_{u_4}(t) = \mathrm{e}^{-t\sqrt{n\alpha}}\left(1 - \frac{t}{\sqrt{n\alpha}}\right)^{-n\alpha}, \quad 1 - \frac{t}{\sqrt{n\alpha}} > 0 \tag{10.15}$$

which shows that $u_4$ is a relocated gamma random variable with parameters $(n\alpha, \frac{1}{\sqrt{n\alpha}})$ and the relocation parameter is $\sqrt{n\alpha}$, for each $n = 1, 2, \ldots$.

$u_5$ is the limiting form of $u_4$ when $n$ goes to infinity. A convenient way of taking this limit is to take the limit of the natural logarithm of the right side of (10.15), then expand and then take the limit. That is,

$$\begin{aligned} \ln M_{u_4}(t) &= -t\sqrt{n\alpha} - n\alpha \ln\left(1 - \frac{t}{\sqrt{n\alpha}}\right), \quad \left|\frac{t}{\sqrt{n\alpha}}\right| < 1 \\ &= -t\sqrt{n\alpha} + n\alpha\left[\frac{t}{\sqrt{n\alpha}} + \frac{1}{2}\frac{t^2}{(\sqrt{n\alpha})^2} + \cdots\right] \\ &= \frac{t^2}{2} + \frac{t^3}{3}O\left(\frac{1}{\sqrt{n}}\right) \to \frac{t^2}{2} \quad \text{as } n \to \infty. \end{aligned} \tag{10.16}$$

Since all terms containing $t^3$ and higher powers will contain $\sqrt{n}$ and its powers in the denominator, all terms will go to zero when $n \to \infty$.

$$\lim_{n\to\infty} \ln M_{u_4}(t) = \frac{t^2}{2} \quad \Rightarrow \quad M_{u_5}(t) = \mathrm{e}^{\frac{t^2}{2}} \tag{10.17}$$

which is the mgf of a standard normal variable. Hence $u_5$ has a standard normal distribution with the density

$$f_{u_5}(u_5) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u_5^2}{2}}, \quad -\infty < u_5 < \infty.$$

Since a chi-square random variable is a particular case of a gamma random variable with $\alpha = \frac{\nu}{2}$ and $\beta = 2$, $\nu = 1, 2, \ldots$ ($\nu$ is the Greek letter nu), if $y \sim \chi^2_\nu$ or if $y$ is a chi-square with $\nu$ degrees of freedom then $E(y) = \alpha\beta = (\frac{\nu}{2}) \times 2 = \nu$ and $\mathrm{Var}(y) = \alpha\beta^2 = (\frac{\nu}{2})(4) = 2\nu$. Hence for a sample of size $n$ from a chi-square distribution, with $\nu$ degrees of freedom, the sample sum $u_1 = x_1 + \cdots + x_n$ is a gamma with parameters $n\alpha = \frac{n\nu}{2}$ and $\beta = 2$ or $u_1$ is a chi-square with $n\nu$ degrees of freedom or $u_1 = \chi^2_{n\nu}$. If $u_2 = \bar{x}$, then $u_2$ is a gamma with the parameters $\alpha = \frac{n\nu}{2}$ and $\beta = \frac{2}{n}$. Therefore, we have the following result

**Result 10.2.** *When the population is a chi-square population with ν degrees of freedom, then as $n \to \infty$*

$$\sqrt{n}\frac{(\bar{x} - \nu)}{\sqrt{2\nu}} \to z \sim N(0,1) \quad \textit{as } n \to \infty \tag{10.18}$$

*where $N(0,1)$ denotes a standard normal population.*

As exponential variable is a gamma variable with $\alpha = 1$ and $\beta = \theta > 0$ and if $y_2$ denotes an exponential variable with parameter $\theta$, then $E(y_2) = \theta$ and $\mathrm{Var}(y_2) = \theta^2$. If a sample of size $n$ comes from an exponential population with parameter $\theta$, then we have the following result.

**Result 10.3.** *For a sample of size $n$ from an exponential population with parameter $\theta$,*

$$\frac{\sqrt{n}(\bar{x} - \theta)}{\theta} \to z \sim N(0,1), \quad \textit{as } n \to \infty. \tag{10.19}$$

Note that in this case the sample sum $u_1 = x_1 + \cdots + x_n$ is a gamma random variable with parameters $\alpha = n$ and $\beta = \theta$. The sample mean $u_2 = \bar{x}$ is a gamma with parameters $\alpha = n$ and $\beta = \frac{\theta}{n}$ for each $n = 1, 2, \ldots$. The above are the illustrations of the central limit theorem when the population is a gamma.

**Example 10.4.** Consider a simple random sample of size $n$, $\{x_1, \ldots, x_n\}$, from a Bernoulli population with probability function

$$f_2(x) = p^x q^{1-x}, \quad x = 0, 1,\ q = 1 - p,\ 0 < p < 1$$

and zero elsewhere. [Note that for $p = 0$ or $p = 1$ we have a deterministic situation or a degenerate random variable.] Evaluate the probability functions of (1) $u_1 = x_1 + \cdots + x_n$; (2) $u_2 = \bar{x} = \frac{x_1 + \cdots + x_n}{n}$; (3) $u_3 = \bar{x} - p$; (4) $u_4 = \sqrt{n}\frac{(\bar{x} - p)}{\sqrt{pq}}$; (5) $u_5 = \lim_{n\to\infty} u_4$.

**Solution 10.4.** Note that the mean value and the variance of a Bernoulli variable $x$ are given by the following: $E(x) = p$, $\mathrm{Var}(x) = pq$ and $\mathrm{Var}(\bar{x}) = \frac{pq}{n}$. The mgf of the Bernoulli variable $x$ is given by

$$M_x(t) = \sum_{x=0}^{1} \mathrm{e}^{tx} p^x q^{1-x} = q + p\mathrm{e}^t, \quad q = 1 - p,\ 0 < p < 1.$$

(1) Therefore, the mgf of $u_1 = x_1 + \cdots + x_n$ is available as

$$M_{u_1}(t) = (q + p\mathrm{e}^t)^n. \tag{10.20}$$

But this is the mgf of a binomial random variable, and hence $u_1$ is a binomial random variable, with the probability function,

$$f_{u_1}(u_1) = \binom{n}{u_1} p^{u_1} q^{n-u_1}, \quad u_1 = 0, 1, \dots, n;\ q = 1 - p,\ 0 < p < 1$$

and zero otherwise.

(2) By using the earlier procedure,

$$M_{u_2}(t) = (q + p\mathrm{e}^{\frac{t}{n}})^n.$$

The probability function in this case is

$$f_{u_2}(u_2) = \binom{n}{nu_2} p^{nu_2} q^{n-nu_2}, \quad u_2 = 0, \frac{1}{n}, \dots, 1$$

and zero elsewhere.

(3) By using the earlier procedure

$$M_{u_3}(t) = \mathrm{e}^{-pt} M_{u_2}(t) = \mathrm{e}^{-pt}(q + p\mathrm{e}^{\frac{t}{n}})^n.$$

This gives a relocated form of the probability function in Case (2).

$$\Pr\{u_1 = y\} = \Pr\left\{u_2 = \frac{y}{n}\right\} = \Pr\left\{u_3 = \frac{y}{n} - p\right\}$$

for $y = 0, 1, \dots, n$.

(4)

$$u_4 = \frac{\sqrt{n}}{\sqrt{pq}}\bar{x} - \frac{p\sqrt{n}}{\sqrt{pq}} = \frac{\sum_{j=1}^n x_j}{\sqrt{npq}} - \frac{\sqrt{np}}{\sqrt{q}}.$$

Then the mgf is given by

$$M_{u_4}(t) = \mathrm{e}^{-t\frac{\sqrt{np}}{\sqrt{q}}}(q + p\mathrm{e}^{\frac{t}{\sqrt{npq}}})^n.$$

(5) Consider the natural logarithm on both sides, then expand the exponential function:

$$\begin{aligned}\ln M_{u_4}(t) &= -\frac{\sqrt{np}}{\sqrt{q}}t + n\ln\left[q + p\left(1 + \frac{t}{\sqrt{npq}} + \frac{t^2}{2!(npq)} + O\left(\frac{1}{n^{3/2}}\right)\right)\right] \\ &= -\frac{\sqrt{np}}{\sqrt{q}}t + n\ln[1 + \epsilon]\end{aligned}$$

where $q + p = 1$ and $\epsilon = \frac{pt}{\sqrt{npq}} + \frac{pt^2}{2!(npq)} + O(\frac{1}{n^{3/2}})$. But

$$\ln(1 + \epsilon) = \epsilon - \frac{\epsilon^2}{2} + \cdots \quad \text{for } |\epsilon| < 1.$$

Without loss of generality, we can assume that $|\epsilon| < 1$ for large $n$. Therefore,

$$\ln M_{u_4}(t) = -\frac{\sqrt{np}}{\sqrt{q}}t + n\epsilon - n\frac{\epsilon^2}{2} + \cdots$$

Now, collecting the coefficients of $t$ on the right we see that it is zero. The coefficient of $t^2$ on the right is $\frac{1}{2}$ and the remaining terms are of the order $O(\frac{1}{n^{\frac{1}{2}}}) \to 0$ as $n \to \infty$. Hence, when $n \to \infty$, we have

$$\ln M_{u_5}(t) = \lim_{n\to\infty} \ln M_{u_4}(t) = \frac{t^2}{2} \quad \Rightarrow \quad M_{u_5}(t) = \mathrm{e}^{\frac{t^2}{2}}$$

which is the mgf of a standard normal variable. Hence we have the following result:

**Result 10.4.** *When the sample of size n comes from a Bernoulli population $p^x q^{1-x}$, $x = 0, 1$, $q = 1-p$, $0 < p < 1$ then the standardized sample mean, which is equivalent to the standardized binomial variable, goes to a standard normal variable when n goes to infinity. That is,*

$$u_5 = \frac{\bar{x} - E(\bar{x})}{\sqrt{\mathrm{Var}(\bar{x})}} = \frac{\sum_{j=1}^{n} x_j - np}{\sqrt{npq}} = \frac{x - np}{\sqrt{npq}} \to z \sim N(0,1)$$

*as $n \to \infty$ where x is the binomial random variable.*

Thus, in the binomial case the standardized variable itself goes to the standard normal variable when the number of Bernoulli trials goes to infinity. This result is also consistent with the central limit theorem where the population is the Bernoulli population.

**Note 10.4.** If $x$ is a real scalar random variable with $E(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$, then $y = \frac{x-\mu}{\sigma}$ is called the standardized $x$, with $E(y) = 0$ and $\mathrm{Var}(y) = 1$.

**Note 10.5.** When a simple random sample of size $n$ comes from a Bernoulli population, then the likelihood function $L$ is given by

$$L = \prod_{j=1}^{n} p^{x_j} q^{1-x_j} = p^x q^{n-x} \tag{10.21}$$

where $x$ is a binomial random variable at the observed sample point. Observe that the number of combinations $\binom{n}{x}$, appearing in the binomial probability function, does not enter into the likelihood function in (10.21).

**Example 10.5.** Let $x_1, \dots, x_n$ be iid real scalar random variables following a normal distribution $N(\mu, \sigma^2)$. Compute the distributions of (1) $u_1 = a_1x_1 + \cdots + a_nx_n$ where $a_1, \dots, a_n$ are constants; (2) $u_2 = \bar{x}$; (3) $u_3 = \bar{x} - \mu$; (4) $u_4 = \frac{\sqrt{n}}{\sigma}(\bar{x} - \mu)$.

**Solution 10.5.** The Gaussian or normal density function for a real scalar random variable $x$ is given by

$$f_2(x) = \frac{\mathrm{e}^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}}, \quad -\infty < x < \infty,\ \sigma > 0,\ -\infty < \mu < \infty$$

and the mgf of $x$ is given by

$$M_x(t) = E[\mathrm{e}^{tx}] = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_2(x)\mathrm{d}x = \mathrm{e}^{t\mu + \frac{1}{2}t^2\sigma^2}. \tag{10.22}$$

(1) In order to compute the distribution of $u_1$, we will compute the mgf of $u_1$ and then try to identify the distribution from this mgf:

$$\begin{aligned} M_{u_1}(t) &= E[\mathrm{e}^{t(a_1x_1+\cdots+a_nx_n)}] = \prod_{j=1}^{n} E[\mathrm{e}^{ta_jx_j}] \\ &= \prod_{j=1}^{n} \mathrm{e}^{ta_j\mu + \frac{1}{2}t^2a_j^2\sigma^2} = \mathrm{e}^{t\mu(\sum_{j=1}^n a_j) + \frac{t^2\sigma^2}{2}(\sum_{j=1}^n a_j^2)} \end{aligned}$$

due to $x_1, \dots, x_n$ being iid normal variables. But this mgf is that of a normal variable with mean value $\mu \sum_{j=1}^n a_j$ and variance $\sigma^2 \sum_{j=1}^n a_j^2$. Therefore,

$$u_1 \sim N\left(\mu \sum_{j=1}^n a_j, \sigma^2 \sum_{j=1}^n a_j^2\right)$$

where "∼" indicates "distributed as".

**Note 10.6.** If $N(\mu_j, \sigma_j^2)$, $j = 1, \dots, n$ are independently distributed and if $u_1 = a_1x_1 + \cdots + a_nx_n$ then from the above procedure, it is evident that

$$u_1 \sim N\left(\sum_{j=1}^n a_j\mu_j, \sum_{j=1}^n a_j^2\sigma_j^2\right). \tag{10.23}$$

**Result 10.5.** *If $x_j \sim N(\mu_j, \sigma_j^2)$, $j = 1, \dots, k$ and mutually independently distributed and if $u = a_1x_1 + \cdots + a_kx_k$ is a linear function, where $a_1, \dots, a_k$ are constants, then*

$$u \sim N\left(\sum_{j=1}^k a_j\mu_j, \sum_{j=1}^k a_j^2\sigma_j^2\right). \tag{10.24}$$

(2) Putting $k = n$, $a_1 = \cdots = a_n = \frac{1}{n}$, $\mu_1 = \cdots = \mu_n = \mu$, $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$ in (10.24) we have

$$u_2 = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad n = 1, 2, \dots; \quad M_{u_2}(t) = \mathrm{e}^{t\mu + \frac{1}{2}\frac{t^2\sigma^2}{n}}.$$

Thus, for each $n$, $\bar{x}$ is again normally distributed with mean value $\mu$ and variance $\sigma^2/n$.

(3)

$$M_{u_3}(t) = \mathrm{e}^{-\mu t} M_{u_2}(t) = \mathrm{e}^{\frac{1}{2}\frac{t^2\sigma^2}{n}}.$$

This means that $u_3 = \bar{x} - \mu \sim N(0, \frac{\sigma^2}{n})$ for every $n = 1, 2, \dots$

(4)

$$M_{u_4}(t) = M_{u_3}\left(\frac{\sqrt{n}}{\sigma}t\right) = \mathrm{e}^{\frac{1}{2}(\frac{\sqrt{n}}{\sigma})^2\frac{t^2\sigma^2}{n}} = \mathrm{e}^{\frac{t^2}{2}}.$$

Therefore, $u_4 \sim N(0,1)$, $n = 1,2,\dots$ Thus, for each $n = 1,2,\dots$ including $n \to \infty$, the standardized sample mean $u_4 = \frac{\bar{x}-E(\bar{x})}{\sqrt{\mathrm{Var}(\bar{x})}}$ is exactly standard normal for each $n$, when the sample comes from a normal population. When the sample comes from other populations with finite variance, we have seen that the standardized sample mean goes to a standard normal variable when the sample size goes to infinity. In the following Figure 10.2, (a) is the density of the sample mean $\bar{x}$, (b) is the density of $\bar{x} - \mu$ and (c) is the density of the standardized variable when the population is normal.

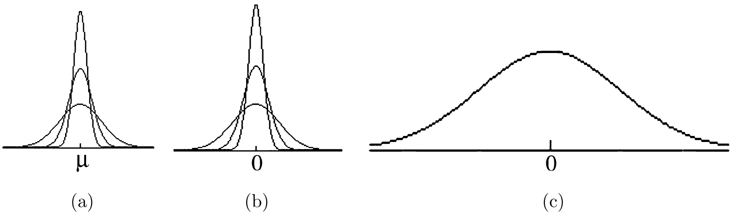


**Figure 10.2:** Density cure for sample mean when the population is Gaussian.

**Example 10.6.** Let $z_1,\dots,z_n$ be iid real scalar random variables following a standard normal distribution $N(0,1)$. Compute the distributions of (1) $u_1 = z_1^2$; (2) $u_2 = z_1^2 + \cdots + z_n^2$.

**Solution 10.6.** (1) This was already done in Module 6. For the sake of completeness, we will repeat here by using transformation of variables. Another method by using the distribution function is given in the exercises. Here, $z_1$ is standard normal and its density is given by

$$f_3(z_1) = \frac{\mathrm{e}^{-\frac{z_1^2}{2}}}{\sqrt{2\pi}}, \quad -\infty < z_1 < \infty.$$

But the transformation $u_1 = z_1^2$ is not one to one since $z_1$ can take negative values also. But in each interval $(-\infty < z_1 < 0)$ and $(0 \le z_1 < \infty)$, the transformation is one to one. Consider the interval $0 \le z_1 < \infty$. Then

$$u_1 = z_1^2 \quad \Rightarrow \quad z_1 = u_1^{\frac{1}{2}} \quad \Rightarrow \quad \mathrm{d}z_1 = \frac{1}{2}u_1^{\frac{1}{2}-1}\mathrm{d}u_1$$

and that part of the density of $u_1$, denoted by

$$\begin{aligned} g_{31}(u_1) &= \frac{1}{2}\frac{1}{\sqrt{2\pi}}u_1^{\frac{1}{2}-1}\mathrm{e}^{-\frac{u_1}{2}} \\ &= \frac{1}{2}\frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}u_1^{\frac{1}{2}-1}\mathrm{e}^{-\frac{u_1}{2}}, \quad 0 \le u_1 < \infty. \end{aligned}$$

But in the interval $(-\infty, 0)$ also the function $f_3(z_1)$ is the same and an even function. Hence the density of $u_1$, denoted by $g_3(u_1)$, is given by

$$g_3(u_1) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} u_1^{\frac{1}{2}-1} e^{-\frac{u_1}{2}}, & 0 \leq u_1 < \infty \\ 0, & \text{elsewhere} \end{cases}$$

which is a gamma density with $\alpha = \frac{1}{2}$ and $\beta = 2$ or it is a chi-square density with one degree of freedom or

$$z_1 \sim N(0,1), \quad u_1 = z_1^2 \sim \chi_1^2; \quad M_{u_1}(t) = (1-2t)^{-\frac{1}{2}}, \quad 1-2t > 0. \tag{10.25}$$

(2) Since the mgf of a sum of independent variables is the product of the individual mgf, we have

$$M_{u_2}(t) = \prod_{j=1}^{n} M_{u_1}(t) = [M_{u_1}(t)]^n = (1-2t)^{-\frac{n}{2}}.$$

Therefore, we have the following result.

**Result 10.6.** *For $z_1, \ldots, z_n$ iid with common distribution $N(0,1)$, then*

$$z_j^2 \sim \chi_1^2 \quad \textit{and} \quad u_3 = \sum_{j=1}^{n} z_j^2 \sim \chi_n^2. \tag{10.26}$$

## Exercises 10.2

**10.2.1.** If $x_1, \ldots, x_n$ are iid from a uniform population over $[0,1]$, evaluate the density of $x_1 + \cdots + x_n$ for (1) $n = 2$; (2) $n = 3$. What is the distribution in the general case?

**10.2.2.** If $x_1, \ldots, x_n$ are iid Poisson distributed with parameter $\lambda$, then (1) derive the probability function of $u_1 = x_1 + \cdots + x_n$; (2) write down the probability function of $\bar{x}$.

**10.2.3.** If $x_1, \ldots, x_n$ are iid type-1 beta distributed with parameters $(\alpha, \beta)$, then compute the density of (1) $u_1 = x_1 + x_2$; (2) $u_2 = x_1 + x_2 + x_3$.

**10.2.4.** Repeat Exercise 10.2.3 if the population is type-2 beta with the parameters $(\alpha, \beta)$.

**10.2.5.** State the central limit theorem explicitly if the sample comes from (1) type-1 beta population; (2) type-2 beta population.

**10.2.6.** Let $x_1, \ldots, x_n$ be iid Bernoulli distributed with parameter $p$, $0 < p < 1$. Let

$$u_1 = x_1 + \cdots + x_n - np; \quad u_2 = \frac{u_1}{\sqrt{np(1-p)}}; \quad u_3 = \frac{u_1 + \frac{1}{2}}{\sqrt{np(1-p)}}.$$

Using a computer, or otherwise, evaluate $y$ so that $\Pr\{|u_2| \geq y\} = 0.05$ for $n = 10, 20, 30, 50$ and compute $n_0$ such that for all $n \geq n_0$, $y$ approximates to the corresponding standard normal value 1.96.

**10.2.7.** Repeat Exercise 10.2.6 with $u_3$ of Example 10.4 and make comments about binomial approximation to a standard normal variable.

**10.2.8.** Let $x$ be a gamma random variable with parameters $(n, \beta)$, $n = 1, 2, \ldots$. Compute the mgf of (1) $u_1 = \bar{x}$; (2) $u_2 = \bar{x} - n\beta$; (3) $u_3 = \frac{\bar{x}-n\beta}{\beta\sqrt{n}}$; (4) show that $u_3$ goes to a standard normal variable when $n \to \infty$.

**10.2.9.** Interpret (4) of Exercise 10.2.8 in terms of the central limit theorem. Which is the population and which is the sample?

**10.2.10.** Is there any connection between central limit theorem and infinite divisibility of real random variables? Explain.

**10.2.11.** Let $z \sim N(0, 1)$. Let $y = z^2$. Compute the following probabilities: (1) $\Pr\{y \leq u\} = \Pr\{z^2 \leq u\} = \Pr\{|z| \leq \sqrt{u}\}$; (2) by using (1) derive the distribution function of $y$ and thereby the density of $y$; (3) show that $y \sim \chi_1^2$.

**10.2.12.** Let $x_1, \ldots, x_n$ be iid with $E(x_j) = \mu_j$, $\operatorname{Var}(x_j) == \sigma_j^2 < \infty$, $j = 1, \ldots, n$. Let $\bar{x} = \frac{x_1+\cdots+x_n}{n}$. Consider the standardized $\bar{x}$,

$$u = \frac{\bar{x} - E(\bar{x})}{\sqrt{\operatorname{Var}(\bar{x})}} = \frac{\sum_{j=1}^{n}(x_j - \mu_j)}{\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}}.$$

Assuming the existence of the mgf of $x_j$, $j = 1, \ldots, n$ work out a condition on $\sigma_1^2 + \cdots + \sigma_n^2$ so that $u \to z \sim N(0, 1)$ as $n \to \infty$.

**10.2.13.** Generalize Exercise 10.2.12 when $u$ is the standardized $v = a_1x_1 + \cdots + a_nx_n$ when $X' = (x_1, \ldots, x_n)$ has a joint distribution with covariance matrix $\Sigma = (\sigma_{ij})$ with $\|\Sigma\| < \infty$ and $a' = (a_1, \ldots, a_n)$ is a fixed vector of constants and $\|(\cdot)\|$ denotes a norm of $(\cdot)$.

## 10.3 Sampling distributions when the population is normal

Here, we will investigate some sampling distributions when the population is normal. We have seen several results in this category already. Let $x \sim N(\mu, \sigma^2)$ be the population and let $x_1, \ldots, x_n$ be a simple random sample from this population. Then we have seen that

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Note that when $n$ becomes larger and larger then the variance of $\bar{x}$ becomes smaller and smaller and finally it goes to zero. In other words, the normal distribution degenerates to the point at $\bar{x} = \mu$ with the total probability mass 1 at this point. That is, $\bar{x}$ converges to $\mu$ with probability 1. This, in fact, is a general result, which was stated as the weak law of large numbers in Chapter 9. We have the following results from the discussions so far.

**Result 10.7.**

$$x_j \sim N(\mu,\sigma^2) \quad \Rightarrow \quad z_j = \frac{x_j-\mu}{\sigma} \sim N(0,1); \quad z_j^2 = \left(\frac{x_j-\mu}{\sigma}\right)^2 \sim \chi_1^2$$

$$\sum_{j=1}^n x_j \sim N(n\mu, n\sigma^2); \quad \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right); \quad u_1 = \frac{\sqrt{n}}{\sigma}(\bar{x}-\mu) \sim N(0,1);$$

$$u_1^2 = \frac{n}{\sigma^2}(\bar{x}-\mu)^2 \sim \chi_1^2; \quad \left(\frac{x_1-\mu}{\sigma}\right)^2 \sim \chi_1^2; \quad \sum_{j=1}^n \left(\frac{x_j-\mu}{\sigma}\right)^2 \sim \chi_n^2.$$

**Result 10.8.** *From Result* 10.7, *we have the following when the population is* $N(\mu,\sigma^2)$*:*

$$E\left[\left(\frac{x_j-\mu}{\sigma}\right)^2\right] = E[\chi_1^2] = 1; E\left[\sum_{j=1}^n \left(\frac{x_j-\mu}{\sigma}\right)^2\right] = E[\chi_n^2] = n;$$

$$\mathrm{Var}\left[\left(\frac{x_j-\mu}{\sigma}\right)^2\right] = \mathrm{Var}(\chi_1^2) = 2;$$

$$\mathrm{Var}\left[\sum_{j=1}^n \left(\frac{x_j-\mu}{\sigma}\right)^2\right] = \mathrm{Var}(\chi_n^2) = 2n. \tag{10.27}$$

**Note 10.7.** Corresponding properties hold even if $x_1,\dots,x_n$ are not identically distributed but independently distributed as $x_j \sim N(\mu_j,\sigma_j^2)$, $j = 1,\dots,n$.

**Result 10.9.** *For* $x_j \sim N(\mu_j,\sigma_j^2)$, $j = 1,\dots,n$ *and independently distributed, we have the following results:*

$$x_j - \mu_j \sim N(0,\sigma_j^2); \quad \frac{x_j-\mu_j}{\sigma_j} \sim N(0,1);$$

$$\left(\frac{x_j-\mu_j}{\sigma_j}\right)^2 \sim \chi_1^2; \quad \sum_{j=1}^n \left(\frac{x_j-\mu_j}{\sigma_j}\right)^2 \sim \chi_n^2. \tag{10.28}$$

**Example 10.7.** Compute the expected value of the sample variance when the sample comes from any population with finite variance and compute the distribution of the sample variance when the sample comes from a normal population.

**Solution 10.7.** Let $x_1,\dots,x_n$ be a simple random sample from any population with mean value $\mu$ and variance $\sigma^2 < \infty$. Then

$$E(x_j)=\mu;\quad E(x_j-\mu)=0;\quad E\left(\frac{x_j-\mu}{\sigma}\right)=0;$$
$$\mathrm{Var}(x_j)=\sigma^2;\quad \mathrm{Var}(x_j-\mu)=\sigma^2;\quad \mathrm{Var}\left(\frac{x_j-\mu}{\sigma}\right)^2=1;\quad E[\bar{x}]=\mu;$$
$$E[\bar{x}-\mu]=0;\quad \mathrm{Var}(\bar{x})=\frac{\sigma^2}{n}.$$

The sample variance can be represented as follows:

$$\begin{aligned} s^2=\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{n}&=\frac{1}{n}\sum_{j=1}^{n}(x_j-\mu+\mu-\bar{x})^2\\ &=\frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)^2+(\bar{x}-\mu)^2+\frac{2}{n}(\mu-\bar{x})\sum_{j=1}^{n}(x_j-\mu)\\ &=\frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)^2+(\bar{x}-\mu)^2-2(\bar{x}-\mu)^2\\ &=\frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)^2-(\bar{x}-\mu)^2. \end{aligned} \tag{10.29}$$

Taking expectations on both sides, we have

$$E(s^2)=\frac{1}{n}\sum_{j=1}^{n}\mathrm{Var}(x_j)-\mathrm{Var}(\bar{x})=\frac{1}{n}n\sigma^2-\frac{\sigma^2}{n}=\frac{n-1}{n}\sigma^2.$$

This shows that

$$E\left[\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{n-1}\right]=\sigma^2, \tag{10.30}$$

a property called unbiasedness, which will be discussed in the chapter on estimation. The above result says that $\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{n-1}$ is unbiased for the population variance, whatever be the population, as long as the population variance is finite.

If the population is normal, then we have shown that

$$\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\sim\chi_n^2\quad\text{and}\quad\frac{n}{\sigma^2}(\bar{x}-\mu)^2\sim\chi_1^2.$$

But from (10.29), it is evident that

$$\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\equiv\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}+\frac{n}{\sigma^2}(\bar{x}-\mu)^2$$

which means

$$\chi_n^2\equiv\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}+\chi_1^2. \tag{10.31}$$

But we have the property that if $\chi_m^2$ and $\chi_n^2$ are independently distributed then

$$\chi_m^2+\chi_n^2\equiv\chi_{m+n}^2. \tag{10.32}$$

Independence of $\bar{x}$ and $s^2$ will guarantee from (10.31), by looking at the mgf of both sides in (10.31), that

$$\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}\sim\chi^2_{n-1}. \tag{10.33}$$

But it can be shown that if the sample comes from a normal population, then $\bar{x}$ and $\sum_{j=1}^{n}(x_j-\bar{x})^2$ are independently distributed. Hence for the normal population, result (10.33) holds. Independence will be proved later on. Independence of $\bar{x}$ and $s^2$, along with some minor conditions, will in fact, characterize a normal population; see the book [11].

**Note 10.8.** If $x$ and $y$ are real scalar random variables and if $x$ and $y$ are independently distributed and if $x_1=a_1x+b_1$, $y_1=a_2y+b_2$, where $a_1\neq 0$, $a_2\neq 0$, $b_1$, $b_2$, are constants, then $x_1$ and $y_1$ are also independently distributed. Are $x$ and $y^2$ independently distributed? Are $x^2$ and $y$ independently distributed? Are $x^2$ and $y^2$ independently distributed?

**Note 10.9.** If $x$ and $y$ are real scalar random variables and if $x^2$ and $y^2$ are independently distributed, then are the following independently distributed: (1) $x^2$ and $y$; (2) $x$ and $y^2$; (3) $x$ and $y$ ?

**Example 10.8** (Non-central chi-square). Let $x_1,\dots,x_n$ be iid variables from a $N(\mu,\sigma^2)$. Evaluate the density of $u=\sum_{j=1}^{n}\frac{x_j^2}{\sigma^2}$, $\mu\neq 0$. [This is known as a non-central chi-square with $n$ degrees of freedom and non-centrality parameter $\lambda=\frac{n\mu^2}{2\sigma^2}$ and it is written as $u\sim\chi^2_n(\lambda)$ because when $\mu$ is present, then $\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\sim\chi^2_n$ or central chi-square with $n$ degrees of freedom.]

**Solution 10.8.** Since the joint density of $x_1,\dots,x_n$ is available, let us compute the mgf of $u$, that is, $M_u(t)=E[\mathrm{e}^{tu}]$.

$$\begin{aligned}M_u(t)=E[\mathrm{e}^{tu}]&=\int\cdots\int\frac{1}{(\sigma\sqrt{2\pi})^n}\\&\quad\times\exp\left\{t\sum_{j=1}^{n}\frac{x_j^2}{\sigma^2}-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\right\}\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_n.\end{aligned}\tag{10.34}$$

Let us simplify the exponent:

$$\begin{aligned}&t\sum_{j=1}^{n}\frac{x_j^2}{\sigma^2}-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\\&\quad=-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{n}(1-2t)x_j^2-2\mu\sum_{j=1}^{n}x_j+n\mu^2\right]\end{aligned}$$

$$= -\lambda + \frac{\lambda}{1-2t} - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(\sqrt{1-2t}x_j - \frac{\mu}{\sqrt{1-2t}}\right)^2$$

for $1-2t > 0$ and $\lambda = \frac{n\mu^2}{2\sigma^2}$. Put $y_j = \sqrt{1-2t}x_j - \frac{\mu}{\sqrt{1-2t}}$ and integrate variables one at a time. For $x_j$, the integral is the following:

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2\sigma^2}(\sqrt{1-2t}x_j - \frac{\mu}{\sqrt{1-2t}})^2}\mathrm{d}x_j = \frac{1}{\sqrt{1-2t}}$$

from the total integral of a normal density. Therefore,

$$M_u(t) = \frac{\mathrm{e}^{-\lambda}}{(1-2t)^{n/2}}\mathrm{e}^{\frac{\lambda}{1-2t}} = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\frac{1}{(1-2t)^{\frac{n}{2}+k}}.$$

But we know that $(1-2t)^{-(\frac{n}{2}+k)}$ is the mgf of a chi-square with $n+2k$ degrees of freedom and its density is a gamma with parameters $(\alpha = \frac{n}{2}+k, \beta = 2)$ and hence the density of $u$, denoted by $g(u)$, is given by

$$g(u) = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\frac{u^{\frac{n}{2}+k-1}}{2^{\frac{n}{2}+k}\Gamma(\frac{n}{2}+k)}\mathrm{e}^{-\frac{u}{2}}, \quad u \ge 0 \tag{10.35}$$

and zero elsewhere. This is the non-central chi-square density, which is in the form of a weighted gamma (chi-square) densities, weights being Poisson probabilities or Poisson-weighted chi-square densities with $n+2k$ degrees of freedom. Observe also that since $\lambda > 0$ we have $P_k = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$ with $\sum_{k=0}^{\infty} P_k = 1$ or the coefficients are from a Poisson distribution. The non-central chi-square density is of the form:

$$g(u) = \sum_{k=0}^{\infty} f_k(u)P_k$$

where

$$f_k(u) = \frac{u^{\frac{n}{2}+k-1}\mathrm{e}^{-\frac{u}{2}}}{2^{\frac{n}{2}+k}\Gamma(\frac{n}{2}+k)}, \quad u \ge 0 \tag{10.36}$$

and

$$P_k = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}.$$

This density is a very important density, which is also connected to Bessel function in the theory of special functions.

## Exercises 10.3

**10.3.1.** If $x_1, x_2, x_3$ are independently distributed so that $x_1 \sim N(0, \sigma^2 = 1)$, $x_2 \sim N(\mu = 2, \sigma^2 = 4)$, $x_3 \sim N(\mu = -1, \sigma^2 = 5)$ evaluate the densities of the following: (1) $u_1 = x_1 + x_2 + x_3$; (2) $u_2 = 2x_1 - 3x_2 + 5x_3$; (3) $u_3 = x_1^2 + \frac{(x_2-2)^2}{4} + \frac{(x_3+1)^2}{5}$; (4) $u_4 = x_1^2 + (x_2-2)^2$; (5) $u_5 = x_1^2 + \frac{x_2^2}{4} + \frac{x_3^2}{5}$.

**10.3.2.** By using the mgf or otherwise compute the exact densities of (1) $u_1 = \chi_n^2 - n$; (2) $u_2 = \frac{\chi_n^2 - n}{\sqrt{2n}}$; (3) show that $u_2 \to z \sim N(0,1)$ as $n \to \infty$.

**10.3.3.** Interpret (3) in Exercise 10.3.2 in terms of the central limit theorem. What is the population and what is the sample?

**10.3.4.** By using a computer (1) compute $y$ so that $\Pr\{|u_2| \geq y\} = 0.05$ for $n = 10, 20, 30$ where $u_2$ is the same $u_2$ in Exercise 10.3.2; (2) determine $n$ so that $y$ approximates well with the corresponding $N(0,1)$ value 1.96 at the 5% level (tail area is 0.05).

**10.3.5.** Find $n_0$ such that for $n \geq n_0$ the standard normal approximation in Exercise 10.3.4 holds well.

## 10.4 Student-*t* and *F* distributions

The distribution of a random variable of the type $u = \frac{z}{\sqrt{y/\nu}}$ where $z$ is a standard normal, $z \sim N(0,1)$, $y$ is a chi-square with $\nu$ degrees of freedom, $y \sim \chi_\nu^2$, where $z$ and $y$ are independently distributed, is known as a Student-*t* variable with $\nu$ degrees of freedom, $t_\nu$.

**Definition 10.6** (A Student-*t* statistic $t_\nu$). A Student-*t* variable with $\nu$ degrees of freedom is defined as

$$t_\nu = u = \frac{z}{\sqrt{y/\nu}}, \quad z \sim N(0,1),\ y \sim \chi_\nu^2 \tag{10.37}$$

where $z$ and $y$ are independently distributed.

The person who derived the density of the variable $t_\nu$, W. Gossett, wrote the paper under the name "a student", and hence the distribution is known in the literature as the Student-*t* distribution. Before deriving the density, let us examine more general structures and derive the density as a special case of such a general structure.

**Example 10.9.** Let $x_1$ and $x_2$ be independently distributed real gamma random variables with the parameters $(\alpha_1, \beta)$ and $(\alpha_2, \beta)$, respectively, that is, the scale parameters are equal to some $\beta > 0$. Let $u_1 = x_1 + x_2$, $u_2 = \frac{x_1}{x_1 + x_2}$, $u_3 = \frac{x_1}{x_2}$. Derive the distributions of $u_1, u_2, u_3$.

**Solution 10.9.** Due to independence, the joint density of $x_1$ and $x_2$, denoted by $f(x_1, x_2)$, is given by

$$f(x_1, x_2) = \frac{x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \mathrm{e}^{-\frac{1}{\beta}(x_1 + x_2)}}{\beta^{\alpha_1 + \alpha_2} \Gamma(\alpha_1)\Gamma(\alpha_2)}, \quad 0 \leq x_i < \infty, \quad i = 1, 2 \tag{10.38}$$

and $f(x_1,x_2) = 0$ elsewhere. Let us make the polar coordinate transformation: $x_1 = r\cos^2\theta$, $x_2 = r\sin^2\theta$. We have taken $\cos^2\theta$ and $\sin^2\theta$ due to the presence of $x_1 + x_2$ in the exponent. The Jacobian of the transformation is the determinant

$$\begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos^2\theta & -2r\cos\theta\sin\theta \\ \sin^2\theta & 2r\cos\theta\sin\theta \end{vmatrix} = 2r\cos\theta\sin\theta.$$

The joint density of $r$ and $\theta$, denoted by $g(r,\theta)$, is given by

$$\begin{aligned} g(r,\theta) &= \frac{(r\cos^2\theta)^{\alpha_1-1}(r\sin^2\theta)^{\alpha_2-1}}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)}\mathrm{e}^{-\frac{r}{\beta}}2r\cos\theta\sin\theta \\ &= \frac{r^{\alpha_1+\alpha_2-1}\mathrm{e}^{-\frac{r}{\beta}}}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2)}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\ &\quad\times(\cos^2\theta)^{\alpha_1-1}(\sin^2\theta)^{\alpha_2-1}2\cos\theta\sin\theta \\ &= g_1(r)g_2(\theta) \end{aligned} \tag{10.39}$$

by multiplying and dividing by $\Gamma(\alpha_1+\alpha_2)$. From (10.39), a few properties are obvious. (a) $x_1 + x_2 = r$ has the density $g_1(r)$, which is a gamma density, and hence $u_1 = x_1 + x_2$ is gamma distributed with parameters $(\alpha_1+\alpha_2,\beta)$; (b) Since (10.39) is in the form of a product of two densities, one is a function of $r$ alone and the other is a function of $\theta$, the variables $r$ and $\theta$ are independently distributed. (c)

$$u_2 = \frac{x_1}{x_1+x_2} = \frac{r\cos^2\theta}{r\cos^2\theta + r\sin^2\theta} = \cos^2\theta$$

is a function of $\theta$ alone. Hence $u_1 = x_1 + x_2$ and $u_2 = \frac{x_1}{x_1+x_2}$ are independently distributed. (d) But $x_1 = u_1u_2 \Rightarrow E(x_1^h) = E(u_1^h)E(u_2^h)$ due to the independence of $u_1$ and $u_2$. Therefore, we have the following result.

**Result 10.10.** *When the real scalar random variables $x_1$ and $x_2$ are independently distributed as gamma variables with parameters $(\alpha_1,\beta)$ and $(\alpha_2,\beta)$, respectively, with the same $\beta$, and when $u_1 = x_1 + x_2$, $u_2 = \frac{x_1}{x_1+x_2}$, then*

$$E(u_1^h) = \frac{E(x_1^h)}{E(u_2^h)} \quad\Rightarrow\quad E\left[\frac{x_1}{x_1+x_2}\right]^h = \frac{E(x_1^h)}{E(x_1+x_2)^h}$$

Note that even if $y_1$ and $y_2$ are independently distributed, $E(\frac{y_1}{y_2})^h \neq \frac{E(y_1^h)}{E(y_2^h)}$. From (10.29), the density of $u_2 = \cos^2\theta$ is the following: The non-zero part of the density of $x_1$ and $x_2$ is in the first quadrant, and hence $0 \le r < \infty$, $0 \le \theta \le \frac{\pi}{2}$. Note that $\mathrm{d}u_2 = 2\cos\theta\sin\theta\mathrm{d}\theta$. But the density of $\theta$ is

$$g_2(\theta) = (\cos^2\theta)^{\alpha_1-1}(\sin^2\theta)^{\alpha_2-1}2\cos\theta\sin\theta, \quad 0 \le \theta \le \frac{\pi}{2} \tag{10.40}$$

and $g_2(\theta) = 0$ elsewhere. Also, for $0 \le \theta \le \frac{\pi}{2}$ means $0 \le u_2 = \cos^2\theta \le 1$. Hence the density of $u_2$ is given by

$$g_{u_2}(u_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u_2^{\alpha_1-1}(1-u_2)^{\alpha_2-1}, \quad 0 \le u_2 \le 1,\ \alpha_1 > 0,\ \alpha_2 > 0 \tag{10.41}$$

and $g_{u_2}(u_2) = 0$ elsewhere. Hence $u_2$ is a type-1 beta variable with parameters $(\alpha_1, \alpha_2)$, and therefore $1 - u_2$ is a type-1 beta with the parameters $(\alpha_2, \alpha_1)$.

**Result 10.11.** *If* $x_1$ *and* $x_2$ *are as in Result* 10.10, *then* $u_2 = \frac{x_1}{x_1+x_2}$ *is a type-1 beta random variable with the parameters* $(\alpha_1, \alpha_2)$.

(e) $u_3 = \frac{x_1}{x_2} = \frac{r\cos^2\theta}{r\sin^2\theta} = \cot^2\theta$. We can evaluate the density of $u_3 = \cot^2\theta$ either from the density of $\theta$ in (10.40) or from the density of $u_2$ in (10.41).

$$u_2 = \frac{x_1}{x_1+x_2} = \frac{x_1/x_2}{1+x_1/x_2} = \frac{u_3}{1+u_3} \quad \Rightarrow \quad u_3 = \frac{u_2}{1-u_2}.$$
$$\mathrm{d}u_3 = \frac{1}{(1-u_2)^2}\mathrm{d}u_2 = (1+u_3)^2\mathrm{d}u_2 \quad \Rightarrow \quad \mathrm{d}u_2 = \frac{1}{(1+u_3)^2}\mathrm{d}u_3.$$

From (10.41), the density of $u_3$, denoted by $g_{u_3}(u_3)$, is given by

$$g_{u_3}(u_3) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\left(\frac{u_3}{1+u_3}\right)^{\alpha_1-1}\left(\frac{1}{1+u_3}\right)^{\alpha_2-1}\frac{1}{(1+u_3)^2}$$
$$= \begin{cases} \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} u_3^{\alpha_1-1}(1+u_3)^{-(\alpha_1+\alpha_2)}, \\ \qquad 0 \le u_3 < \infty,\ \alpha_1 > 0,\ \alpha_2 > 0 \\ 0, \quad \text{elsewhere.} \end{cases} \tag{10.42}$$

Therefore, $u_3 = \frac{x_1}{x_2}$ is type-2 beta distributed with parameters $(\alpha_1, \alpha_2)$ and then $u_4 = \frac{x_2}{x_1}$ is type-2 beta distributed with the parameters $(\alpha_2, \alpha_1)$.

**Result 10.12.** *If* $x_1$ *and* $x_2$ *are as in Result* 10.10, *then* $u_3 = \frac{x_1}{x_2}$ *is a type-2 beta with the parameters* $(\alpha_1, \alpha_2)$.

Now, consider a particular case of a gamma variable, namely the chi-square variable. Let $x_1 \sim \chi^2_m$ and $x_2 \sim \chi^2_n$ be independently distributed. $x_1 \sim \chi^2_m$ means a gamma with the parameters $(\alpha = \frac{m}{2}, \beta = 2)$. Then

$$u_3 = \frac{x_1}{x_2}$$

has the density, for $m, n = 1, 2, \ldots$,

$$f_{u_3}(u_3) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} u_3^{\frac{m}{2}-1}(1+u_3)^{-(\frac{m+n}{2})}, \quad 0 \le u_3 < \infty, \tag{10.43}$$

and $f_{u_3}(u_3) = 0$ elsewhere. A constant multiple, namely,

$$\frac{n}{m}u_3 = \frac{\chi_m^2/m}{\chi_n^2/n}$$

or a chi-square with $m$ degrees of freedom divided by its degrees of freedom and the whole thing is divided by a chi-square with $n$ degrees of freedom, divided by its degrees of freedom, where the two chi-squares are independently distributed, is known as a $F$ random variable with $m$ and $n$ degrees of freedom and usually written as $F_{m,n}$.

**Definition 10.7** ($F$ random variable). A $F = F_{m,n}$ random variable is defined, and it is connected to $u_3$, as follows:

$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n} = \frac{n}{m}u_3 \quad \Rightarrow \quad u_3 = \frac{m}{n}F_{m,n}.$$

Then the $F$-density is available from the type-2 beta density or from (10.43), and it is the following:

$$f_F(F_{m,n}) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{\frac{m}{2}} F_{m,n}^{\frac{m}{2}-1}\left(1+\frac{m}{n}F_{m,n}\right)^{-(\frac{m+n}{2})} \tag{10.44}$$

for $0 \le F_{m,n} < \infty$, $m, n = 1, 2, \dots$ and $f_F(F_{m,n}) = 0$ elsewhere.

**Special Case for $m = 1, n = \nu$.** Let $F_{1,\nu} = t_\nu^2$. Then putting $m = 1$, $n = \nu$ in (10.44) we have the density of $F_{1,\nu} = \frac{\chi_1^2}{\chi_\nu^2/\nu} = t_\nu^2$ which will then be the density of the square of Student-$t$ with $\nu$ degrees of freedom. Denoting the density by $f_t(t_\nu^2)$, we have the following:

$$f_t(t_\nu^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}(t_\nu^2)^{\frac{1}{2}-1}\left(1+\frac{t_\nu^2}{\nu}\right)^{-(\frac{\nu+1}{2})}, \quad 0 \le t_\nu^2 < \infty.$$

Note that

$$t_\nu^2 = F_{1,\nu} \quad \Rightarrow \quad |t_\nu| = \frac{\sqrt{\chi_1^2}}{\sqrt{\chi_\nu^2/\nu}} = \frac{|z_1|}{\sqrt{\chi_\nu^2/\nu}}, \quad t_\nu = \frac{z_1}{\sqrt{\chi_\nu^2/\nu}}$$

where $z_1$ is a standard normal variable. For $t_\nu > 0$, it is a one to one transformation and

$$\mathrm{d}t_\nu = \frac{1}{2}F_{1,\nu}^{-\frac{1}{2}}\mathrm{d}F_{1,\nu} \quad \text{or} \quad 2\mathrm{d}t_\nu = F_{1,\nu}^{-\frac{1}{2}}\mathrm{d}F_{1,\nu}.$$

Hence for $t_\nu > 0$ the folded Student-$t$ density is given by

$$f_t(t_\nu) = 2\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{t_\nu^2}{\nu}\right)^{-(\frac{\nu+1}{2})}, \quad 0 \le t_\nu < \infty. \tag{10.45}$$

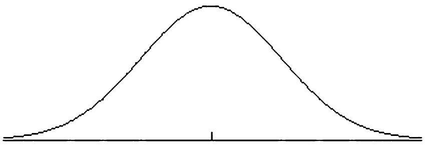


**Figure 10.3:** Student-*t* density.

Since it is symmetric about $t_\nu = 0$, the Student-*t* density is given by

$$f_t(t_\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1 + \frac{t_\nu^2}{\nu}\right)^{-(\frac{\nu+1}{2})}, \quad -\infty < t_\nu < \infty.$$

A graph of Student-t density is given in Figure 10.3. Another way of deriving the density directly from the joint density of independently distributed standard normal variable and a chi-square variable is by using transformation of variables. Since $t_\nu = \frac{z}{\sqrt{y/\nu}}$, where $z \sim N(0,1)$ and $y \sim \chi_\nu^2$ where $z$ and $y$ are independently distributed, for $z > 0$ the transformation is one to one, and similarly for $z < 0$ also the transformation is one to one. Let $z > 0$. Take $u = \frac{z}{\sqrt{y/\nu}}$ and $v = y$. Then $\mathrm{d}z \wedge \mathrm{d}y = \sqrt{\frac{y}{\nu}}\mathrm{d}u \wedge \mathrm{d}v$. The joint density of $z$ and $y$, denoted by $f(z,y)$, is given by

$$f(z,y) = \frac{\mathrm{e}^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \times \frac{y^{\frac{\nu}{2}-1}\mathrm{e}^{-\frac{y}{2}}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}, \quad z = u\sqrt{\frac{y}{\nu}}, \ y = v$$

and then the joint density of $u$ and $v$, denoted by $g(u,v)$, is given by

$$\begin{aligned} g(u,y) &= \frac{\mathrm{e}^{-\frac{u^2y}{2\nu}}}{\sqrt{2\pi}} \frac{y^{\frac{\nu}{2}-1}\mathrm{e}^{-\frac{y}{2}}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} \frac{y^{\frac{1}{2}}}{\nu^{\frac{1}{2}}} \\ &= \frac{y^{\frac{\nu+1}{2}-1}\mathrm{e}^{-y(\frac{1}{2}+\frac{u^2}{2\nu})}}{2^{\frac{\nu+1}{2}}\Gamma(\frac{\nu}{2})\sqrt{\nu}}. \end{aligned}$$

Integrating out $y$, we have that part of the marginal density for $u$ given by

$$\begin{aligned} g_1(u) &= \frac{1}{2^{\frac{\nu+1}{2}}\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \int_0^\infty y^{\frac{\nu+1}{2}-1}\mathrm{e}^{-\frac{1}{2}y}\left(1 + \frac{u^2}{\nu}\right)\mathrm{d}y \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}\left(1 + \frac{u^2}{\nu}\right)^{-(\frac{\nu+1}{2})}, \quad u > 0 \end{aligned}$$

The same is the function for $u < 0$, and hence for $-\infty < t_\nu < \infty$, $u^2 = t_\nu^2$ we have the density of $t_\nu$ given by

$$f_t(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-(\frac{\nu+1}{2})}, \quad -\infty < t_\nu < \infty.$$

Observe that when the sample $x_1, \ldots, x_n$ comes from a normal population $N(\mu, \sigma^2)$ we have

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0,1)$$

and

$$\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma^2}\sim\chi^2_{n-1}$$

and these two are independently distributed. Hence

$$t_{n-1}=\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\Big/\left[\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{(n-1)\sigma^2}\right]^{\frac{1}{2}}=\frac{\sqrt{n}(\bar{x}-\mu)}{\hat{\sigma}}$$

is a Student-$t$ with $n-1$ degrees of freedom, where $\hat{\sigma}^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n-1}$, which is the unbiased estimator for $\sigma^2$. Thus, if $\sigma^2$ is replaced by its unbiased estimator $\hat{\sigma}^2$ then the standardized normal variable changes to a Student-$t$ variable with $n-1$ degrees of freedom.

$$\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\sim N(0,1)\quad\text{and}\quad\frac{\sqrt{n}(\bar{x}-\mu)}{\hat{\sigma}}\sim t_{n-1}. \tag{10.46}$$

We also have a corresponding distribution on variance ratios. Consider two independent populations $N(\mu_1,\sigma_1^2)$ and $N(\mu_2,\sigma_2^2)$ and let $x_1,\dots,x_m$ and $y_1,\dots,y_n$ be iid variables from these two populations respectively. Then

$$\begin{aligned}\frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2}&=\frac{\sum_{i=1}^{m}(x_i-\bar{x})^2/((m-1)\sigma_1^2)}{\sum_{i=1}^{n}(y_i-\bar{y})^2/((n-1)\sigma_2^2)}\sim F_{m-1,n-1}\\&=\frac{\sum_{i=1}^{m}(x_i-\bar{x})^2/(m-1)}{\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)}\sim F_{m-1,n-1}\quad\text{for }\sigma_1^2=\sigma_2^2.\end{aligned} \tag{10.47}$$

This $F$-density is also known as the density for the "variance ratio", which is useful in testing hypotheses of the type $\sigma_1^2=\sigma_2^2$. Observe that the results in (10.46) and (10.47) do not hold when the populations are not independent normal populations.

## Exercises 10.4

**10.4.1.** Let $x_1,\dots,x_m$ be iid variables from the population $N(\mu_1,\sigma_1^2)$ and let $y_1,\dots,y_n$ be iid variables from the population $N(\mu_2,\sigma_2^2)$ and let all variables be mutually independently distributed. [This is also known as samples coming from independent populations.] Let $\bar{x}=\sum_{j=1}^{m}\frac{x_j}{m}$ and $\bar{y}=\sum_{j=1}^{n}\frac{y_j}{n}$. Then show that the following results hold:

(1) $\left(\frac{x_j-\mu_1}{\sigma_1}\right)^2\sim\chi_1^2$; (2) $s_1^2=\sum_{j=1}^{m}\left(\frac{x_j-\mu_1}{\sigma_1}\right)^2\sim\chi_m^2$;

(3) $\left(\frac{y_j-\mu_2}{\sigma_2}\right)^2\sim\chi_1^2$; (4) $s_2^2=\sum_{j=1}^{n}\left(\frac{y_j-\mu_2}{\sigma_2}\right)^2\sim\chi_n^2$;

(5) $\frac{s_1^2}{s_2^2}\sim\text{type-2 beta}\left(\frac{m}{2},\frac{n}{2}\right)$;

(6) $\frac{s_1^2}{s_1^2+s_2^2} \sim \text{type-1 beta}\left(\frac{m}{2},\frac{n}{2}\right)$;

(7) $\frac{n}{m}\frac{s_1^2}{s_2^2} \sim F_{m,n}$; (8) $s_3^2 = \sum_{j=1}^{m}\left(\frac{x_j-\bar{x}}{\sigma_1}\right)^2 \sim \chi_{m-1}^2$;

(9) $s_4^2 = \sum_{j=1}^{n}\left(\frac{y_j-\bar{y}}{\sigma_2}\right)^2 \sim \chi_{n-1}^2$;

(10) $\frac{s_3^2}{s_4^2} \sim \text{type-2 beta}\left(\frac{m-1}{2},\frac{n-1}{2}\right)$;

(11) $\frac{s_3^2}{s_3^2+s_4^2} \sim \text{type-1 beta}\left(\frac{m-1}{2},\frac{n-1}{2}\right)$;

(12) $\frac{n-1}{m-1}\frac{s_3^2}{s_4^2} \sim F_{m-1,n-1}$.

Note that when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then all the variances, $\sigma_1^2, \sigma_2^2$, will disappear from all the ratios above. Hence the above results are important in testing hypotheses of the type $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

**10.4.2.** For the same samples in Exercise 10.4.1, evaluate the densities of the following variables:

(1) $u_1 = \frac{1}{2\sigma_1^2}(x_1-x_2)^2$; (2) $u_2 = \frac{1}{2\sigma_2^2}(y_1-y_2)^2$;

(3) $u_3 = \frac{1}{2}(x_1-x_2)^2$; (4) $u_4 = \frac{1}{2}(y_1-y_2)^2$;

(5) $u_5 = \frac{\sigma_2^2}{\sigma_1^2}\frac{(x_1-x_2)^2}{(y_1-y_2)^2}$; (6) $u_6 = \frac{(x_1-x_2)^2}{(y_1-y_2)^2}$;

(7) $u_7 = \sqrt{u_5}$; (8) $\sqrt{u_6}$.

**10.4.3.** Show that $F_{m,n} = \frac{1}{F_{n,m}}$. Let $x = F_{m,n}$ with density $f_1(x)$ and let $y = \frac{1}{x} = F_{n,m}$ with density $f_2(y)$. The notation $F_{m,n,\alpha}$ means the point from where onward to the right the area under the curve $f_1(x)$ is $\alpha$. Then $F_{n,m,1-\alpha}$ means the point from where onward to the right the area under the curve $f_2(y)$ is $1-\alpha$. By using the densities $f_1(x)$ and $f_2(y)$ and then by transforming $y = \frac{1}{x}$ show that

$$F_{m,n,\alpha} = \frac{1}{F_{n,m,1-\alpha}}.$$

[Note that, due to this property, only the right tail areas are tabulated in the case of *F* distribution. Such numerical tables, called *F*-tables, are available.]

**10.4.4.** Notations $\chi_{\nu,\alpha}^2$, $t_{\nu,\alpha}$, $F_{m,n,\alpha}$ mean the point from where onward to the right the area under the curve in the case of chi-square density, Student-*t* density and *F*-density, is $\alpha$. By using a computer, compute $\chi_{\nu,\alpha}^2$, $t_{\nu,\alpha}$, $F_{m,n,\alpha}$ for $\alpha = 0.05$ (5% tables), $\alpha = 0.01$ (1% tables) for various values of $\nu, m, n = 1, 2, \ldots$. [This is equivalent to creating 5% and 1% chi-square, Student-*t* and *F*-tables.]

**10.4.5.** Derive the density of a non-central $F$, where the numerator chi-square is non-central with $m$ degrees of freedom and non-centrality parameter $\lambda$, and the denominator chi-square is central with $n$ degrees of freedom.

**10.4.6.** Derive the density of a doubly non-central $F_{m,n}(\lambda_1, \lambda_2)$ with degrees of freedom $m$ and $n$ and non-centrality parameters $\lambda_1$ and $\lambda_2$.

**10.4.7.** For the standard normal distribution $x \sim N(0,1)$, $\Pr\{|x| \geq y\} = 0.05$ means $y \approx 1.96$. By using a computer, calculate $y$ such that $\Pr\{|t_\nu| \geq y\} = 0.05$ for $\nu = 10, 20, 30, 100$. Then show that a Student-$t$ does not approximate well to a standard normal variable even for $\nu = 100$. Hence conclude that reading from standard normal tables, when the degrees of freedom of a Student-$t$, is greater than or equal to 30 is not a valid procedure.

**10.4.8.** For a type-2 beta variable $x$ with parameters $\alpha$ and $\beta$ show that

$$E(x^h) = \frac{\Gamma(\alpha+h)}{\Gamma(\alpha)} \frac{\Gamma(\beta-h)}{\Gamma(\beta)}, \quad -\alpha < h < \beta$$

when real and $-\Re(\alpha) < \Re(h) < \Re(\beta)$ when in the complex domain. What are the corresponding conditions for (1) $F_{m,n}$ random variable; (2) Student-$t$ variable with $\nu$ degrees of freedom. [Hint: $x = \frac{m}{n} F_{m,n}$.]

**10.4.9.** Show that (1) $E(t_\nu)$ does not exist for $\nu = 1, 2$; (2) $E(t_\nu^2)$ does not exist for $\nu = 3, 4$.

**10.4.10.** Evaluate the $h$-th moment of a non-central chi-square with $\nu$ degrees of freedom and non-centrality parameter $\lambda$, and write down its conditions for existence. Write it as a hypergeometric function, if possible.

**10.4.11.** Evaluate the $h$-th moment of a (1): singly non-central $F_{m,n}(\lambda)$ with numerator chi-square $\chi_m^2(\lambda)$; (2): doubly non-central $F_{m,n}(\lambda_1, \lambda_2)$, and write down the conditions for its existence.

## 10.5 Linear forms and quadratic forms

We have already looked into linear functions of normally distributed random variables in equation (10.24). We have seen that arbitrary linear functions of independently distributed normal variables are again normally distributed. We will show later that this property holds even if the variables are not independently distributed but having a certain form of a joint normal distribution. First, we will look into some convenient ways of writing linear forms and quadratic forms by using vector and matrix notations.

Consider a set of real scalar random variables $x_1, \dots, x_k$ and real scalar constants $a_1, \dots, a_k$. Then linear form is of the type $y$ and a linear expression is of the type $y_1 = y + b$ where $b$ is a constant, where

$$y = a_1x_1 + \cdots + a_kx_k; \quad y_1 = a_1x_1 + \cdots + a_kx_k + b.$$

These can also be written as

$$y = a'X = X'a, \quad y_1 = a'X + b = X'a + b \tag{10.48}$$

where a prime denotes the transpose and

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}, \quad a' = (a_1, a_2, \dots, a_k), \quad X' = (x_1, \dots, x_k).$$

For example,

$$u_1 = 2x_1 - x_2 + x_3 = [2, -1, 1] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = [x_1, x_2, x_3] \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}; \quad u_2 = 2x_1 - x_2 + x_3 + 5 = a'X + 5.$$

Here, $b = 5$.

A simple quadratic form is of the type $X'X = x_1^2 + x_2^2 + \dots + x_k^2$ where $X' = (x_1, \dots, x_k)$ and the prime denotes the transpose. A general quadratic form is of the type

$$X'AX = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}x_ix_j \tag{10.49}$$

$$= \sum_{j=1}^{k} a_{jj}x_j^2 + \sum_{i}\sum_{j,i\neq j} a_{ij}x_ix_j \tag{10.50}$$

$$= \sum_{j=1}^{k} a_{jj}x_j^2 + 2\sum_{i<j} a_{ij}x_ix_j = \sum_{j=1}^{k} a_{jj}x_j^2 + 2\sum_{i>j} a_{ij}x_ix_j \tag{10.51}$$

where the matrix $A = A'$ without loss of generality, $X' = (x_1, \dots, x_k)$. The coefficient of $x_ix_j$ is $a_{ij}$ for all $i$ and $j$, including $i = j$. In (10.49), all terms, including the case $i = j$, are written in a single expression. In (10.50), the diagonal terms and all non-diagonal terms are separated. Due to symmetry, $A = A'$, we have $a_{ij} = a_{ji}$ and hence the coefficients of $x_ix_j$ will be the same as that of $x_jx_i$. Thus some of the terms appear twice and this is reflected in (10.51). For example, for $k = 2$ the above representations are equivalent to the following:

$$\begin{aligned} X'AX &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2 \\ &= a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 \quad \text{since } a_{12} = a_{21} \\ &= a_{11}x_1^2 + a_{22}x_2^2 + 2a_{21}x_2x_1. \end{aligned}$$

**Definition 10.8** (Linear Forms, Quadratic Form, Linear Expressions and Quadratic Expressions). A linear form is where all terms are of degree one each. A linear expression is one where the maximum degree is one. A quadratic form is where all terms are of degree 2 each. A quadratic expression is such that the maximum degree of the terms is two.

Thus, a quadratic expression has a general representation $X'AX + a'X + b$ where $a'X$ is a linear form and $b$ is a scalar constant, $a' = (a_1, \dots, a_k)$ a set of scalar constants. Examples will be of the following types:

$$
\begin{aligned}
&u_1 = x_1^2 + \cdots + x_k^2 \quad \text{(a quadratic form);}\\
&u_2 = 2x_1^2 - 5x_2^2 + x_3^2 - 2x_1x_2 - 6x_2x_3 \quad \text{(a quadratic form);}\\
&u_3 = x_1^2 + 5x_2^2 - 3x_1x_2 + 4x_1 - x_2 + 7 \quad \text{(a quadratic expression).}
\end{aligned}
$$

**Example 10.10.** Write the following in vector, matrix notation:

$$
\begin{aligned}
u_1 &= 2x_1^2 - 5x_2^2 + x_3^2 - 2x_1x_2 + 3x_1 - x_2 + 4;\\
u_2 &= x_1 - x_2 + x_3;\\
u_3 &= x_1^2 + 2x_2^2 - x_3^2 + 4x_2x_3.
\end{aligned}
$$

**Solution 10.10.** Here, $u_1$ is a quadratic expression

$$
\begin{aligned}
&u_1 = X'AX + a'X + b \quad \text{where}\\
&X' = (x_1, x_2, x_3), \quad a' = (3, -1, 0), \quad b = 4, \quad A = \begin{bmatrix} 2 & -1 & 0\\ -1 & -5 & 0\\ 0 & 0 & 1 \end{bmatrix}
\end{aligned}
$$

It is a quadratic expression.

$$u_2 = a'X, \quad a' = (1, -1, 1), \quad X' = (x_1, x_2, x_3).$$

It is a linear form.

$$u_3 = X'AX, \quad X' = (x_1, x_2, x_3), \quad A = \begin{bmatrix} 1 & 0 & 0\\ 0 & 2 & 2\\ 0 & 2 & -1 \end{bmatrix}.$$

This is a quadratic form. In all these cases, the matrix of the quadratic form is written in the symmetric form. Any quadratic form in real variables will be of the form $X'AX$, where $X' = (x_1, \dots, x_k)$, $A = (a_{ij})$ and $X'AX$ is a scalar quantity or a $1 \times 1$ matrix, and hence it is equal to its transpose. That is, $X'AX = (X'AX)' = X'A'X$ and, therefore,

$$
\begin{aligned}
X'AX &= \frac{1}{2}[X'AX + X'A'X] = X'\Big(\frac{A + A'}{2}\Big)X = X'BX,\\
B &= \frac{1}{2}(A + A') = B'
\end{aligned}
$$

and hence the result.

When we have a sample from a normal population, we can derive many interesting and useful results. For a full discussion of quadratic forms and bilinear forms in real random variables, see the books [12] and [13]. We will need only two main results on quadratic forms, one is the chi-squaredness of quadratic forms and the other is the independence of two quadratic forms.

**Result 10.13** (Chi-squaredness of quadratic forms). *Let $x_1,\dots,x_p$ be iid random variables from a normal population $N(0,\sigma^2)$. Let $y = X'AX$ be a quadratic form, where $X' = (x_1,\dots,x_p)$ and $A = (a_{ij}) = A'$ be a matrix of constants. Then the necessary and sufficient condition for $u = \frac{y}{\sigma^2}$ to be chi-square distributed with r degrees of freedom or $u \sim \chi_r^2$ is that A is idempotent and of rank r.*

**Proof.** For any real symmetric matrix $A$, there exists an orthonormal matrix $Q$, $QQ' = I$, $Q'Q = I$, such that $Q'AQ = D = \operatorname{diag}(\lambda_1,\dots,\lambda_p)$ where $\lambda_1,\dots,\lambda_p$ are the eigenvalues of $A$. Hence by making the transformation $Y = Q'X$ we have

$$X'AX = Y'DY = \lambda_1 y_1^2 + \cdots + \lambda_p y_p^2. \tag{10.52}$$

Also the orthonormal transformation, being linear in $X$, will still have $y_j$, $j = 1,\dots,p$ independently distributed as $N(0,\sigma^2)$. Hence $\frac{y_j^2}{\sigma^2} \sim \chi_1^2$, $j = 1,\dots,p$. If $A$ is idempotent of rank $r$ then $r$ of the $\lambda_j$'s are unities and the remaining ones are zeros, and hence $\frac{X'AX}{\sigma^2} = \frac{1}{\sigma^2}(y_1^2 + \cdots + y_r^2) \sim \chi_r^2$. Thus, if$A$ is idempotent of rank $r$ then $u \sim \chi_r^2$. For proving the converse, we assume that $\frac{X'AX}{\sigma^2} \sim \chi_r^2$ and show that then $A$ is idempotent and of rank $r$. The mgf of a chi-square with $r$ degrees of freedom is $(1-2t)^{-\frac{r}{2}}$ for $1-2t > 0$. But, from (10.41) each $\frac{y_j^2}{\sigma^2}$ is $\chi_1^2$, with mgf $(1-2t)^{-\frac{1}{2}}$ with $1-2t > 0$, for $j = 1,\dots,p$ and mutually independently distributed. Further, $\lambda_j y_j^2/\sigma^2$ has the mgf $(1 - 2\lambda_j t)^{-\frac{1}{2}}$ with $1 - 2\lambda_j t > 0$, $j = 1,\dots,p$. Then from (10.52) and the $\chi_r^2$, we have the identity

$$\prod_{j=1}^{p}(1 - 2\lambda_j t)^{-\frac{1}{2}} \equiv (1 - 2t)^{-\frac{r}{2}}. \tag{10.53}$$

Take the natural logarithm on both sides of (10.53), expand and equate the coefficients of $(2t), (2t)^2, \dots$ we obtain the following:

$$\sum_{j=1}^{p} \lambda_j^m = r, \quad m = 1, 2, \dots$$

The only solution for the above sequence of equations is that $r$ of the $\lambda_j$'s are unities and the remaining ones are zeros. This condition, together with the property that our matrix $A$ is real symmetric will guarantee that $A$ is idempotent of rank $r$. This establishes the result.

**Note 10.10.** Observe that if a matrix has eigenvalues 1's and zeros that does not mean that the matrix is idempotent. For example, take a triangular matrix with diagonal elements zeros and ones. But this property, together with real symmetry will guarantee idempotency.

**Note 10.11.** Result 10.13 can be extended to a dependent case also. When $x_1,\dots,x_p$ have a joint normal distribution $X \sim N_p(O,\Sigma)$, $\Sigma = \Sigma' > 0$, $X' = (x_1,\dots,x_p)$, a corresponding result can be obtained. Make the transformation $Y = \Sigma^{-\frac{1}{2}}X$ then the problem will reduce to the situation in Result 10.13. If $X \sim N_p(\mu,\Sigma)$, $\mu \neq O$ then also a corresponding result can be obtained but in this case the chi-square will be a non-central chi-square. Even if $X$ is a singular normal, that is, $|\Sigma| = 0$ then also a corresponding result can be obtained. For such details, see [12].

As a consequence of Result 10.13, we have the following result.

**Result 10.14.** *Let $x_1,\dots,x_n$ be iid variables distributed as $N(\mu,\sigma^2)$. Let*

$$u = \sum_{j=1}^{n} \frac{(x_j - \bar{x})^2}{\sigma^2} = \sum_{j=1}^{n} \frac{[(x_j - \mu) - (\bar{x} - \mu)]^2}{\sigma^2} = \sum_{j=1}^{n} (y_j - \bar{y})^2$$

*where $y_j = \frac{x_j-\mu}{\sigma}$. Then*

$$u = \sum_{j=1}^{n} (y_j - \bar{y})^2, \quad y_j \sim N(0,1) \tag{10.54}$$

*and*

$$u = \chi^2_{n-1}. \tag{10.55}$$

**Proof.** Writing $x_j - \bar{x} = (x_j - \mu) - (\bar{x} - \mu)$ and then taking $y_j = \frac{x_j-\mu}{\sigma}$, we have the representation in (10.54). But (10.54) is a quadratic form of the type $Y'AY$ where $A = I - \frac{1}{n}LL'$, $L' = (1,1,\dots,1)$ which is idempotent of rank $n-1$. Then from Result 10.13 the result follows.

Another basic result, which is needed for testing hypotheses in model building situations, design of experiments, analysis of variance, regression problems, etc. is the result on independence of quadratic forms. This will be stated next.

**Result 10.15** (Independence of two quadratic forms). *Let $x_1,\dots,x_p$ be iid variables following a $N(0,\sigma^2)$ distribution. [This is also the same as saying $X \sim N_p(O,\sigma^2 I)$, $X' = (x_1,\dots,x_p)$ where $I$ is the identity matrix and $\sigma^2 > 0$ is a scalar quantity.] Let $u = X'AX$, $A = A'$ and $v = X'BX$, $B = B'$ be two quadratic forms in $X$. Then these quadratic forms $u$ and $v$ are independently distributed if and only if (iff) $AB = O$.*

**Proof.** Since $A$ and $B$ are real symmetric matrices, from Result 10.13 we have the representations

$$u = X'AX = \lambda_1 y_1^2 + \cdots + \lambda_p y_p^2 \tag{10.56}$$

and

$$v = \nu_1 y_1^2 + \cdots + \nu_p y_p^2 \tag{10.57}$$

where $\lambda_1, \dots, \lambda_p$ are the eigenvalues of $A$; $\nu_1, \dots, \nu_p$ are the eigenvalues of $B$, and $y_j \sim N(0,1), j = 1, \dots, p$ and mutually independently distributed. Let us assume that $AB = O$. Then due to symmetry, we have

$$AB = O = O' = (AB)' = B'A' = BA \quad \Rightarrow \quad AB = BA$$

which means that $A$ and $B$ commute. This commutativity and symmetry will guarantee that there exists an orthonormal matrix $Q$, $QQ' = I$, $Q'Q = I$ such that both $A$ and $B$ are reduced to their diagonal forms by the same $Q$. That is,

$$\begin{aligned} O = AB \quad &\Rightarrow \quad Q'AQQ'BQ = D_1D_2, \\ D_1 = \mathrm{diag}(\lambda_1, \dots, \lambda_p), &\quad D_2 = \mathrm{diag}(\nu_1, \dots, \nu_p). \end{aligned}$$

But $D_1D_2 = O$ means that whenever a $\lambda_j \neq 0$ the corresponding $\nu_j = 0$ and vice versa. In other words, all terms in $u$ and $v$ are mathematically separated. Once a set of statistically independent variables are mathematically separated then the two sets are statistically independent also. Hence $u$ and $v$ are independently distributed. This is the sufficiency part of the proof. For proving the converse, the "necessary" part, we assume that $u$ and $v$ are independently distributed. We can use this property and the representations in (10.56) and (10.57). By retracing the steps in the "sufficiency" part, we cannot prove the "necessary" part. There are many incorrect proofs of this part in the literature. The correct proof is a little more lengthy and makes use of a number of properties of matrices, and hence we will not give here. The students may refer to Mathai and Provost [12].

**Remark 10.2.** One consequence of the above result with respect to a simple random sample from a normal population is the following: Let $x_1, \dots, x_n$ be iid $N(\mu, \sigma^2)$ variables. Let $u = \frac{1}{\sigma^2}\sum_{j=1}^n (x_j - \bar{x})^2$ and $v = \frac{n}{\sigma^2}(\bar{x} - \mu)^2$. Taking $y_j = \frac{(x_j - \mu)}{\sigma}$, we have $y_j \sim N(0,1)$, $j = 1, \dots, n$ and iid. Then

$$\begin{aligned} u &= \sum_{j=1}^m (y_j - \bar{y})^2 = Y'AY, \quad A = I - \frac{1}{n}LL', \quad L' = (1, \dots, 1) \\ v &= (\bar{y})^2 = Y'BY, \quad B = \frac{1}{n}LL', \quad Y' = (y_1, \dots, y_n). \end{aligned}$$

Observe that $AB = O \Rightarrow u$ and $v$ are independently distributed, thereby one has the independence of the sample variance and the square of the sample mean when the sample comes from a $N(\mu, \sigma^2)$ population. One can extend this result to the independence of the sample variance and the sample mean when the sample is from a $N(\mu, \sigma^2)$ population.

## Exercises 10.5

**10.5.1.** Let the $p \times 1$ vector $X$ have a mean value vector $\mu$ and positive definite covariance matrix $\Sigma$, that is, $E(X) = \mu$, $\mathrm{Cov}(X) = \Sigma = \Sigma' > 0$. Show that $Y = \Sigma^{-\frac{1}{2}}X \Rightarrow E(Y) = \Sigma^{-\frac{1}{2}}\mu$, $\mathrm{Cov}(X) = I$ and for $Z = Y - \Sigma^{-\frac{1}{2}}\mu$, $E(X) = O, \mathrm{Cov}(Z) = I$.

**10.5.2.** Consider the quadratic form $Q(X) = X'AX$ and $Y$ and $Z$ as defined in Exercise 10.5.1. Then show that

$$\begin{aligned} Q(X) &= X'AX = Y'\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}Y \\ &= (Z + \Sigma^{-\frac{1}{2}}\mu)'\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}(Z + \Sigma^{-\frac{1}{2}}\mu). \end{aligned} \tag{10.58}$$

**10.5.3.** Let $P$ be an orthonormal matrix which will diagonalize the symmetric matrix of Exercise 10.5.2, $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$, into the form

$$P'\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}P = \mathrm{diag}(\lambda_1, \dots, \lambda_p), \quad PP = I, \quad P'P = I$$

where $\lambda_1, \dots, \lambda_p$ are the eigenvalues of $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$. Then show that $Q(X)$, the quadratic form, has the following representations:

$$\begin{aligned} Q(X) = X'AX &= \sum_{j=1}^{p} \lambda_j(u_j + b_j)^2, \quad A = A', \quad \mu \neq O \\ &= \sum_{j=1}^{p} \lambda_j u_j^2, \quad A = A', \quad \mu = O \end{aligned} \tag{10.59}$$

where $b' = (b_1, \dots, b_p) = \mu'\Sigma^{-\frac{1}{2}}P$.

**10.5.4.** Illustrate the representation in Exercise 10.5.3 for $Q(X) = 2x_1^2 + 3x_2^2 - 2x_1x_2$ and $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$.

**10.5.5. Singular case.** Let the $p \times 1$ vector $X$ have the mean value $E(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$ of rank $r \le p$. Since $\Sigma$ here is at least positive semi-definite, we have a representation $\Sigma = BB'$ where $B$ is $p \times r$ of rank $r$. Then one can write $X = \mu + BY$ with $E(Y) = O$ and $\mathrm{Cov}(Y) = I$. Show that any quadratic form $Q(X) = X'AX$ has the representation

$$\begin{aligned} Q(X) &= X'AX = (\mu + BY)'A(\mu + BY) \\ &= \mu'A\mu + 2Y'B'A\mu + Y'B'ABY \quad \text{for } A = A'. \end{aligned} \tag{10.60}$$

Obtain a representation for $Q(X)$, corresponding to the one in Exercise 10.5.3 for the singular case.

**10.5.6.** Let the $p \times 1$ vector $X$, $X' = (x_1, \dots, x_p)$, be distributed as a multivariate normal of the following type:

$$f(X) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{p/2}} \mathrm{e}^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)}$$

where $-\infty < x_j < \infty$, $-\infty < \mu_j < \infty$, $\Sigma > 0$, $\mu' = (\mu_1, \dots, \mu_p)$. Show that for this non-singular normal the mgf is given by

$$M_X(T) = \mathrm{e}^{T'\mu + \frac{1}{2}T'\Sigma T} \tag{10.61}$$

where $T' = (t_1, \dots, t_p)$ is a parametric vector. (Hint: $M_X(T) = E[\mathrm{e}^{T'X}]$.)

**10.5.7.** Taking the mgf in Exercise 10.5.6 as the mgf for both the non-singular case $\Sigma > O$ and the singular case $\Sigma \geq O$ show by calculating the mgf, or otherwise, that an arbitrary linear function $y = a'X$, $a' = (a_1, \dots, a_p)$, $X' = (x_1, \dots, x_p)$ has a univariate normal distribution.

**10.5.8.** If an arbitrary linear function $y = a'X$, $a' = (a_1, \dots, a_p)$, $X' = (x_1, \dots, x_p)$, has a univariate normal distribution, for all constant vectors $a$, then show that the $p \times 1$ vector $X$ has a multivariate normal distribution of the type determined by (10.61) and in the non-singular case, has the density as in Exercise 10.5.6.

**10.5.9.** Let the $p \times 1$ vector $X$ have a singular normal distribution (which also includes the non-singular case). Let the covariance matrix $\Sigma$ of $X$ be such that $\Sigma = BB'$ where $B$ is a $p \times q$, $q \geq p$ matrix of rank $r \leq p$. Let $Q(X) = X'AX$. Show that the mgf of $Q = Q(X)$ has the representation

$$\begin{aligned} M_Q(t) &= \left\{\prod_{j=1}^{r}(1-2t\lambda_j)^{-\frac{1}{2}}\right\} \exp\left\{\alpha t + 2t^2 \sum_{j=1}^{r} b_j^2 (1-2t\lambda_j)^{-1}\right\}, \quad \mu \neq O \\ &= \prod_{j=1}^{r}(1-2t\lambda_j)^{-\frac{1}{2}}, \quad \mu = O \end{aligned} \tag{10.62}$$

where the $\lambda_j$'s are the eigenvalues of $B'AB$, $b' = (b_1, \dots, b_r) = \mu'A'BP$, $\alpha = \mu'A\mu$.

**10.5.10.** Let $X \sim N_p(O, I)$ and $X'X = X'A_1X + X'A_2X$, $A_1 = A_1'$, $A_2 = A_2'$, where $A_1$ is idempotent of rank $r < p$. Then show that $X'A_1X \sim \chi_r^2$, $X'A_2X \sim \chi_{p-r}^2$ and $X'A_1X$ and $X'A_2X$ are independently distributed. [This result can also be extended when we have the representation $I = A_1 + \dots + A_k$ and this will help to split the total variation to sum of individual variations due to different components in practical situations such as analysis of variance problems.]

## 10.6 Order statistics

In a large number of practical situations, the items of interest may be largest value or the smallest value of a set of observations. If we are watching the flood in the local river, then the daily water level in the river is not that important but the highest water level is most important or water levels over a threshold value are all important. If you are watching the grain storage in a silo or water storage in a water reservoir serving a city, over the years, then both the highest level and lowest levels are very important.

If you are running an insurance firm then the maximum damage due to vehicular collision, largest number of accidents, largest number of thefts of properties are all very important. The theory of order statistics deals with such largest values or smallest values or the $r$-th largest values, etc. Since numbers are simply numbers and there is not much to study there, we will be studying some random variables corresponding to such ordered observations.

Let $x_1, \dots, x_n$ be a simple random sample of size $n$ from some population. Then $\{x_1, \dots, x_n\}$ is a collection of random variables. Consider one set of observations on $\{x_1, \dots, x_n\}$. For example, let $x$ be the waiting time for a particular bus at a local bus stop. Assume that this bus never comes earlier than the scheduled time but it can only be on time or late. On 5 (here $n = 5$) randomly selected occasions let the waiting times be $3, 10, 15, 0, 2$, time being measured in minutes. If we write these observations in ascending order of their magnitudes, then we have

$$0 < 2 < 3 < 10 < 15 \tag{i}$$

Again, suppose that another 5 (same $n = 5$) occasions are checked. The waiting times may be $3, 5, 8, 5, 10$. If we order these observations, then we have

$$3 < 5 \le 5 < 8 < 10 \tag{ii}$$

If we keep on taking such 5 observations each then each such set of 5 observations can be ordered as in (i) and (ii). There will be a set of observations which will be the smallest in each set, a set of observations which will be the second smallest and so on, and finally there will be a set of observation corresponding to the largest in each set. Now think of the set of smallest observations as coming from a random variable denoted by $x_{n:1}$, the set of second smallest numbers coming from a random variable $x_{n:2}$, etc. and finally the set of largest observations as coming from the random variable $x_{n:n}$. Thus, symbolically we may write

$$x_{n:1} \le x_{n:2} \le \cdots \le x_{n:n} \tag{10.63}$$

Since these are random variables, defined on the whole real line $(-\infty, \infty)$, there is no meaning of the statement that the variables are ordered or one variable is less than another variable. What it means is that if we have an observed sample, then the $n$ observations can be ordered. Once they are ordered, then the smallest will be the observation on $x_{n:1}$, the second smallest will be the observation on $x_{n:2}$ and so on, and the largest will be an observation on $x_{n:n}$. From the ordering in (i) and (ii) note that the number 3 is the smallest in (ii) whereas it is the 3rd smallest in (i). Thus, for example, every observation on $x_{n:1}$ need not be smaller than every observation on $x_{n:2}$ but for every observed sample of size $n$ we have one observation each corresponding to $x_{n:r}$ for $r = 1, 2, \dots, n$. Now, we will have some formal definitions.

Let us consider a continuous population. Let the iid variables $x_1, \dots, x_n$ come from a population with density function $f(x)$ and distribution function $F(x)$ (cumulative density). How can we compute the density function or distribution function of $x_{n:r}$, the $r$-th largest variable or the $r$-th order statistic? For example, how can we compute the density of the smallest order statistic?

### 10.6.1 Density of the smallest order statistic $x_{n:1}$

We are considering continuous random variables here. We may use the argument that if the smallest is bigger than a number $y$ then all observations on the variables $x_1, \dots, x_n$ must be bigger than $y$. Since the variables are iid, the required probability will be a product. Therefore,

$$\Pr\{x_{n:1} > y\} = \Pr\{x_1 > y\}\Pr\{x_2 > y\} \cdots \Pr\{x_n > y\} = [\Pr\{x_j > y\}]^n$$

since the variables are iid. But

$$\Pr\{x_j > y\} = 1 - \Pr\{x_j \le y\} = 1 - F(y) \tag{10.64}$$

where $F(y)$ is the distribution function of $x$ evaluated at the point $y$. But the left side is 1– the distribution function of $x_{x:1}$, denoted by $1 - F_{(1)}(y)$. Therefore, the density function of $x_{n:1}$, denoted by $f_{(1)}(y)$, is given by

$$f_{(1)}(y)\big|_{y=x_{n:1}} = -\frac{\mathrm{d}}{\mathrm{d}y}[1 - F_{(1)}(y)]\bigg|_{y=x_{n:1}} = -\frac{\mathrm{d}}{\mathrm{d}y}[1 - F(y)]^n\bigg|_{y=x_{n:1}}.$$

$$f_{(1)}(x_{n:1}) = n[1 - F(x_{n:1})]^{n-1} f(x_{n:1}), \quad -\infty < x_{n:1} < \infty. \tag{10.65}$$

Here, $f(x_{n:1})$ indicates the population density evaluated at the observed point of $x_{n:1}$ and $F(x_{n:1})$ means the population distribution function evaluated at the observed $x_{n:1}$.

### 10.6.2 Density of the largest order statistic $x_{n:n}$

Again we are considering continuous random variables. Here, we may use the argument that if the largest of the observations is less than or equal to $y$ then every observation must be $\le y$. This statement, translated in terms of the random variables is the following:

$$\begin{aligned}\Pr\{x_{n:n} \le y\} &= \Pr\{x_1 \le y\} \cdots \Pr\{x_n \le y\} \\ &= [\Pr\{x_j \le y\}]^n = [F(y)]^n.\end{aligned}$$

Hence the density of $x_{n:n}$, denoted by $f_{(n)}(\cdot)$, is given by

$$\begin{aligned}f_{(n)}(x_{n:n}) &= \frac{\mathrm{d}}{\mathrm{d}y}\Pr\{x_{n:n} \le y\}\bigg|_{y=x_{n:n}} = n[F(y)]^{n-1} f(y)\big|_{y=x_{n:n}} \\ &= n[F(x_{n:n})]^{n-1} f(x_{n:n}).\end{aligned} \tag{10.66}$$

### 10.6.3 The density of the $r$-th order statistic $x_{n:r}$

Here also one can use an argument similar to the one in Sections 10.6.1 and 10.6.2. But it will be easier to use the following argument. Think of the subdivision of the $x$-axis into the following intervals: $(-\infty, x_{n:r})$, $(x_{n:r}, x_{n:r} + \Delta x_{n:r})$, $(x_{n:r} + \Delta x_{n:r}, \infty)$, where $\Delta x_{n:r}$ is a small increment in $x_{n:r}$. When we say that an observation is the $r$-th largest that means that $r-1$ are below that or in the interval $(-\infty, x_{n:r})$, one is in the interval $(x_{n:r}, x_{n:r} + \Delta x_{n:r})$ and $n-r$ observations are in the interval $(x_{n:r} + \Delta x_{n:r}, \infty)$. Let $p_1, p_2$ and $p_3$ be the respective probabilities. These probabilities can be computed from the population density. Note that $p_i > 0$, $i = 1, 2, 3$ and $p_1 + p_2 + p_3 = 1$ because we have $n \geq 1$ observations. Then from the multinomial probability law the density of $x_{n:r}$, denoted by $f_{(r)}(x_{n:r})$, is given by the following multinomial probability law:

$$f_{(r)}(x_{n:r})\mathrm{d}x_{n:r} = \lim_{\Delta x_{n:r} \to 0} \left[ \frac{n!}{(r-1)!(n-r)!} p_1^{r-1} p_2^1 p_3^{n-r} \right].$$

But

$$\begin{aligned}
p_1 &= \Pr\{-\infty < x_j \leq x_{n:r}\} = F(x_{n:r}) \\
\lim_{\Delta x_{n:r} \to 0} p_2 &= \lim_{\Delta x_{n:r} \to 0} \Pr\{x_{n:r} \leq x_j \leq x_{n:r} + \Delta x_{n:r}\} = f(x_{n:r})\mathrm{d}x_{n:r} \\
\lim_{\Delta x_{n:r} \to 0} p_3 &= \lim_{\Delta x_{n:r} \to 0} \Pr\{x_{n:r} + \Delta x_{n:r} \leq x_j < \infty\} = 1 - F(x_{n:r}).
\end{aligned}$$

Substituting these values, we get

$$\begin{aligned}
f_{(r)}(x_{n:r}) &= \frac{n!}{(r-1)!(n-r)!} [F(x_{n:r})]^{r-1} [1 - F(x_{n:r})]^{n-r} f(x_{n:r}) \\
&= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)} [F(x_{n:r})]^{r-1} \\
&\quad \times [1 - F(x_{n:r})]^{n-r} f(x_{n:r}).
\end{aligned} \tag{10.67}$$

Note that $f_{(r)}(\cdot)$ is the density of the $r$-th order statistic, $f(x_{n:r})$ is the population density evaluated at $x_{n:r}$ and $F(x_{n:r})$ is the population distribution function evaluated at $x_{n:r}$. Note that the above procedure is the most convenient one when we want to evaluate the joint density of any number of order statistics, that is, divide the real line into intervals accordingly and then use the multinomial probability law to evaluate the joint density.

**Example 10.11.** Evaluate the densities of (1) the largest order statistic, (2) the smallest order statistic, (3) the $r$-th order statistic, when the population is (i) uniform over $[0, 1]$, (ii) exponential with parameter $\theta$.

**Solution 10.11.** (i) Let the population be uniform over $[0, 1]$. Then the population density and distribution function are the following:

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{elsewhere}; \end{cases} \qquad F(x) = \begin{cases} x, & 0 \le x \le 1 \\ 1, & x \ge 1. \end{cases}$$

From (10.65), the density of the smallest order statistic is given by

$$f_{(1)}(y) = n[1-y]^{n-1}, \quad 0 \le y \le 1,\ y = x_{n:1}$$

and zero elsewhere. From (10.66), the density of the largest order statistic is given by

$$f_{(n)}(y) = ny^{n-1}, \quad 0 \le y \le 1,\ y = x_{n:n}$$

and zero elsewhere. From (10.67), the density of the $r$-th order statistic is given by

$$f_{(r)}(y) = \frac{n!}{(r-1)!(n-r)!} y^{r-1}(1-y)^{n-r}, \quad 0 \le y \le 1,\ y = x_{n:r}.$$

(ii) When the population is exponential the density and distribution function are the following:

$$f(x) = \begin{cases} \frac{\mathrm{e}^{-\frac{x}{\theta}}}{\theta} & \\ 0, & \text{elsewhere}; \end{cases} \qquad F(x) = \begin{cases} 0, & -\infty < x < 0 \\ 1-\mathrm{e}^{-x/\theta}, & 0 \le x < \infty. \end{cases}$$

Hence the density for the largest order statistic is given by

$$f_{(n)}(y) = n[1-\mathrm{e}^{-y/\theta}]^{n-1}\mathrm{e}^{-y/\theta}\frac{1}{\theta}, \quad y = x_{n:n}.$$

The density for the smallest order statistic is given by

$$f_{(1)}(y) = n[\mathrm{e}^{-y/\theta}]^{n-1}\mathrm{e}^{-y/\theta}\frac{1}{\theta} = \frac{n}{\theta}\mathrm{e}^{-ny/\theta}, \quad y = x_{n:1}.$$

It is interesting to note that the density of the smallest order statistic in this case is again an exponential density with parameter $\theta/n$ or if the original population density is taken as $f(x) = \theta \mathrm{e}^{-\theta x}$, $x \ge 0$, $\theta > 0$ then the density of the smallest order statistic is the same with $\theta$ replaced by $n\theta$. This, in fact, is a property which can be used to characterize or uniquely determine the exponential density.

**Example 10.12.** A traveler taking a commuter train every morning for five days every week has to wait in a queue for buying the ticket. If the waiting time is exponentially distributed with the expected waiting time 4 minutes, then what is the probability that for any given week (1) the shortest waiting time is less than one minute, (2) the longest waiting time is more than 10 minutes, time being measured in minutes?

**Solution 10.12.** From the given information, the population density is of the form:

$$f(t) = \frac{1}{4}\mathrm{e}^{-t/4}, \quad t \ge 0$$

and zero elsewhere, and hence the distribution function will be of the form $F(x) = 1 - \mathrm{e}^{-x/4}$ for $x \geq 0$. Then the density for the smallest order statistic is of the form:

$$f_{(1)} = n[1 - F(y)]^{n-1} f(y) = \frac{5}{4}\mathrm{e}^{-5y/4}, \quad y \geq 0,\ y = x_{n:n}.$$

The probability that we need is $\Pr\{y \leq 1\}$. That is,

$$\Pr\{y \leq 1\} = \int_0^1 \frac{5}{4}\mathrm{e}^{-5y/4}\mathrm{d}y = 1 - \mathrm{e}^{-5/4}.$$

Similarly, the density for the largest order statistic is

$$f_{(n)}(y) = n[F(y)]^{n-1} f(y) = 5[1 - \mathrm{e}^{-y/4}]^4 \frac{1}{4}\mathrm{e}^{-y/4}.$$

The probability that we need is $\Pr\{y \geq 10\}$. That is,

$$\begin{aligned}\Pr\{y \geq 10\} &= \frac{5}{4}\int_{10}^{\infty} [1 - \mathrm{e}^{-y/4}]^4 \mathrm{e}^{-y/4}\mathrm{d}y \\ &= 5\int_{2.5}^{\infty} [1 - 4\mathrm{e}^{-u} + 6\mathrm{e}^{-2u} - 4\mathrm{e}^{-3u} + \mathrm{e}^{-4u}]\mathrm{e}^{-u}\mathrm{d}u \\ &= 5\mathrm{e}^{-2.5}\left[1 - 2\mathrm{e}^{-2.5} + 2\mathrm{e}^{-5} - \mathrm{e}^{-7.5} + \frac{1}{5}\mathrm{e}^{-10}\right].\end{aligned}$$

**Example 10.13.** For the $r$-th order statistic $x_{n:r}$ show that the density can be transformed to a type-1 beta density.

**Solution 10.13.** Let the population density and distribution function be denoted by $f(x)$ and $F(x)$ respectively and let the density for the $r$-th order statistic be denoted by $f_{(r)}(y)$, $y = x_{n:r}$. Then it is given by

$$f_{(r)}(y) = \frac{n!}{(r-1)!(n-r)!}[F(y)]^{r-1}[1 - F(y)]^{n-r} f(y), \quad y = x_{n:r}.$$

Let $u = F(y) \Rightarrow \mathrm{d}u = \mathrm{d}F(y) = f(y)\mathrm{d}y$. Then the density of $u$, denoted by $g(u)$, is given by

$$\begin{aligned}g(u) &= \frac{n!}{(r-1)!(n-r)!}u^{r-1}(1-u)^{n-r}, \quad 0 \leq u \leq 1 \\ &= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n+1-r)}u^{r-1}(1-u)^{n+1-r-1}, \quad 0 \leq u \leq 1\end{aligned}$$

and zero elsewhere, which is a type-1 beta density with the parameters $r$ and $n + 1 - r$.

**Remark 10.3.** From the structure of the density function of the $r$-th order statistic, it may be noted that the distribution function $F(x)$ (cumulative density) of the original population is involved. Hence if the population is gamma, generalized gamma, Raleigh, Maxwell–Boltzmann, general pathway density, etc. then $F(x)$ may not have a simple explicit analytic form or it may go into incomplete gamma functions, and then the explicit evaluation of the moments, etc. of the $r$-th order statistic for $r = 1, 2, \dots, n$ may not be possible. Even if the population is standard normal still the distribution function $F(x)$ is not available analytically. One may have to evaluate in terms of incomplete gamma function or go for numerical evaluations. But the transformation $u = F(x)$, as in Example 10.13, will reduce the density in terms of type-1 beta density or in terms of type-1 Dirichlet density in the joint distribution of several order statistics. Hence properties of order statistics can be studied easily in all situations where one can write $x = F^{-1}(u)$ explicitly. This is possible in some cases, for example, for uniform and exponential situations. If $x$ is uniform over $[a, b]$, then

$$F(x) = \frac{x}{b-a} \quad \text{and} \quad u = F(x) \quad \Rightarrow \quad x = (b-a)u. \tag{10.68}$$

If $x$ is exponential with parameter $\theta$, then

$$F(x) = 1 - \mathrm{e}^{-x/\theta} \quad \text{and} \quad u = F(x) \quad \Rightarrow \quad x = -\theta \ln(1-u). \tag{10.69}$$

**Example 10.14.** Evaluate the $h$-th moment of the $r$-th order statistic coming from a sample of size $n$ and from a uniform population over $[0, 1]$.

**Solution 10.14.** When the population is uniform over $[0, 1]$, then the distribution function

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \int_{0}^{x} \mathrm{d}t = x, \quad 0 \le x \le 1.$$

Hence the cumulative density at the point $x_{n:r}$ is $u = x_{n:r}$, and the density for $x_{n:r}$ is given by

$$f_{(r)}(y) = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n+1-r)} y^{r-1}(1-y)^{n+1-r-1}, \quad y = x_{n:r}, \ 0 \le y \le 1$$

and zero elsewhere. Hence the $h$-th moment of the $r$-th order statistic $x_{n:r}$ is given by

$$\begin{aligned} E[x_{n:r}]^h = E[y^h] &= \int_0^1 y^h f_{(r)}(y)\mathrm{d}y \\ &= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n+1-r)} \int_0^1 y^{r+h-1}(1-y)^{n+1-r-1}\mathrm{d}y \\ &= \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n+1-r)} \frac{\Gamma(r+h)\Gamma(n+1-r)}{\Gamma(n+1+h)} \\ &= \frac{\Gamma(n+1)}{\Gamma(n+1+h)} \frac{\Gamma(r+h)}{\Gamma(r)}, \quad \Re(h) > -r \end{aligned}$$

or $h > -r$ if $h$ is real.

### 10.6.4 Joint density of the $r$-th and $s$-th order statistics $x_{n:r}$ and $x_{n:s}$

Let us consider two order statistics $x_{n:r}$, the $r$-th and $x_{n:s}$, the $s$-th order statistics for $r < s$. Then divide the real axis into intervals $(-\infty, x_{n:r})$, $(x_{n:r}, x_{n:r} + \Delta x_{n:r})$, $(x_{n:r} + \Delta x_{n:r}, x_{n:s})$, $(x_{n:s}, x_{n:s} + \Delta x_{n:s})$, $(x_{n:s} + \Delta x_{n:s}, \infty)$. Proceed exactly as in the derivation of the density of the $r$-th order statistic. Let the joint density of $x_{n:r}$ and $x_{x:s}$ be denoted by $f(y,z)$, $y = x_{n:r}$, $z = x_{n:s}$. Then we have

$$\begin{aligned} f(y,z)\mathrm{d}y \wedge \mathrm{d}z = {} & \frac{n!}{(r-1)!1!(s-r-1)!1!(n-s)!}[F(y)]^{r-1} \\ & \times [F(z)-F(y)]^{s-r-1}[1-F(z)]^{n-s} \\ & \times f(y)f(z)\mathrm{d}y \wedge \mathrm{d}z, y = x_{n:r}, \quad z = x_{n:s}, \end{aligned} \tag{10.70}$$

for $-\infty < y < z < \infty$ and zero elsewhere. As in Example 10.13, if we make the transformation $u = F(y)$, $v = F(z)$ then the joint density of $u$ and $v$, denoted by $g(u,v)$ is given by the following:

$$\begin{aligned} g(u,v) = {} & \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(s-r)\Gamma(n+1-s)} u^{r-1}(v-u)^{s-r-1} \\ & \times (1-v)^{n+1-s-1}, \quad 0 \le u < v \le 1 \end{aligned} \tag{10.71}$$

and zero elsewhere. If we make a further transformation $u = u_1$, $v - u = u_2$, then the joint density of $u$ and $v$ is changed to $g_1(u_1,u_2)$, given by

$$g_1(u_1,u_2) = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(s-r)\Gamma(n+1-s)} u_1^{r-1}u_2^{s-r-1}(1-u_1-u_2)^{n+1-s-1} \tag{10.72}$$

for $0 \le u_i \le 1$, $i = 1,2$, $0 \le u_1 + u_2 \le 1$ and $g_1(u_1,u_2) = 0$ elsewhere, which is a Dirichlet density. If we make a further simplification $r = r_1$, $s - r = r_2$, $n+1-s = n+1-r_1-r_2 = r_3$, then $r_1 + r_2 + r_3 = n+1$. Thus the density $g_1(u_1,u_2)$ becomes

$$g_1(u_1,u_2) = \frac{\Gamma(r_1+r_2+r_3)}{\Gamma(r_1)\Gamma(r_2)\Gamma(r_3)} u_1^{r_1-1}u_2^{r_2-1}(1-u_1-u_2)^{r_3-1} \tag{10.73}$$

for $0 \le u_j \le 1$, $0 \le u_1 + u_2 \le 1$. Then (10.73) is in the usual format of a type-1 Dirichlet density. Now we can extend to consider the joint density of the $y_1 = r_1$-th, $y_2 = (r_1 + r_2)$-th, ..., $y_p = (r_1 + \cdots + r_p)$-th order statistics. Let the joint density be denoted by $g(y_1,\dots,y_p)$. Then we have

$$\begin{aligned} g(y_1,\dots,y_p) = {} & \frac{\Gamma(n+1)}{\Gamma(r_1)\Gamma(r_2)\cdots\Gamma(r_{p+1})}[F(y_1)]^{r_1-1} \\ & \times [F(y_2)-F(y_1)]^{r_2-1}\cdots[1-F(y_p)]^{r_{p+1}-1} \\ & \times f(y_1)\cdots f(y_p). \end{aligned} \tag{10.74}$$

Now, make the transformations $v_1 = F(y_1)$, $v_2 = F(y_2),\ldots,v_p = F(y_p)$ and then make the transformation $u_1 = v_1$, $u_2 = v_2 - v_1$, ..., $u_p = v_p - v_{p-1}$. Then the joint density of $u_1,\ldots,u_p$ will reduce to a type-1 Dirichlet density of the form:

$$\begin{aligned} g_1(u_1,\ldots,u_p) &= \frac{\Gamma(n+1)}{\Gamma(r_1)\Gamma(r_2)\cdots\Gamma(r_{p+1})}u_1^{r_1-1} \\ &\quad\times u_2^{r_2-1}\cdots u_p^{r_p-1}(1-u_1-\cdots-u_p)^{r_{p+1}-1} \end{aligned} \tag{10.75}$$

for $r_1 + \cdots + r_{p+1} = n+1$, $0 \le u_j \le 1$, $j = 1,\ldots,p$, $0 \le u_1 + \cdots + u_p \le 1$.

**Example 10.15.** Construct the joint density of $x_{n:1}$, the smallest order statistic, and $x_{n:n}$ the largest order statistic for a sample of size $n$ from a population with distribution function $F(x)$. Show that it is a density. Then construct the density of the range = largest order statistic minus the smallest order statistic, when the population is uniform over $[0,1]$.

**Solution 10.15.** In the general formula in (10.70) put $r = 1$ and $s = n$ to obtain the joint density of the largest order statistic $y_2 = x_{n:n}$ and the smallest order statistic $y_1 = x_{n:1}$. Let the joint density be denoted by $f(y_1,y_2)$. Then

$$\begin{aligned} f(y_1,y_2) &= \frac{n!}{0!(n-1-1)!0!}[F(y_1)]^0 \\ &\quad\times[F(y_2)-F(y_1)]^{n-1-1}[1-F(y_2)]^0 f(y_1)f(y_2) \\ &= n(n-1)[F(y_2)-F(y_1)]^{n-2}f(y_1)f(y_2). \end{aligned} \tag{10.76}$$

Let $u = F(y_1)$, $v = F(y_2)$. Then the joint density of $u$ and $v$, denoted by $g(u,v)$, is given by

$$g(u,v) = n(n-1)[v-u]^{n-2}, \quad 0 \le u \le v \le 1,\ n \ge 2. \tag{10.77}$$

Integrating out over $y_1$ and $y_2$ in (10.76) is equivalent to integrating out $u$ and $v$ in (10.77). Let us compute the total integral, denoted by $q$.

$$q = \int_u\int_v g(u,v)\mathrm{d}u\wedge\mathrm{d}v = n(n-1)\int_{v=0}^1\left[\int_{u=0}^v (v-u)^{n-2}\mathrm{d}u\right]\mathrm{d}v.$$

Put $z = \frac{u}{v}$ and take out $v$. Then

$$\begin{aligned} q &= n(n-1)\int_{v=0}^1 v^{n-1}\left[\int_{z=0}^1(1-z)^{n-2}\mathrm{d}z\right]\mathrm{d}v \\ &= \frac{n(n-1)}{n-1}\int_{v=0}^1 v^{n-1}\mathrm{d}v = 1 \quad \text{for } n \ge 2. \end{aligned}$$

Hence $f(y_1,y_2)$ is a density since it is non-negative with total integral unity. When the population is uniform over $[0,1]$ then the joint density of $y_1$ and $y_2$ is that in (10.77) with $y_1 = u$ and $y_2 = v$. Let $w = y_2 - y_1$ and $y = y_2$. Then the Jacobian is 1 and the joint density of $w$ and $y$, denoted by $g_1(w,y)$, is given by

$$g_1(w,y) = n(n-1)w^{n-2}, \quad 0 \le w \le y \le 1$$

and zero elsewhere. Integrating out $y$, $w < y < 1$, the marginal density of $w$, denoted by $h(w)$, is the following:

$$h(w) = n(n-1)w^{n-2}\int_w^1 \mathrm{d}y = n(n-1)w^{n-2}(1-w), \quad 0 \le w \le 1$$

and zero elsewhere. It is type-1 beta density with parameters $(n-1, 2)$.

One can make another interesting observation. Consider a simple random sample $x_1, \dots, x_n$ from a population with density $f(x)$. Suppose that we consider the joint density of all order statistics $y_1 = x_{n:1}$, $y_2 = x_{n:2}$, ..., $y_n = x_{n:n}$. If all variables are involved, then the collection of the original variables $\{x_1, \dots, x_n\}$ and the collection of all order statistics $\{y_1 = x_{n:1}, \dots, y_n = x_{n:n}\}$ are one and the same. That is, $\{x_1, \dots, x_n\} = \{y_1, \dots, y_n\}$. The only difference is that in the set $\{x_1, \dots, x_n\}$ the variables are free to vary. But in the set $\{y_1, \dots, y_n\}$ the variables are ordered, that is, $y_1 \le y_2 \le \dots \le y_n$. Given a set of variables $x_1, \dots, x_n$ how many such ordered sets are possible? This number is the number of permutations, which is $n!$. Hence $n!$ ordered sets are possible. If integration is to be done for computing some probabilities, then in the set $\{y_1 = x_{n:1}, \dots, y_n = x_{n:n}\}$ the integration is to be done as follows: If the original variables have non-zero density in $[a, b]$, then $y_1$ goes from $a$ to $y_2$. Then $y_2$ goes from $a$ to $y_3$ and so on or if we are integrating from the other end then $y_n$ goes from $y_{n-1}$ to $b$, $y_{n-1}$ from $y_{n-2}$ to $b$ and so on. The idea will be clear from the following example.

**Example 10.16.** Let $x_1, \dots, x_n$ be iid as exponential with parameter $\theta$ or with density function

$$f(x) = \frac{\mathrm{e}^{-x/\theta}}{\theta}, \quad x \ge 0,\ \theta > 0$$

and zero elsewhere. Compute the joint density of all order statistics $y_1 = x_{n:1}, \dots, y_n = x_{n:n}$.

**Solution 10.16.** The joint density of $x_1, \dots, x_n$, denoted by $f(x_1, \dots, x_n)$, is given by

$$f(x_1, \dots, x_n) = \frac{1}{\theta^n}\exp\left\{-\frac{1}{\theta}(x_1 + \dots + x_n)\right\}, \quad 0 \le x_j < \infty,\ \theta > 0.$$

The joint density of $y_1, \dots, y_n$, denoted by $g(y_1, \dots, y_n)$, is then

$$g(y_1, \dots, y_n) = \frac{n!}{\theta^n}\exp\left\{-\frac{1}{\theta}(y_1 + \dots + y_n)\right\}, \quad 0 \le y_1 \le y_2 \le \dots \le y_n < \infty,$$

and $\theta > 0$. Let us verify that it is a density. It is a non-negative function and let us compute the total integral. For convenience, we can integrate out starting from $y_n$. Integration over $y_n$ is given by

$$\frac{n!}{\theta^n}\int_{y_n=y_{n-1}}^{\infty}\exp\left\{-\frac{y_n}{\theta}\right\}\mathrm{d}y_n=\frac{n!}{\theta^{n-1}}\exp\left\{-\frac{y_{n-1}}{\theta}\right\}.$$

In the joint density there is already a $y_{n-1}$ sitting in the exponent. Then for the next integral the coefficient of $y_{n-1}$ is 2. Then the integral over $y_{n-1}$ gives

$$\frac{n!}{\theta^{n-1}}\int_{y_{n-2}}^{\infty}\exp\left\{-\frac{2y_{n-1}}{\theta}\right\}\mathrm{d}y_{n-1}=\frac{n!}{\theta^{n-2}(2)}\exp\left\{-\frac{2y_{n-2}}{\theta}\right\}.$$

Proceeding like this the last integral over $y_1$ is the following:

$$\frac{n!}{\theta(2)(3)\cdots(n-1)}\int_{0}^{\infty}\exp\left\{-\frac{ny_1}{\theta}\right\}\mathrm{d}y_1=\frac{n!\theta}{\theta(2)(3)\cdots(n)}=1.$$

This shows that it is a joint density.

## Exercises 10.6

**10.6.1.** Let $x_1,\dots,x_n$ be independently distributed but not identically distributed. Let the distribution function of $x_j$ be $F_j(x_j)$, $j=1,\dots,n$. Consider observations on $x_1,\dots,x_n$ and ordering them in ascending values. Let $y_1=x_{n:1}$ be the smallest order statistic and $y_2=x_{n:n}$ be the largest order statistic here. By using the ideas in Sections 10.6.1 and 10.6.2 derive the densities of $y_1$ and $y_2$.

**10.6.2.** Let $x$ have the density $f(x)=\frac{5x^4}{2^5}$, $0\le x\le 2$ and $f(x)=0$ elsewhere. Consider a simple random sample of size $n$ from this population. Construct the densities of (1) the smallest order statistic, (2) the largest order statistic, (3) the 5-th order statistic, $n>5$.

**10.6.3.** For the problem in Exercise 10.6.2 evaluate the probability that $x_{n:1}\le 0.5$, $x_{n:4}\ge 1.5$, $x_{n:5}\le 1$ for $n=10$.

**10.6.4.** (1) Compute $c$ so that $f(x)=c(1-x)^3$, $0\le x\le 1$ and $f(x)=0$ elsewhere, is a density; (2) Repeat Exercise 10.6.2 if the density in (1) here is the density there.

**10.6.5.** Repeat Exercise 10.6.3 for the density in Exercise 10.6.4.

**10.6.6.** During the raining season of June–July at Palai it is found that the duration of a rain shower, $t$, is exponentially distributed with expected duration 5 minutes, time being measured in minutes. If 5 such showers are randomly selected, what is the probability that (1) the shortest shower lasted for more than 2 minutes; (2) the longest shower lasted for less than 5 minutes; (3) the longest shower lasted for more than 10 minutes?

**10.6.7.** For the Exercise in 10.6.6, what is the probability that the second shortest shower lasted for more than 2 minutes?

**10.6.8.** Let $x_{n:1}, x_{n:2}, \dots, x_{n:n}$ be all the order statistics for a simple random sample of size $n$ coming from a population with density function $f(x)$ and distribution function $F(x)$. Construct the joint density of all these order statistics $x_{n:1}, \dots, x_{n:n}$.

**10.6.9.** The annual family income $x$ of households in a city is found to be distributed according to the density $f(x) = \frac{c}{x^2}$, $1 \leq x \leq 10$, $x$ being measured in Rs 10 000 units which means, for example, $x = 2$ means of Rs 20 000. (1) Compute $c$; (2) If a simple random sample of 6 households is taken, what is the probability that the largest of the household income is more than Rs 80 000 or $x_{n:n} > 8$?

**10.6.10.** If $x_1, \dots, x_n$ are iid Poisson distributed with parameter $\lambda = 2$, construct the probability function of (1) the smallest order statistic, (2) the largest order statistic.

**10.6.11.** Let $x_1, \dots, x_n$ be iid as uniform over $[0, \theta]$. Compute the joint density of the order statistics $y_1 = x_{n:1}, \dots, y_n = x_{n:n}$ and verify that it is a density.

# 11 Estimation

## 11.1 Introduction

Statistical inference part consists of mainly estimation, testing of statistical hypotheses and model building. In Chapter 10, we developed some tools which will help in statistical inference problems. Inference about a statistical population is usually made by observing a representative sample from that population and then making some decisions based on some statistical procedures. Inference may be of the following nature: Suppose that a farmer is planting corn by preparing the soil as suggested by the local agricultural scientist. The item of interest to the farmer is whether the yield per plot of land is going to be bigger than the usual yield that the farmer is getting by the traditional method of planting. Then the hypothesis is that the expected yield under the new method is greater than or equal to the expected yield under the traditional method of planting. This will be something like a hypothesis $E(x_1) \geq E(x_2)$ where $x_1$ is the yield under the new method and $x_2$ is the yield under the traditional method. The variables $x_1$ and $x_2$ are the populations here, having their own distributions. If $x_1$ and $x_2$ are independently gamma distributed with the parameters $(\alpha_1, \beta_1)$ and $(\alpha_2, \beta_2)$, then $E(x_1) = \alpha_1\beta_1$ and $E(x_2) = \alpha_2\beta_2$. Then our hypothesis is that $\alpha_1\beta_1 \geq \alpha_2\beta_2$. How do we carry out a statistical test of the above hypothesis? We have to conduct experiments under the traditional method of planting and under the new method of planting and collect a sample of observations under both of these methods. This is the first step. Hence the very basic aspect of inference is a sampling procedure or to take a representative sample, someway representing the whole population. There are many types of samples and sampling procedures. We have already discussed one type of sample, in Chapter 10, namely a simple random sample. There are other types of samples such as multistage samples, stratified samples, proportional samples and so on, and the corresponding sampling plans are there. We will only consider inference based on simple random samples here.

The student may be wondering whether it is essential to go for samples and why not look into the whole population itself and then take a decision. Sometimes it is possible to check the whole population and then take a decision. Suppose that a firm, such as HMT, has produced only 10 printing units. Suppose that the claimed weight per unit is 10 tons. It is not difficult to check this claim by weighing all the 10 units. Sometimes, even if we have only a finite number of units in a population it may not be possible to check each and every unit and come up with a decision. Suppose that a car manufacturer has produced 25 expensive cars, such as a new model of Ferrari. The manufacturer's claim is that in case of frontal collision the damage to the car will be less than 10%. One cannot test this claim by a 100% checking or 100% testing because checking each item involves destroying the car itself. Suppose that the manufacturer of a new brand of electric bulb says that the expected life time of the new brand is

https://doi.org/10.1515/9783110562545-011

1 000 hours. Here, the life-time $x$ may be exponentially distributed with mean value $\theta$. Then the claim is that $\theta = 1000$. One cannot do a hundred percent testing of this claim because each observation can be made only by burning the bulb until it is burnt out. There will not be anything left for sale. Hence one has to go for a representative sample, study the sample and test the claims by using this sample observations only.

As another example, suppose that you want to test the claim that in the local Meenachil River the chance (probability) of flood (water level going above a threshold value) in any given year is less than or equal to 0.2. Here, even if you wish to do a 100% checking it is not possible because you do not have all the past records and you do not have access to the future data on flood. Hence one has to go for some sort of a sampling scheme and collect some data and based on this data make inference or decisions. Sometimes it may not be worth going for a 100% survey. Suppose that the claim is that the average annual income per household in Kerala is less than or equal to Rs 10 000. Suppose that there are about one lakh households. It is not wise of you to go for a 100% survey to collect this single item about the expected annual income because it is not worth the money spent. Thus, even if it is possible to conduct a 100% survey, it may be unwise and may not be worth it to do so.

The first topic in statistical inference that we are going to consider is statistical estimation problem. The idea is to use a representative sample from a given population and then estimate some aspects of the population. If the population under consideration is waiting time, $t$, at a given bus stop and if $t$ is exponentially distributed with expected waiting time some unknown quantity $\theta$ then our aim here may be to come up with an estimate of the expected waiting time, or to come up with an estimate of this unknown quantity $\theta$. Here, we want to estimate a parameter in a specified population. Suppose that in the same situation of waiting time being exponentially distributed, someone is interested to estimate the probability that the waiting time there is greater than or equal to 5 minutes, time being measured in minutes. Then this probability, $p$, is given by the following:

$$p = \Pr\{t \geq 5\} = \int_5^\infty \frac{1}{\theta} e^{-\frac{t}{\theta}} dt = e^{-\frac{5}{\theta}} = p(\theta).$$

Here, this is a function of the unknown parameter $\theta$ and hence we can look upon this problem as the problem of estimating a probability or estimating a function of a given parameter.

As another example, we can look at the growth of a child, measured in terms of height. The growth pattern may be more or less a straight line model until the age of 10 to 11. Then all of a sudden, the child shoots up and attains the maximum height by 11 or 12. There is a change-point in the growth pattern, somewhere between 10 and 11 years of age. One may be interested to estimate this change point. As another example, consider $x$, the amount of gold in every ton of gold bearing rock in a mountain range. The density function of $x$ itself is unknown and we would like to estimate the density function itself. Such an estimation problem is called “density estimation”. Thus, so far, we have considered the problem of estimating a parameter in a well-defined distribution, estimation of a certain probability, estimation of a parametric function, estimation of a density itself and the estimation of a change-point in a model building situation. Then there are problems where one may be interested in testing whether two variables $x$ and $y$ are statistically independently distributed (this is called testing independence), one may want to check and see whether the occurrence of a defective item in a production process is a random phenomenon or occurring according to some specific pattern (this is called testing for randomness), one may want to measure the association or affinity between two variables, etc. All these problems involve some estimation of some characteristics before a testing procedure can be devised.

## 11.2 Parametric estimation

Out of the various situations that we considered above, one situation is that we have a well-defined population or populations and well-defined parameter or parameters therein. The estimation process involves estimating either such a parameter or parameters or a parametric function.

**Definition 11.1** (Parametric estimation)**.** Estimation of a parameter or parameters or functions of parameters, coming from well-defined populations, is known as a parametric estimation problem.

Even here, there are different situations. We may want to estimate the expected life-time of a machine part, the life-time may be exponentially distributed with expected value $\theta$ hours. In this case, we want to estimate this $\theta$. Theoretically this expected life time could be any real non-negative number, or in this case $0 < \theta < \infty$.

**Definition 11.2** (Parameter space)**.** The set of all possible values a well-defined parameter or parameters can take is called the parameter space, usually denoted by $\Omega$.

For example, in an exponential density there is usually one parameter $\theta$ where $0 < \theta < \infty$. Here, the parameter space $\Omega = \{\theta \mid 0 < \theta < \infty\}$. In a normal population $N(\mu, \sigma^2)$, there are two parameters $\mu$ and $\sigma^2$ and here the parameter space is $\Omega = \{(\mu, \sigma^2) \mid -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty\}$. Consider a gamma density with shape parameter $\alpha$, scale parameter $\beta$ and location parameter $\gamma$. Then the parameter space is given by $\Omega = \{(\alpha, \beta, \gamma) \mid 0 < \alpha < \infty,\ 0 < \beta < \infty,\ -\infty < \gamma < \infty\}$. Thus the set of all possible values the parameters can take will constitute the parameter space, whatever be the number of parameters involved.

You see advertisements by commercial outfits, such as a toothpaste manufacturer claiming that their toothpaste will reduce cavities, (whatever that may be), by 21 to 46%. If $p$ is the true percentage reduction, the manufacturer is giving an interval saying that this true unknown quantity is somewhere on the interval $[0.21, 0.46]$. A single number is not given but an interval is given. A saree shop may be advertising saying that in that shop the customer will save between Rs 500 and Rs 1 000 on the average per saree, compared to other shops. If the expected saving is $\theta$, then the claim is that this unknown $\theta$ is somewhere on the interval $[500, 1000]$. An investment company may be advertising that the expected return in this company will be 10% for any money invested with that firm. If the true percentage return is $p$, then the estimate given by that firm is that $\hat{p} = 0.10$, where $\hat{p}$ is the estimated value of $p$. A frequent traveler is claiming that the expected waiting time at a particular bus stop for a particular bus is 10 minutes. If $\theta$ is the unknown expected waiting time, then the traveler is claiming that an estimate for this $\theta$ is 10 or $\hat{\theta} = 10$.

We have looked into several situations where sometime points or single units are given as estimates for a given parameter and sometime an interval is given, claiming that the unknown parameter is somewhere on this interval. The first type of parameter estimates, where single points are given as estimates for given parameters, are called *point estimates* and the procedure is called *point estimation procedure*. When an interval is given saying that the unknown parameter is somewhere on that interval, then such an estimate will be called an *interval estimate* and the procedure is called an *interval estimation procedure* or the procedure for setting up *confidence intervals*. In the case of interval estimation problem if two parameters are involved, then a region will be given saying that the two-dimensional point is somewhere on this region. Similarly, if $k$ parameters are involved, then $k$-dimensional Euclidean region will be given, which will be then called a *confidence region* for $k = 2, 3, \dots$. These will be defined properly in the next chapter.

We will start with point estimation procedure in this chapter. But before discussing point estimation procedures we will look into some desirable properties that we would like to have for point estimates. We have already defined what is meant by a statistic in Chapter 10.

**Definition 11.3** (Point estimators). If a statistic is used to estimate a parameter or a parametric function then that statistic is called a point estimator for that parametric function and a specific value assumed by that estimator is called a point estimate. An estimator is a random variable and an estimate is a value assumed by that random variable.

A desirable property of a point estimator is the property known as unbiasedness. The property comes from the desire of having the estimator coinciding with the parametric function, on the average, in the long run. For example, if someone is throwing a dart at a small circle on a dart board then an estimate of the probability of hit is available from the relative frequency of hit. Repeat the experiment 100 times and if 45 hits are there then the relative frequency of hit or the average is $\frac{45}{100} = 0.45$. Repeat the experiment another 100 times. If 52 hits are there, then an estimate of the true probability of hit $p$ is the estimate $\hat{p} = \frac{52}{100} = 0.52$. We would like to have this average or the relative frequency agreeing with the true value $p$ in the long run in the sense when the same experiment of throwing 100 times is repeated such batches of 100 throws infinitely many times.

**Definition 11.4** (Unbiasedness of estimators). Let $T = T(x_1, \dots, x_n)$ be an estimator of a parametric function $g(\theta)$, where $\theta \in \Omega$ some parameter space. If $E[T] = g(\theta)$ for all values of $\theta \in \Omega$, then we say that the estimator $T$ is unbiased for $g(\theta)$ or $T$ is an unbiased estimator for $g(\theta)$.

If $E(T) = g(\theta)$ holds for some values of $\theta$ only, such as $\theta > \theta_0 = 5$, then $T$ is not unbiased for $g(\theta)$. The condition must hold for each and every value of $\theta$ in the whole parameter space $\Omega$. In Chapter 10, we have seen that when we have iid variables or simple random sample from a population with mean value $\mu$ and variance $\sigma^2$ then the sample mean $\bar{x}$ is unbiased for the population mean value $\mu$, as long as $\mu$ is finite, and $\sum_{j=1}^{n}(x_j - \bar{x})^2/(n-1)$ is unbiased for the population variance, as long as $\sigma^2 < \infty$. That is, for all values of $\mu < \infty$, whatever be the population,

$$E[\bar{x}] = \mu < \infty \tag{11.1}$$

and

$$E\left[\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1}\right] = \sigma^2 < \infty, \tag{11.2}$$

for $\sigma^2$ finite. These are some general results, irrespective of the populations.

**Remark 11.1.** Due to unbiasedness of the statistic $T = \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1}$ for the population variance $\sigma^2$, some people are tempted to define this $T$ as the sample variance. But this approach is inconsistent with the definition of population variance as $\sigma^2 = E[x - E(x)]^2$ take, for example, a discrete distribution where the probability masses $1/n$ each are distributed at the points $x_1, \dots, x_n$ then

$$E[x - E(x)]^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n} \neq \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1}. \tag{11.3}$$

Besides, variance is the square of a "distance" measure, measuring per unit scatter from a point of location, namely, $E(x)$ if we consider the population and $\bar{x}$ if we consider sample values. Thus the dividing factor should be $n$, not $n - 1$. Also, unbiasedness may not be a desirable property in many situations. Suppose that someone wants to cross a river. If he has the depths at every meter across the river, the average may be half a meter and this information is of no use to cross the river. The deepest point may be 5 meters deep.

**Example 11.1.** Compute the expected value of (1): $(-2)^x$ when $x$ is a Poisson random variable with parameter $\lambda > 0$; (2): $(-1)^x$ when $x$ is a binomial random variable with parameters $(n,p)$, $0 < p < 1$, $q = 1-p$, $p < \frac{1}{2}$, and construct unbiased estimators for $e^{-3\lambda}$ and $(1-2p)^n$, respectively, and make comments.

**Solution 11.1.** (1) For the Poisson case the probability function is

$$f_1(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad \lambda > 0,\ x = 0,1,2,\dots$$

and $f_1(x) = 0$ elsewhere. Then the expected value of $(-2)^x$ is given by the following:

$$\begin{aligned} E[(-2)^x] &= \sum_{x=0}^{\infty}(-2)^x\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(-2\lambda)^x}{x!} \\ &= e^{-\lambda}e^{-2\lambda} = e^{-3\lambda}, \quad 0 < e^{-3\lambda} < 1. \end{aligned}$$

Thus the statistic $(-2)^x$ is unbiased for $e^{-3\lambda}$ in this case. Note that for $\lambda > 0$, $0 < e^{-3\lambda} < 1$ whereas $(-2)^x$ is 1 or 2 or $\geq 1$ or a negative quantity. Hence this statistic $T(x) = (-2)^x$ is a meaningless estimator for $e^{-3\lambda}$, even though it is an unbiased estimator for $e^{-3\lambda}$.
(2) For the binomial case, the probability function is given by

$$f_2(x) = \binom{n}{x}p^x q^{n-x}, \quad 0 < p < 1,\ q = 1-p,\ n = 0,1,\dots,n$$

and $f_2(x) = 0$ elsewhere. Then the expected value of $(-1)^x$ is given by the following:

$$\begin{aligned} E((-1)^x) &= \sum_{x=0}^{n} f_2(x) \\ &= (q-p)^n = (1-2p)^n. \end{aligned}$$

But for $p < \frac{1}{2}$, $(1-2p)^n > 0$ and not equal to $\pm 1$. The estimator is $(-1)^x = 1$ or $-1$, and hence this unbiased estimator is a meaningless estimator for the parametric function $(1-2p)^n$.

**Remark 11.2.** Many such unbiased estimators for parametric functions can be constructed as in Example 11.1. Hence one should not take unbiasedness as a universally very desirable property for estimators.

**Definition 11.5** (Estimable function). Let $g(\theta)$ be a parametric function. If there exists at least one unbiased estimator for $g(\theta)$, then we say that $g(\theta)$ is estimable. If there is no $T = T(x_1,\dots,x_n)$ such that $E(T) = g(\theta)$ for all $\theta$, then we say that $g(\theta)$ is not estimable.

Suppose that we are constructing an estimator for $E(x_j) = \mu$ as a linear function of the iid variables $x_1,\dots,x_n$, then any estimator in this class of linear functions will be of the form $u = a_1x_1 + \cdots + a_nx_n$. Then estimation of $\mu$ in the class of linear functions will imply that

$$E(u) = \mu = a_1\mu + \cdots + a_n\mu \quad \Rightarrow \quad a_1 + \cdots + a_n = 1.$$

That is, any unbiased estimator for $\mu$ in this linear class must satisfy the condition $a_1 + \cdots + a_n = 1$, which will then be the estimability condition here.

**Example 11.2.** Construct an unbiased estimator for (1) $p^2$ where $p$ is the expected binomial proportion; (2) $\lambda^2$ where $\lambda$ is the mean value in a Poisson population; (3) $\theta^2$ where $\theta$ is the mean value in an exponential population.

**Solution 11.2.** (1) Let $x$ be binomially distributed with parameters $(n,p)$, $0 < p < 1$. Then we know that

$$E[x(x-1)] = n(n-1)p^2 \quad \Rightarrow \quad E[u] = E\left[\frac{x(x-1)}{n(n-1)}\right] = p^2.$$

Hence $u = \frac{x(x-1)}{n(n-1)}$ here is an unbiased estimator for $p^2$.

(2) Let $x$ be a Poisson random variable with parameter $\lambda$. Then we know that

$$E[x(x-1)] = \lambda^2 \quad \Rightarrow \quad u = x(x-1)$$

is unbiased for $\lambda^2$ when $x$ is a Poisson variable.

(3) Let $x$ be an exponential random variable with mean value $\theta$. Then we know that the population variance is $\theta^2$. Let $x_1,\dots,x_n$ be iid variables distributed as the exponential variable $x$. For any population with finite variance $\sigma^2$, we know that

$$E\left[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}\right] = \sigma^2.$$

Hence $u = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$ is unbiased for $\theta^2$.

**Definition 11.6** (Bias in an estimator). Let $T = T(x_1,\dots,x_n)$ be an estimator for a parametric function $g(\theta)$. Let the expected value of $T$ be $g_1(\theta)$. If $T$ is unbiased, then $g_1(\theta) = g(\theta)$. Otherwise $b(\theta) = g_1(\theta) - g(\theta)$ is called the *bias* in this estimator for estimating $g(\theta)$.

Another desirable property can be defined in terms of a measure of distance. Let $x_1,\dots,x_n$ be a simple random sample from a population with parameter $\theta$ and let $g(\theta)$ be a function of $\theta$ to be estimated. Let $T_1 = T_1(x_1,\dots,x_n)$ and $T_2 = T_2(x_1,\dots,x_n)$ be two estimators for $g(\theta)$. A measure of distance between $T_1$ and $g(\theta)$ and that between $T_2$ and $g(\theta)$ are the following:

$$E[|T_i(x_1,\dots,x_n) - g(\theta)|^r]^{\frac{1}{r}}, \quad r \geq 1,\ i = 1,2. \tag{11.4}$$

If the distance between an estimator $T$ and a parametric function $g(\theta)$ is small, then $T$ is close to $g(\theta)$ or we say that it is a good estimator or smaller the distance more efficient the estimator. If the distance is zero, then that is the best estimator but distance is zero only for deterministic situations or for degenerate random variables. Hence we will adopt the terminology "smaller the distance better the estimator". A criterion called relative efficiency is defined in terms of the above distance measure in (11.4) for $r = 2$ and by using the square of the distance measure when $r = 2$. It is only for convenience that $r = 2$ is taken.

**Definition 11.7** (Relative efficiency of estimators). Let $T_1$ and $T_2$ be estimators for the same parametric function $g(\theta)$. If $E[T_1 - g(\theta)]^2 < E[T_2 - g(\theta)]^2$ for all $\theta$, then we say that $T_1$ is relatively more efficient than $T_2$ for estimating $g(\theta)$.

**Remark 11.3.** If $T_1$ and $T_2$ are unbiased for $g(\theta)$ also, then $E[T_i - g(\theta)]^2 = \mathrm{Var}(T_i)$ and then the criterion becomes the following: If $\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$, then we say that $T_1$ is relatively more efficient than $T_2$ in estimating $g(\theta)$.

**Example 11.3.** Let $x_1, \ldots, x_n$ be iid variables with $E(x_j) = \mu$ and $\mathrm{Var}(x_j) = \sigma^2 < \infty$. Consider the estimates (a) $u_1 = 2x_1 - x_2$, $u_2 = \frac{1}{2}(x_1 + x_2)$; (b) $u_1 = x_1 + x_2$, $u_2 = x_1 - x_2$; (c) $u_1 = 2x_1 - x_2$, $u_2 = x_1 + x_2$. Which is more efficient in estimating $\mu$?

**Solution 11.3.** (a) Let us compute the squared distances from $\mu$. Let the distances be denoted by $d_1$ and $d_2$, respectively. Then

$$\begin{aligned} d_1^2 &= E[u_1 - \mu]^2 = E[2x_1 - x_2 - \mu]^2 = E[2(x_1 - \mu) - (x_2 - \mu)]^2 \\ &= E[4(x_1 - \mu)^2 + (x_2 - \mu)^2 - 4(x_1 - \mu)(x_2 - \mu)] \\ &= 4\sigma^2 + \sigma^2 - 0 = 5\sigma^2 \end{aligned}$$

since the covariance is zero. Similarly,

$$\begin{aligned} d_2^2 &= E[u_2 - \mu]^2 = E\left[\frac{1}{2}(x_1 + x_2) - \mu\right]^2 = \frac{1}{4}E[(x_1 - \mu) + (x_2 - \mu)]^2 \\ &= \frac{1}{4}E[(x_1 - \mu)^2 + (x_2 - \mu)^2 + 2(x_1 - \mu)(x_2 - \mu)] \\ &= \frac{1}{4}[\sigma^2 + \sigma^2 + 0] = \frac{1}{2}\sigma^2. \end{aligned}$$

Hence $u_2$ is relatively more efficient than $u_1$ in this case. Note that both $u_1$ and $u_2$ are unbiased for $\mu$ also in this case.

(b)

$$\begin{aligned} d_1^2 &= E[u_1 - \mu]^2 = E[x_1 + x_2 - \mu]^2 = E[(x_1 - \mu) + (x_2 - \mu) + \mu]^2 \\ &= E[(x_1 - \mu)^2] + E[(x_2 - \mu)^2] + \mu^2 + 2E[(x_1 - \mu)(x_2 - \mu)] \\ &\quad + 2\mu E[(x_1 - \mu)] + 2\mu E[(x_2 - \mu)] \end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 + \sigma^2 + \mu^2 + 0 = 2\sigma^2 + \mu^2.\\
d_2^2 &= E[u_2 - \mu]^2 = E[x_1 - x_2 - \mu]^2 = E\big[(x_1 - \mu) - (x_2 - \mu) - \mu\big]^2\\
&= \sigma^2 + \sigma^2 + \mu^2 - 0 = 2\sigma^2 + \mu^2.
\end{aligned}
$$

Here, both $u_1$ and $u_2$ are equally efficient in estimating $\mu$ and both are not unbiased also.

(c)

$$d_1^2 = E[u_1 - \mu]^2 = E[2x_1 - x_2 - \mu]^2 = 5\sigma^2.$$

This is already done above. Note that $u_1$ is unbiased for $\mu$ also here.

$$d_2^2 = E[u_2 - \mu]^2 = E[x_1 + x_2 - \mu]^2 = 2\sigma^2 + \mu^2.$$

This is computed above. Note that $u_2$ is not unbiased for $\mu$. If $5\sigma^2 > 2\sigma^2 + \mu^2$ or $3\sigma^2 > \mu^2$, then $u_2$, even though not unbiased, is relatively more efficient in estimating $\mu$.

The smaller the distance of an estimator $T$ from a parametric function $g(\theta)$ better the estimator $T$ for estimating $g(\theta)$. If $T$ is unbiased also then it is equivalent to saying: smaller the variance better the estimator. Now, if we consider the class of unbiased estimators then there may be an estimator having the least variance. Then that estimator will be a very good estimator. Then, combining the two properties of unbiasedness and relative efficiency one can come up with a desirable estimator called *minimum variance unbiased estimator or MVUE*. We will discuss this aspect later on.

Observe that all the desirable properties of point estimators that we have described so far, namely unbiasedness and relative efficiency, are fixed sample size properties or the sample size $n$ remains the same. But we may want to look at estimators if we take more and more sample values or when the sample size $n$ goes to infinity. Then such a property can be called "large sample property", compared to fixed sample size or "small sample properties". If we are estimating a parametric function $g(\theta)$ and if $T = T(x_1, \dots, x_n)$ is an estimator for $g(\theta)$, then it is desirable to have the probability of having $|T - g(\theta)| < \epsilon$, where $\epsilon$ is a very small positive quantity, is one or nearly one for every sample size $n$, that is, if we have one variable $x_1$ or two variables $x_1$, $x_2$, etc. But this may not be possible. Then the next best thing is to have this property holding when $n \to \infty$. This means that the probability that $T$ goes to $g(\theta)$ goes to one when $n$ goes to infinity or

$$\Pr\{|T - g(\theta)| < \epsilon\} \to 1 \quad \text{as } n \to \infty \quad \text{or} \quad \lim_{n\to\infty} \Pr\{|T - g(\theta)| < \epsilon\} = 1.$$

**Definition 11.8** (Consistency of estimators). Let $T = T(x_1, \dots, x_n)$ be an estimator for a parametric function $g(\theta)$. Then if $\lim_{n\to\infty} \Pr\{|T_n - g(\theta)| > \epsilon\} = 0$ for every $\epsilon > 0$ or $\Pr\{T_n \to g(\theta)\} \to 1$ as $n \to \infty$ or $\lim_{n\to\infty} \Pr\{|T_n - g(\theta)| < \epsilon\} = 1$ for every $\epsilon > 0$ however small it may be, or $T_n$ converges to $g(\theta)$ in probability then $T_n$ is called a *consistent estimator* for $g(\theta)$ and the property is called *consistency of estimators*.

Hence one can say that consistency is nothing but stochastic convergence of an estimator to the parametric function to be estimated, when the sample size $n$ goes to infinity. From Chebyshev's inequality of Chapter 9, we have the following result: Let $T_n = T(x_1,\dots,x_n)$ be an estimator for $g(\theta)$. Let $E(T_n) = g_1(\theta)$ if it is not unbiased for $g(\theta)$. Then from Chebyshev's inequality,

$$\Pr\{|T_n - E(T_n)| \geq k\} \leq \frac{\operatorname{Var}(T_n)}{k^2}. \tag{11.5}$$

Then we have the following possibilities. Either $E[T_n] = g_1(\theta) = g(\theta)$ and $\operatorname{Var}(T_n) \to 0$ as $n \to \infty$ or $E(T_n) = g_1(\theta) \to g(\theta)$ and $\operatorname{Var}(T_n) \to 0$ as $n \to \infty$. In both of these situations, $T_n$ goes to $g(\theta)$ with probability 1. Hence a practical way of checking for consistency of an estimator is the following: Check for

$$E[T(x_1,\dots,x_n)] \to g(\theta) \quad \text{as } n \to \infty \tag{i}$$

$$\operatorname{Var}(T_n) \to 0 \quad \text{as } n \to \infty \tag{ii}$$

then $T_n$ is consistent for $g(\theta)$.

**Example 11.4.** Check for the consistency of (1) sample mean as an estimator for the population mean value $\mu$ for any population with finite variance $\sigma^2 < \infty$, (2) sample variance as an estimator for the population variance when the population is $N(\mu,\sigma^2)$.

**Solution 11.4.** (1) Let $x_1,\dots,x_n$ be iid variables with $E(x_j) = \mu$ and $\operatorname{Var}(x_j) = \sigma^2 < \infty$. Let the sample mean $\bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$. Then we know that $E(\bar{x}) = \mu$ and $\operatorname{Var}(\bar{x}) = \frac{\sigma^2}{n}$. Then from Chebyshev's inequality,

$$\Pr\{|\bar{x} - \mu| \geq k\} \leq \frac{\operatorname{Var}(\bar{x})}{k^2} = \frac{\sigma^2}{nk^2} \to 0 \quad \text{as } n \to \infty.$$

Therefore, $\bar{x}$ is consistent for $\mu$ whatever be the population as long as $\sigma^2 < \infty$.

(2) The sample variance

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \quad \text{and} \quad E[s^2] = \frac{(n-1)}{n}\sigma^2 \to \sigma^2 \quad \text{as } n \to \infty$$

for all populations with finite variance. For computing the variance of $s^2$, we will consider the case when the population is normal, $N(\mu,\sigma^2)$. For other populations, one has to work out the variance separately. When the population is normal one can use the properties of chi-square variables because (see Chapter 10)

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{n-1}$$

with

$$E[\chi^2_{n-1}] = (n-1) \quad \text{and} \quad \operatorname{Var}(\chi^2_{n-1}) = 2(n-1).$$

Hence

$$
\begin{aligned}
\mathrm{Var}(s^2) &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{j=1}^{n}(x_j-\bar{x})^2\right) = \frac{(\sigma^2)^2}{n^2}\mathrm{Var}\left(\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}\right)\\
&= \frac{\sigma^4}{n^2}\mathrm{Var}(\chi^2_{n-1}) = \frac{2(n-1)\sigma^4}{n^2} \to 0 \quad \text{as } n\to\infty.\\
E(s^2) &= \frac{\sigma^2}{n}E\left[\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}\right] = \frac{\sigma^2}{n}E[\chi^2_{n-1}]\\
&= \frac{\sigma^2(n-1)}{n} \to \sigma^2 \quad \text{as } n\to\infty.
\end{aligned}
$$

Hence from (i) and (ii) above, $s^2$ is consistent for the population variance $\sigma^2$. Also from (i) and (ii), the following result is obvious.

**Result 11.1.** *If $T_n$ is a consistent estimator for $g(\theta)$, then $b_nT_n$ is consistent for $bg(\theta)$ when $b_n \to b$ and $b_n^2$ goes to a finite constant when $n\to\infty$.*

**Example 11.5.** Consider a uniform population over $[0,\theta]$ and a simple random sample of size $n$ from this population. Are the largest order statistic $x_{n:n}$ and the smallest order statistic $x_{n:1}$ consistent for $\theta$?

**Solution 11.5.** See Section 10.5 in Chapter 10 for a discussion of order statistics. Let $y_1 = x_{n:1}$ and $y_2 = x_{n:n}$. The distribution function $F(x)$ of $x$, when $x$ is uniform over $[0,\theta]$, is given by

$$
F(x) = \begin{cases} 0, & -\infty < x < 0\\ \frac{x}{\theta}, & 0 \le x \le \theta\\ 1, & x \ge \theta. \end{cases}
$$

The densities for the largest order statistic, denoted by $f_{(n)}(y_2)$, and the smallest order statistic, denoted by $f_{(1)}(y_1)$, are given by the following:

$$
\begin{aligned}
f_{(n)}(y_2) &= \frac{\mathrm{d}}{\mathrm{d}x}[F(x)]^n\Big|_{x=y_2} = \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{x}{\theta}\right]^n\Big|_{x=y_2}\\
&= \frac{nx^{n-1}}{\theta^n}\Big|_{x=y_2} = \frac{ny_2^{n-1}}{\theta^n}, \quad 0 \le y_2 \le \theta,\ n \ge 1
\end{aligned}
$$

and zero elsewhere, and

$$
\begin{aligned}
f_{(1)}(y_1) &= -\frac{\mathrm{d}}{\mathrm{d}x}[1-F(x)]^n\Big|_{x=y_1} = \frac{n}{\theta}\left[1-\frac{x}{\theta}\right]^{n-1}\Big|_{x=y_1}\\
&= \frac{n}{\theta}\left[1-\frac{y_1}{\theta}\right]^{n-1}, \quad 0 \le y_1 \le \theta
\end{aligned}
$$

and zero elsewhere.

$$\begin{aligned}
E[y_2] &= \frac{n}{\theta^n}\int_0^\theta y_2(y_2)^{n-1}\mathrm{d}y_2\\
&= \frac{n}{\theta^n}\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta \to \theta \quad \text{as } n\to\infty.\\
E[y_2^2] &= \frac{n}{\theta^n}\int_0^\theta y_2^2(y_2)^{n-1}\mathrm{d}y_2 = \frac{n}{\theta^n}\frac{\theta^{n+2}}{n+2}\\
&= \frac{n}{n+2}\theta^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(y_2) = E[y^2]-[E(y_2)]^2 &= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2\\
&= \theta^2\left[\frac{n}{n+2}-\frac{n^2}{(n+1)^2}\right]\to 0 \quad \text{as } n\to\infty
\end{aligned}$$

since $\lim_{n\to\infty}\frac{n}{n+2}=1$ and $\lim_{n\to\infty}\frac{n^2}{(n+1)^2}=1$. Thus, from properties (i) and (ii) of Definition 11.8, the largest order statistic here is consistent for $\theta$. Now, let us examine $y_1 = x_{n:1}$, the smallest order statistic.

$$\begin{aligned}
E[y_1] &= \frac{n}{\theta}\int_0^\theta y_1\left[1-\frac{y_1}{\theta}\right]^{n-1}\mathrm{d}y_1\\
&= n\theta\int_0^1 u(1-u)^{n-1}\mathrm{d}u, \quad u=\frac{y_1}{\theta}\\
&= n\theta\frac{\Gamma(2)\Gamma(n)}{\Gamma(n+2)} = \frac{\theta}{n+1}\to 0 \quad \text{as } n\to\infty.\\
E[y_1^2] &= \frac{n}{\theta}\int_0^\theta y_1^2\left[1-\frac{y_1}{\theta}\right]^{n-1}\mathrm{d}y_1 = n\theta^2\int_0^1 u^2(1-u)^{n-1}\mathrm{d}u\\
&= n\theta^2\frac{\Gamma(3)\Gamma(n)}{\Gamma(n+3)} = \frac{2n\theta^2}{n(n+1)(n+2)}\to 0 \quad \text{as } n\to\infty.
\end{aligned}$$

Here, $E(y_1)\to 0$ as well as $\mathrm{Var}(y_1)\to 0$ as $n\to\infty$. Hence $y_1$ is not consistent for $\theta$.

Another desirable property of an estimator is known as *sufficiency*. Let $T = T(x_1,\dots,x_n)$ be an estimator for a parameter $\theta$. If the joint density/probability function of the sample values $x_1,\dots,x_n$, given the statistic $T$, meaning the conditional distribution of $x_1,\dots,x_n$, given the statistic $T$, is free of the parameter $\theta$, then all the information about $\theta$, that can be obtained from the sample $x_1,\dots,x_n$, is contained in $T$ because once $T$ is known the conditional distribution of $x_1,\dots,x_n$ becomes free of $\theta$. Once $T$ is known there is no need to know the whole sample space if our aim is to say something about $\theta$. For example, for $n=2$, knowing the sample means knowing every point $(x_1,x_2)$ in 2-space where the joint density/probability function of $(x_1,x_2)$ has the non-zero form. Let $T=x_1+x_2$, the sample sum. Then $x_1+x_2=$ a given quantity such as

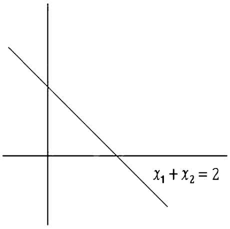


**Figure 11.1:** Reduction of data.

$x_1 + x_2 = 2$ is only a line, as shown in Figure 11.1. Knowing $T = x_1 + x_2$ means knowing only the points on this line. This enables us a large reduction of the data points. If $T$ is sufficient for $\theta$, then it means that if our aim is to say something about $\theta$ then we need to confine our attention only on the line $x_1 + x_2 = 2$ and need not worry about the whole $(x_1, x_2)$-plane. This is a very strong property.

**Definition 11.9** (Sufficient estimators). An estimator $T = T(x_1, \dots, x_n)$ is called sufficient for a parameter $\theta \epsilon \Omega$ if the conditional joint distribution of the sample values $x_1, \dots, x_n$, given $T$, is free of $\theta$ for all values of $\theta \epsilon \Omega$, where $\Omega$ is the parameter space. If $r$ statistics $T_1, \dots, T_r$ are such that the conditional distribution of $x_1, \dots, x_n$, given $T_1, \dots, T_r$, is free of the parameters $\theta_1, \dots, \theta_s$, $r \geq s$ or $r < s$, then we say that $\{T_1, \dots, T_r\}$ are *jointly sufficient* for $\{\theta_1, \dots, \theta_s\}$.

**Example 11.6.** Let $x_1, \dots, x_n$ be a simple random sample from $N(\mu, 1)$. Let $\bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$. Is $\bar{x}$ sufficient for $\mu$?

**Solution 11.6.** Joint density of $x_1, \dots, x_n$, denoted by $f(x_1, \dots, x_n)$, is given by the following since the variables are iid.

$$\begin{aligned} f(x_1, \dots, x_n) &= \frac{1}{(\sqrt{2\pi})^n} \prod_{j=1}^{n} \mathrm{e}^{-\frac{1}{2}(x_j-\mu)^2} \\ &= \frac{1}{(\sqrt{2\pi})^n} \mathrm{e}^{-\frac{1}{2}[\sum_{j=1}^{n}(x_j-\mu)^2]}. \end{aligned}$$

Note that

$$\sum_{j=1}^{n}(x_j - \mu)^2 = \sum_{j=1}^{n}(x_j - \bar{x})^2 + n(\bar{x} - \mu)^2.$$

When $x_j \sim N(\mu, 1)$, we know that $\bar{x} \sim N(\mu, \frac{1}{n})$. The density function of $\bar{x}$, denoted by $f_1(\bar{x})$, is the following:

$$f_1(\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}} \mathrm{e}^{-\frac{n}{2}(\bar{x}-\mu)^2}.$$

The joint density of $x_1, \dots, x_n$ and $\bar{x}$ is the same as the joint density of $x_1, \dots, x_n$ because no independent variable is added by incorporating $\bar{x}$, which is a linear function of the sample values $x_1, \dots, x_n$. Hence the conditional density of $x_1, \dots, x_n$, given $\bar{x}$, denoted by $g(x_1,\dots,x_n|\bar{x})$, is given by the following:

$$
\begin{aligned}
g(x_1,\dots,x_n|\bar{x}) &= \frac{\text{joint density of } x_1,\dots,x_n \text{ and } \bar{x}}{f_1(\bar{x})} \\
&= \frac{\text{joint density of } x_1,\dots,x_n}{f_1(\bar{x})} = \frac{f(x_1,\dots,x_n)}{f_1(\bar{x})} \\
&= \frac{[(\sqrt{2\pi})^{-n}\exp\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\bar{x})^2-\frac{n}{2}(\bar{x}-\mu)^2\}]}{\sqrt{n}[(\sqrt{2\pi})^{-1}\exp\{-\frac{n}{2}(\bar{x}-\mu)^2\}]} \\
&= \sqrt{n}(\sqrt{2\pi})^{-(n-1)}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\bar{x})^2\right\}
\end{aligned} \tag{11.6}
$$

which is free of $\mu$ and hence $\bar{x}$ is sufficient for $\mu$ here.

**Example 11.7.** Let $x_1,\dots,x_n$ be iid as a uniform over $[0,\theta]$. Let $y_2=\max\{x_1,\dots,x_n\}$ and $y_1=\min\{x_1,\dots,x_n\}$ (other notations for largest and smallest order statistics). Is $y_2$ sufficient for $\theta$?

**Solution 11.7.** From Section 10.5 of Chapter 10, the density for the largest order statistic is given by the following:

$$
f_{(n)}(y_2)=\frac{ny_2^{n-1}}{\theta^n},\quad 0\le y_2\le\theta,\ n\ge 1.
$$

The joint density of $x_1,\dots,x_n$, which is the same as the joint density of $x_1,\dots,x_n$ and $\max\{x_1,\dots,x_n\}=y_2$, is given by

$$
f(x_1,\dots,x_n)=\frac{1}{\theta^n},\quad 0\le x_j\le\theta,\ j=1,\dots,n.
$$

Hence the conditional density, $g(x_1,\dots,x_n|T)$, of $x_1,\dots,x_n$, given $T=y_2$, is the following:

$$
g(x_1,\dots,x_n|T=m)=\frac{f(x_1,\dots,x_n)}{f_{(n)}(y_2)}\bigg|_{T=m}=\frac{1}{\theta^n}\frac{\theta^n}{ny_2^{n-1}}\bigg|_{y_2=m}=\frac{1}{nm^{n-1}}, \tag{11.7}
$$

for $m\neq 0$, which is free of $\theta$. Hence the largest order statistic, namely, $x_{n:n}$ or $\max\{x_1,\dots,x_n\}$ is sufficient for $\theta$ here.

Observe that if $T$ is given, that is, $T=c$ where $c$ is a given constant, then $c_1T=c_1c$ is also given for $c_1\neq 0$ because $c_1T=c_1c=$ a constant. Also if we take a one to one function of $T$, say, $h(T)$, $T\leftrightarrow h(T)$ then when $T$ is given $h(T)$ is also given and vice versa. This shows that if $T$ is sufficient for $\theta$ then $h(T)$ is also sufficient for $\theta$.

**Result 11.2.** *If $T$ is a sufficient estimator for $\theta$ and if $T$ to $h(T)$ is a one to one function, then $h(T)$ is also sufficient for $\theta$.*

For example, consider Example 11.6. Let $h(T) = 2\bar{x} - 5 = \text{given} = 7$, say, then this means that $\bar{x} = \frac{7+5}{2} = 6 = \text{given}$ or vice versa.

**Example 11.8.** Let $x_1, \dots, x_n$ be iid as a Poisson with probability function

$$f(x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \quad x = 0, 1, 2, \dots, \; \lambda > 0$$

and zero elsewhere. Let $u = x_1 + \dots + x_n$. Is $u$ sufficient for $\lambda$?

**Solution 11.8.** The joint probability function of $x_1, \dots, x_n$ is given by

$$\begin{aligned} f(x_1, \dots, x_n) &= \prod_{j=1}^{n} \frac{\lambda^{x_j}}{x_j!}\mathrm{e}^{-\lambda} \\ &= \frac{\lambda^{x_1+\dots+x_n}\mathrm{e}^{-n\lambda}}{x_1!x_2!\cdots x_n!}, \quad x_j = 0, 1, 2, \dots, \; j = 1, \dots, n \end{aligned}$$

and zero elsewhere. But the moment generating function (mgf) of $x_j$ is given by

$$M_{x_j}(t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}$$

which shows that the mgf of $u = x_1 + \dots + x_n$ is a Poisson with parameter $n\lambda$, and the probability function of $u$ is the following:

$$f_1(u) = \frac{(n\lambda)^u}{u!}\mathrm{e}^{-n\lambda}, \quad u = 0, 1, \dots$$

and zero elsewhere. Note that the joint probability function of $x_1, \dots, x_n$ and $u$ is the same as the joint probability function of $x_1, \dots, x_n$ since $u$ is a function of $x_1, \dots, x_n$ and not an independent variable. The conditional probability function of $x_1, \dots, x_n$, given $u$, is then

$$\begin{aligned} f(x_1, \dots, x_n | u) &= \frac{f(x_1, \dots, x_n)}{f_1(u)} \quad \text{for } f_1(u) \neq 0 \\ &= \frac{\lambda^u \mathrm{e}^{-n\lambda}}{x_1! \cdots x_n!} \frac{u!}{(n\lambda)^u \mathrm{e}^{-n\lambda}} \\ &= \frac{m!}{x_1! \cdots x_n!}\left(\frac{1}{n}\right)^m \quad \text{for } u = m \end{aligned} \tag{11.8}$$

which is free of $\lambda$, and hence $u = x_1 + \dots + x_n$ is sufficient for $\lambda$.

**Note 11.1.** Note that for a given $u = m$ the last part of (11.8) is a multinomial probability function with $p_1 = \dots = p_n = \frac{1}{n}$ and $x_1 + \dots + x_n = m$. Hence the conditional probability function is a multinomial probability law here.

**Example 11.9.** Consider a simple random sample $x_1, \dots, x_n$ of size $n$ from a $N(\mu, \sigma^2)$ population. Show that the sample mean $\bar{x}$ and $\sum_{j=1}^n (x_j - \bar{x})^2$ are jointly sufficient for the population parameters $\mu$ and $\sigma^2$.

**Solution 11.9.** When the population is $N(\mu, \sigma^2)$, we know from Chapter 10 that $u_1 = \sum_{j=1}^n (x_j - \bar{x})^2$ and $u_2 = \bar{x}$ are independently distributed. Hence the joint density of $u_1$ and $u_2$ is the product of the marginal densities, denoted by $f_1(u_1)$ and $f_2(u_2)$, respectively. Further, we know that $u_2 = \bar{x} \sim N(\mu, \frac{\sigma^2}{n})$. Hence

$$f_2(u_2) = \frac{\sqrt{n}}{\sigma(\sqrt{2\pi})} \mathrm{e}^{-\frac{n}{2\sigma^2}} (\bar{x} - \mu)^2.$$

We also know from Chapter 10 that

$$\begin{aligned} u = \frac{u_1}{\sigma^2} &= \frac{\sum_{j=1}^n (x_j - \bar{x})^2}{\sigma^2} \sim \chi^2_{n-1} \\ \Rightarrow \quad u_1 &\sim \sigma^2 \chi^2_{n-1}. \end{aligned}$$

The density of $u$, denoted by $g(u)$, is given by the following:

$$g(u) = \frac{u^{\frac{n-1}{2}-1}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \mathrm{e}^{-\frac{u}{2}}, \quad u \geq 0$$

and free of all parameters. Put $u_1 = \sigma^2 u \Rightarrow \mathrm{d}u = \frac{1}{\sigma^2} \mathrm{d}u_1$. Hence the density of $u_1$ denoted by $f_1(u_1)$ is the following:

$$\begin{aligned} f_1(u_1)\mathrm{d}u_1 &= \frac{u_1^{\frac{n-1}{2}-1} \mathrm{e}^{-\frac{u_1}{2\sigma^2}}}{\sigma^{n-1} 2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \mathrm{d}u_1 \\ &= \frac{[\sum_{j=1}^n (x_j - \bar{x})^2]^{\frac{n-1}{2}-1}}{\sigma^{(n-1)} 2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \mathrm{e}^{-\frac{1}{2\sigma^2}} \sum_{j=1}^n (x_j - \bar{x})^2 \mathrm{d}u_1. \end{aligned}$$

The joint density of $u_1$ and $u_2$ is $f_1(u_1)f_2(u_2)$. The joint density of $x_1, \dots, x_n$ and $u_1$ and $u_2$ is the same as the joint density of $x_1, \dots, x_n$, which has the form in (11.6). Hence dividing by $f_1(u_1)f_2(u_2)$ we obtain the conditional density of $x_1, \dots, x_n, u_1, u_2$, given $u_1 = a, u_2 = b$, denoted by $g(x_1, \dots, x_n | u_1 = a, u_2 = b)$, is given by the following: Note that the exponential part in (11.6) will be canceled.

$$g(x_1, \dots, x_n | u_1 = a, u_2 = b) = \frac{\sqrt{n} 2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})}{a^{\frac{n-1}{2}-1}}$$

which is free of the parameters $\mu$ and $\sigma^2$. Hence $u_1 = \sum_{j=1}^n (x_j - \bar{x})^2$ and $u_2 = \bar{x}$ are jointly sufficient for $\mu$ and $\sigma^2$.

**Note 11.2.** Another concept associated with sufficiency is *minimal sufficiency*. There may be different sets of statistics which are jointly sufficient for a set of parameters $\theta_1, \dots, \theta_r$. The first set of statistics may contain $s_1$ statistics, the second set may contain $s_2$ statistics and so on. Which set we should take in such a situation? One may be tempted to go for the set containing the minimum number of statistics. But observe that our aim of selecting a sufficient statistic is the reduction in data. Hence that set which allows the maximal reduction of data is the preferable set and such a set which allows the maximal reduction of data is called the set of *minimal sufficient statistics*. But we will not go further into the properties of sufficient statistics.

Before concluding this section, we will introduce a factorization theorem given by Neyman and it is known as Neyman factorization theorem. The students are likely to misuse this factorization theorem when checking for sufficiency by using this factorization, and hence I have postponed the discussion to the end. The students are advised to use the definition of conditional distributions rather than going for the factorization theorem.

**Result 11.3** (Neyman factorization theorem). *Let $x_1, \dots, x_n$ be a sample coming from some population and let the joint probability/density function of $x_1, \dots, x_n$ be denoted by $f(x_1, \dots, x_n, \theta)$, where $\theta$ stands for all the parameters in the population. Let the support or the range of the variables with non-zero probability/density function be free of the parameters $\theta$. Let $T = T(x_1, \dots, x_n)$ be a statistic or an observable function of the sample values. (The sample need not be a simple random sample). If the joint probability/density function allows a factorization of the form*

$$f(x_1, \dots, x_n, \theta) = f_1(T, \theta) f_2(x_1, \dots, x_n), \tag{11.9}$$

*where $f_1(T, \theta)$ is a function of $T$ and $\theta$ alone and $f_2(x_1, \dots, x_n)$ is free of all parameters $\theta$ in the parameter space $\Omega$, then $T$ is called a sufficient statistic for $\theta$.*

The proof is beyond the scope of this book, and hence deleted. The students are advised to use the rule only in populations such as Gaussian where $-\infty < x < \infty$ or gamma where $0 \le x < \infty$ etc where the range does not depend on the parameters, and not in situations such as uniform over $[a, b]$ where $a \le x \le b$ with $a$ and $b$ being parameters.

## Exercises 11.2

**11.2.1.** Consider iid variables $x_1, \dots, x_n$ distributed as uniform over $[a, b]$, $b > a$. (1) Show that the largest order statistic is a consistent estimator for $b$; (2) Is the smallest order statistic consistent for $a$? Prove your assertion.

**11.2.2.** Let $f(x) = cx^2$, $0 \le x \le \theta$ and zero elsewhere. (1) Compute $c$ if $f(x)$ is a density; (2) compute $\Pr\{y_2 \ge 0.8\}$ and $\Pr\{y_1 \le 0.1\}$ where $y_1$ and $y_2$ are the smallest and largest order statistics for a sample of size 5 coming from the population in (1).

**11.2.3.** Let $f(x) = \frac{3}{\theta^3}x^2$, $0 \le x \le \theta$ and zero elsewhere. For a sample of size 6 from this population, compute (1) $\Pr\{y_1 \ge \frac{\theta}{4}\}$; (2) $\Pr\{y_2 \ge \frac{3}{4}\theta\}$, where $y_1$ and $y_2$ are the smallest and largest order statistics, respectively.

**11.2.4.** Consider the same population as in Exercise 11.2.3. Is the largest order statistic for a sample of size $n$ from this population (1) unbiased for $\theta$; (2) consistent for $\theta$?

**11.2.5.** Consider the function

$$f(x) = c\begin{cases} x^2, & 0 \le x \le \theta \\ \theta(2\theta - x), & \theta \le x \le 2\theta \\ 0, & \text{elsewhere.} \end{cases}$$

If $f(x)$ is a density function then (1) compute $c$; (2) the density of (a) the smallest order statistic, (b) the largest order statistic, for a sample of size 3 from this population.

**11.2.6.** For the population in Exercise 11.2.5 compute the probabilities (1) $\Pr\{y_1 \le \frac{\theta}{2}\}$; (2) $\Pr\{y_2 \ge \frac{3}{2}\theta\}$, where $y_1$ and $y_2$ are the smallest and largest order statistics, respectively.

**11.2.7.** For the order statistics in Exercise 11.2.6 is $y_1$ or $y_2$ unbiased for $\theta$; (2) consistent for $\theta$?

**11.2.8.** Show that (11.7) is a density function for $\max\{x_1, \dots, x_n\} = m$ = a given number.

**11.2.9.** Show that the conditional statement in (11.6) is in fact a density function for $\bar{x} = m$ = a given number.

**11.2.10.** Let $x_1, \dots, x_n$ be a sample from some population with all parameters denoted by $\theta$. Let $u_1 = x_1, \dots, u_n = x_n$ be $n$ statistics. Then show that $u_1, \dots, u_n$ are jointly sufficient for all the parameters $\theta$.

## 11.3 Methods of estimation

There are many methods available in the literature for getting point estimates for parameters of a given population. Some of the most frequently used methods will be described here. Usually one selects a particular method of estimation in a given situation based on the following criteria: (1) Convenience and simplicity in the use of the particular method; (2) The estimators available through that method have some desirable properties that you would like to have. Hence no particular method can be better than another method or no method is the best method universally. Each method has its own motivating factors in being selected as a method of estimation.

### 11.3.1 The method of moments

From the weak law of convergence in Chapter 9, we have seen that for iid variables a sample moment converges to the corresponding population moment as the sample size goes to infinity. Taking this property as the motivational factor the method of moments suggests to equate the sample moments to the corresponding population moments and by using such equations estimate the parameters involved in the population moments. Take as many equations as necessary. Let $x_1,\dots,x_n$ be a simple random sample. Then the sample moments are given by $\sum_{j=1}^{n}\frac{x_j^r}{n} = m_r$, $r = 1,2,\dots$. These are the sample integer moments. If $r$ here is replaced by a general $h$ where $h$ could be fractional or even complex, then we have the general sample moments. We are going to take only integer moments. Let the corresponding population moments be denoted by $\mu_r' = E[x^r]$ where $E$ denotes the expected value. Now the principle of method of moments says to equate $m_r$ and $\mu_r'$. Naturally $\mu_r'$ will contain the parameters from the population. That is, consider the equation

$$m_r = \mu_r', \quad r = 1,2,\dots \tag{11.10}$$

Take as many equations as needed from (11.10) and estimate the parameters. Note that (11.10) does not hold universally or for all values of the parameters. It holds only at the estimated points. Hence when writing the right side in (11.10) replace the parameters $\theta$ by $\hat{\theta}$, meaning the estimated point. We will use the same notation $\hat{\theta}$ to denote the estimator (random variable) as well as the estimate (a number), the use will be clear from the context.

**Example 11.10.** Estimate the parameters in (1) an exponential population, (2) gamma population, (3) in a Bernoulli population, when there is a simple random sample of size $n$. Construct the estimates when there is an observed sample $\{2,0,5\}$ in (1); $\{5,1,0,2\}$ in (2); $\{1,0,0,0,1\}$ in (3).

**Solution 11.10.** (1) Let the population density be

$$f_1(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, \quad x \ge 0,\ \theta > 0$$

and zero elsewhere. Then we know that $E(x) = \theta = \mu_1'$. The first sample moment $m_1 = \sum_{j=1}^{n}\frac{x_j}{n} = \bar{x}$. Consider the equation (11.10) for $r = 1$ at the estimated point. That is,$\bar{x} = \hat{\theta}$ or the parameter $\theta$ here is estimated by $\bar{x}$ and $\bar{x}$ is the estimator here or the estimator by using the method of moments. For the observed sample point, $\{2,0,5\}$ the value of $\bar{x} = \frac{1}{3}(2+0+5) = \frac{7}{3}$, which is the estimate by using the method of moments.

(2) Here, the population is gamma with density function, denoted by $f_2(x)$, is the following:

$$f_2(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, \quad \alpha > 0,\ \beta > 0,\ x \geq 0$$

and zero elsewhere. We know that

$$E(x) = \alpha\beta, \quad E(x^2) = \alpha(\alpha+1)\beta^2 \quad \text{and} \quad m_1 = \bar{x}, \quad m_2 = \sum_{j=1}^{n} \frac{x_j^2}{n}.$$

We may take two equations from (11.10) for $r = 1, 2$ and perhaps we may be able to estimate the two parameters $\alpha$ and $\beta$ by using these two equations. Since the equations are non-linear, there is no guarantee that two equations will enable us to solve for $\alpha$ and $\beta$. Let us try.

$$\bar{x} = \hat{\alpha}\hat{\beta} \tag{i}$$

$$\sum_{j=1}^{n} \frac{x_j^2}{n} = \hat{\alpha}(\hat{\alpha}+1)(\hat{\beta})^2. \tag{ii}$$

From (i), we have $[\bar{x}]^2 = [\hat{\alpha}\hat{\beta}]^2$ and from (ii)

$$[\hat{\alpha}\hat{\beta}]^2 + \hat{\alpha}[\hat{\beta}]^2 = \sum_{j=1}^{n} \frac{x_j^2}{n}$$

and hence

$$\hat{\alpha}[\hat{\beta}]^2 = \sum_{j=1}^{n} \frac{x_j^2}{n} - [\bar{x}]^2 = \sum_{j=1}^{n} \frac{(x_j - \bar{x})^2}{n} = s^2. \tag{iii}$$

Hence from (i) and (iii), we have

$$\hat{\beta} = \frac{s^2}{\bar{x}} \quad \text{and from (i) and (iii)} \quad \hat{\alpha} = \frac{[\bar{x}]^2}{s^2} \tag{iv}$$

where $s^2$ is the sample variance. The quantities given in (iv) are the estimators (random variables). For the observed sample point $\{5, 1, 0, 2\}$, we have

$$\bar{x} = \frac{1}{4}(5 + 1 + 0 + 2) = 2$$

and

$$\begin{aligned} s^2 &= \frac{1}{4}[(5-2)^2 + (1-2)^2 + (0-2)^2 + (2-2)^2] \\ &= \frac{14}{4} = \frac{7}{2}. \end{aligned}$$

Hence the estimates are $\hat{\beta} = \frac{7}{(2)(2)} = \frac{7}{4}$ and $\hat{\alpha} = 4(\frac{2}{7}) = \frac{8}{7}$.

(3) Here, the population is Bernoulli with the probability function

$$f_3(x) = p^x q^{1-x}, \quad x = 0, 1, \ 0 < p < 1, \ q = 1 - p$$

and zero elsewhere. Here, $E(x) = \sum_{x=0}^{1} x p^x q^{1-x} = p$. Perhaps one population moment is sufficient to estimate $p$ since there is only one parameter. The first sample moment is $m_1 = \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$, which is actually the binomial proportion because the sample sum in the Bernoulli case is the binomial random variable $x$ and then $\bar{x}$ becomes the binomial proportion. Therefore, the estimator by the method of moment, also called the *moment estimator*, is the binomial proportion $\hat{p} = \frac{x}{n}$ where $x$ is the binomial variable with $n$ number of trials. For the observed sample point $\{1, 0, 0, 0, 1\}$, the sample proportion is $\frac{1}{5}(1 + 0 + 0 + 0 + 1) = \frac{2}{5}$. Hence the estimate in this case or the *moment estimate* is $\hat{p} = \frac{2}{5}$.

### 11.3.2 The method of maximum likelihood

Another method of estimation is the method of maximum likelihood. The likelihood function $L$ is defined in Chapter 10 (Definition 10.2). It is the joint density/probability function of the sample values at an observed sample point. This procedure requires the maximization of the likelihood function, with respect to all the parameters there, and thus select the estimates. The motivation behind this method is to assign those values to the parameters where the likelihood or the joint density/probability function of getting that sample point is the maximum.

**Example 11.11.** Obtain the maximum likelihood estimators of the parameters (1) $p$ in a Bernoulli population; (2) $\theta$ in an exponential population.

**Solution 11.11.** (1) The probability function in a Bernoulli population is

$$f_1(x_j) = p^{x_j} q^{1-x_j}, \quad x_j = 0, 1, \ 0 < p < 1, \ q = 1 - p$$

and zero elsewhere. The likelihood function

$$L = \prod_{j=1}^{n} p^{x_j} q^{1-x_j} = p^{\sum_{j=1}^{n} x_j} q^{n - \sum_{j=1}^{n} x_j} = p^x q^{n-x}, \quad x = \sum_{j=1}^{n} x_j.$$

Since $L$ to $\ln L$ is a one to one function, maximization of $L$ is the same as maximization of $\ln L$ and vice versa. In this case,

$$\ln L = x \ln p + (n - x) \ln(1 - p).$$

Here, the technique of calculus is applicable. Hence, differentiate $\ln L$ with respect to $p$, equate to zero and solve for $p$ to obtain the critical points.

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L = 0 \quad \Rightarrow \quad \frac{x}{p} - \frac{(n - x)}{1 - p} = 0 \qquad \text{(i)}$$

$$\Rightarrow \quad \hat{p} = \frac{x}{n}. \tag{ii}$$

Here, $x$ is the sample sum in the iid Bernoulli case, and hence $x$ is the binomial random variable and, therefore, the only critical point here is the binomial proportion. Since (i) does not hold for all $p$ the point where it holds is denoted by $\hat{p}$. Now, consider the second-order derivative.

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ln L = \left[-\frac{x}{p^2} - \frac{(n-x)}{(1-p)^2}\right] < 0$$

for $p = \hat{p}$. Hence $\hat{p} = \frac{x}{n}$ corresponds to a maximum, and this $\hat{p}$ is the maximum likelihood estimator (MLE) of $p$ here and an observed value of $x$ will give an observed value of $\hat{p}$ and then it will be called a maximum likelihood estimate (MLE). We will use the same abbreviation MLE for the estimator as well as for the estimate. Similarly, a hat, that is, $\hat{\theta}$ will be used to denote the estimator as well as estimate of $\theta$, the difference in the use will be clear from the context.

**Note 11.3.** If we take a sample of size 1 from a binomial population then the population probability function is

$$L_1 = f(x) = \binom{n}{x} p^x q^{n-x},$$

$x = 0, 1, \dots, n$, $0 < p < 1$, $q = 1 - p$. Then the MLE of $p$ here is given by $\hat{p} = \frac{x}{n}$, the same as the one obtained in the Solution 11.11 where the population is Bernoulli and a sample of size $n$ is available. But the likelihood function in the Bernoulli case is

$$L = p^x q^{n-x}, \quad x = 0, 1, \dots, n,\ 0 < p < 1,\ q = 1 - p.$$

Observe that the number of combinations $\binom{n}{x}$ is missing here, but both $L$ and $L_1$ will lead to the same MLE for $p$.

(2) Here the density function is given by

$$f(x) = \frac{\mathrm{e}^{-\frac{x}{\theta}}}{\theta}, \quad \theta > 0,\ x \geq 0$$

and zero elsewhere. Therefore,

$$L = \prod_{j=1}^{n} f(x_j) = \frac{1}{\theta^n}\mathrm{e}^{-\sum_{j=1}^{n}\frac{x_j}{\theta}} \quad \Rightarrow \quad \ln L = -n\ln\theta - \left(\sum_{j=1}^{n} x_j\right)\left(\frac{1}{\theta}\right).$$

By using calculus, we differentiate and equate to zero to obtain the critical points.

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ln L = -\frac{n}{\theta} + \frac{(\sum_{j=1}^{n} x_j)}{\theta^2} = 0 \tag{i}$$

$$\Rightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} x_j = \bar{x} = \text{sample mean} \tag{ii}$$

Taking the second-order derivative

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln L &= \left[\frac{n}{\theta^2} - \frac{2n\bar{x}}{\theta^3}\right]\Big|_{\theta=\bar{x}} \\ &= \frac{n}{(\bar{x})^2}[1-2] = -\frac{n}{(\bar{x})^2} < 0.\end{aligned}$$

Hence the only critical point $\hat{\theta} = \bar{x}$ corresponds to a maximum or $\bar{x}$ is the MLE of $\theta$ here.

**Example 11.12.** Evaluate the MLE of $\mu$ and $\sigma^2$ in $N(\mu, \sigma^2)$ population, assuming that a simple random sample of size $n$ is available from here.

**Solution 11.12.** The likelihood function in this case is the following:

$$L = \prod_{j=1}^{n} \frac{1}{\sigma(\sqrt{2\pi})} \mathrm{e}^{-\frac{1}{2\sigma^2}(x_j-\mu)^2} = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2}.$$

Then

$$\ln L = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2.$$

Here, there are two parameters $\mu$ and $\theta = \sigma^2$. We can apply calculus here. Consider the partial derivatives of $\ln L$, with respect to $\mu$ and $\theta = \sigma^2$, and equate to zeros to obtain the critical points.

$$\begin{aligned}\frac{\partial}{\partial\mu}\ln L = 0 \quad &\Rightarrow \quad \frac{1}{\theta}\sum_{j=1}^{n}(x_j - \mu) = 0 \\ &\Rightarrow \quad \hat{\mu} = \bar{x} && \text{(i)} \\ \frac{\partial}{\partial\theta}\ln L = 0 \quad &\Rightarrow \quad -\frac{n}{2\hat{\theta}} + \frac{1}{2\hat{\theta}^2}\sum_{j=1}^{n}(x_j - \bar{x})^2 = 0 \\ &\Rightarrow \quad \hat{\theta} = \hat{\sigma^2} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 = s^2 && \text{(ii)}\end{aligned}$$

where $s^2$ is the sample variance. Hence there is only one critical point $(\mu, \sigma^2) = (\bar{x}, s^2)$. Taking all the second-order derivatives we have the following:

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\ln L &= -\frac{n}{\theta} \\ \frac{\partial^2}{\partial\theta\partial\mu}\ln L\Big|_{\mu=\bar{x}} &= -\frac{1}{\theta^2}\sum_{j=1}^{n}(x_j - \bar{x}) = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\ln L &= \left[\frac{n}{2\theta^2} - \frac{1}{\theta^3}\sum_{j=1}^{n}(x_j-\mu)^2\right]\Bigg|_{\mu=\bar{x},\theta=s^2}\\ &= \frac{n}{\hat{\theta}^2}\left[\frac{1}{2}-\frac{s^2}{\hat{\theta}}\right] = -\frac{n}{2(s^2)^2} < 0.\end{aligned}$$

Hence the matrix of second-order derivatives, evaluated at the critical point $(\mu,\theta) = (\bar{x}, s^2)$, is the following:

$$\begin{bmatrix} -\frac{n}{s^2} & 0 \\ 0 & -\frac{n}{2(s^2)^2}\end{bmatrix}$$

which is negative definite and hence the critical point $(\mu,\theta) = (\bar{x}, s^2)$ corresponds to a maximum.

**Note 11.4.** Instead of computing all second-order derivatives and checking for the negative definiteness of the matrix of second-order derivatives, evaluated at the critical point, one can also use the following argument in this case. Examine $\ln L$ for all possible values of $\mu$, $-\infty < \mu < \infty$ and $\theta = \sigma^2 > 0$. We see that $\ln L$ goes from $-\infty$ to $-\infty$ through some finite values. Hence the only critical point must correspond to a maximum because $\ln L$ does not stay at $-\infty$ all the time.

**Note 11.5.** The students are likely to ask the following question: If we had differentiated with respect to $\sigma$ instead of $\theta = \sigma^2$ would we have ended up with the same MLE $\hat{\mu} = \bar{x}$ and $\widehat{\sigma^2} = s^2$? The answer is in the affirmative and the student may verify this by treating $\sigma$ as the parameter and going through the same steps. This is coming from a general result on differentiation. For any function $g(\theta)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta) = \left[\frac{\mathrm{d}}{\mathrm{d}h(\theta)}g(\theta)\right]\left[\frac{\mathrm{d}}{\mathrm{d}\theta}h(\theta)\right] \qquad \text{(i)}$$

and hence as long as $\frac{\mathrm{d}}{\mathrm{d}\theta}h(\theta) \neq 0$ we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta) = 0 \quad \Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}h(\theta)}g(\theta) = 0 \qquad \text{(ii)}$$

and vice versa. Hence both the procedures should arrive at the same result as long as $h(\theta)$ is not a trivial function of $\theta$. The following is a general result.

**Result 11.4.** *If $\hat{\theta}$ is the MLE of $\theta$ and if $h(\theta)$ is a non-trivial function of $\theta$, that is, $\frac{\mathrm{d}}{\mathrm{d}\theta}h(\theta) \neq 0$, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.*

**Example 11.13.** Evaluate the MLE of $a$ and $b$ in a uniform population over $[a,b]$, $a < b$, assuming that a simple random sample of size $n$ is available.

**Solution 11.13.** The likelihood function in this case is the following:

$$L = \frac{1}{(b-a)^n}, \quad a \le x_j \le b,\ j = 1, \dots, n$$

and zero elsewhere. Note that the method of calculus fails here. Hence we may use other arguments. Observe that $a \le x_j \le b$ for each $j$. Maximum of $\frac{1}{(b-a)^n}$ means the minimum of $(b-a)^n$, which means the minimum possible value that can be assigned for $b-a$. When we substitute $\hat{b}$ = a function of $x_1, \dots, x_n$ and $\hat{a}$ = a function of $x_1, \dots, x_n$ we should get the minimum possible value for $\hat{b} - \hat{a}$. This means that we should assign the minimum possible value for $b$ and the maximum possible value for $a$. Since all observations are greater than or equal to $a$ the maximum possible value that can be assigned to $a$ is the smallest of the observations. Similarly, the smallest possible value that can be assigned to $b$ is the largest of the observations. Then the MLE are

$$\hat{a} = x_{n:1} = \text{smallest order statistic}, \quad \hat{b} = x_{n:n} = \text{largest order statistic} \tag{11.11}$$

**Note 11.6.** If the uniform density is written as $f(x) = \frac{1}{b-a}$, $a < x < b$ what are the MLE of $a$ and $b$? If $a < x < b$, that is, $x$ is in the open interval $(a, b)$ then there is no MLE for either $a$ or $b$ because the observations can never attain $a$ or $b$ and hence no value from the observations can be assigned to $a$ or $b$. If $a < x \le b$, then the MLE for $b$ is $\hat{b} = x_{n:n}$ = the largest order statistic but there is no MLE for $a$. If $a \le x < b$, then there is no MLE for $b$ but the MLE for $a$ is $\hat{a} = X_{n:1}$ = smallest order statistic.

**Example 11.14.** Consider an exponential density with scale parameter $\theta$ and location parameter $\gamma$. That is, the density is of the form

$$f(x) = \frac{e^{-\frac{(x-\gamma)}{\theta}}}{\theta}, \quad \theta > 0,\ x \ge \gamma$$

and zero elsewhere. Evaluate the MLE of $\theta$ and $\gamma$, if they exist.

**Solution 11.14.** The likelihood function in this case is the following:

$$L = \frac{1}{\theta^n} e^{-\sum_{j=1}^n (x_j - \gamma)}, \quad x_j \ge \gamma,\ j = 1, \dots, n,\ \theta > 0.$$

By using calculus, we can obtain the MLE of $\theta$, as done before, and $\hat{\theta} = \bar{x}$ but calculus fails to obtain the MLE of $\gamma$. We may use the following argument: For any fixed $\theta$, maximum of $L$ means the minimum possible value for $\sum_{j=1}^n (x_j - \gamma)$, which means the maximum possible value that can be assigned to $\gamma$ because all observations $x_1, \dots, x_n$ are fixed. Since each observation $x_j \ge \gamma$, the maximum possible value that can be assigned to $\gamma$ is the smallest of the observations or the MLE of $\gamma$ is $\hat{\gamma} = x_{n:1}$ = smallest order statistic.

There are several large sample properties for maximum likelihood estimators, which we will consider later on, after introducing some more methods of estimation.

**Note 11.7.** The student is likely to ask the question that if the likelihood function $L(\theta)$ has several local maxima then what is to be done? Then we take $\hat{\theta} = \text{MLE}$ as that point corresponding to the largest of the maxima or $\sup_{\theta\in\Omega} L(\theta)$ or the supremum.

Another method of estimation, which is applicable in situations where we have data which is already classified into a number of groups and we have only the information of the type that $n_1$ observations are there in the first group, $n_2$ observations are in the second group or $n_i$ observations in the $i$-th group, $i = 1, 2, \dots, k$ for a given $k$. An example may be of the following type: One may be observing the life-times of the same type of a machine component. $n_1$ may be the number of components which lasted 0 to 10 hours, $n_2$ may be the number of components which lasted between 10 and 20 hours and so on, and $n_{10}$ may be the number of components which lasted between 90 and 100 hours. Let $t$ denote the life-time then the intervals are $0 < t \leq 10, \dots, 90 < t \leq 100$. Suppose that there is a total of $n_1 + \cdots + n_k = n$ observations, say, for example, $n = 50$ components in the above illustration with $n_1 = 1$, $n_2 = 5$, $n_3 = 3$, $n_4 = 6$, $n_5 = 5$, $n_6 = 6$, $n_7 = 6$, $n_8 = 5$, $n_9 = 10$, $n_{10} = 3$. If the life-time is exponentially distributed with density $f(t) = \frac{1}{\theta}\mathrm{e}^{-\frac{t}{\theta}}$, $t \geq 0$, then the true probability of finding a life-time between 0 and 10, denoted by $p_1 = p_1(\theta)$ is given by

$$p_1 = p_1(\theta) = \int_0^{10} \frac{\mathrm{e}^{-\frac{t}{\theta}}}{\theta} = 1 - \mathrm{e}^{-\frac{10}{\theta}}.$$

In general, let $p_i(\theta)$ be the true probability of finding an observation in the $i$-th group. The actual number of observations or frequency in the $i$-th group is $n_i$ and the expected frequency is $np_i(\theta)$ where $n$ is the total frequency. We have a multinomial probability law with the true probabilities $p_1(\theta), \dots, p_k(\theta)$ with $p_1(\theta) + \cdots + p_k(\theta) = 1$ and the frequencies as $n_1, \dots, n_k$ with $n_1 + \cdots + n_k = n$. The probability function, denoted by $f(n_1, \dots, n_k)$ is the following:

$$f(n_1, \dots, n_k) = \frac{n!}{n_1! \cdots n_k!}[p_1(\theta)]^{n_1} \cdots [p_k(\theta)]^{n_k},$$

with $p_i(\theta) \geq 0$, $i = 1, \dots, k$, $p_1(\theta) + \cdots + p_k(\theta) = 1$, for all $\theta \in \Omega$ (parameter space), $n_i = 0, 1, \dots, n$, $n_1 + \cdots + n_k = n$ or it is a $(k-1)$-variate multinomial law. We have the vector of true probabilities $P' = (p_1(\theta), \dots, p_k(\theta))$ and we have the corresponding relative frequencies $Q' = (\frac{n_1}{n}, \dots, \frac{n_k}{n})$, where a prime denotes transpose. A good method of estimation of $\theta$ is to minimize a distance between the two vectors $P$ and $Q$. A measure of generalized distance between $P$ and $Q$ is Karl Pearson's $X^2$ statistic, which is given by

$$X^2 = \sum_{j=1}^{k} \left[\frac{(n_i - np_i(\theta))^2}{np_i(\theta)}\right] \approx \chi^2_{k-s-1} \tag{11.12}$$

which can be shown to be approximately a chi-square with $k - s - 1$ degrees of freedom under some conditions on $n, k, p_1(\theta), \dots, p_k(\theta)$.

**Note 11.8.** The quantity in (11.12) is only approximately a chi-square under some conditions such as $np_i(\theta) \geq 5$ for all $\theta$ and for each $i = 1, \dots, k$, and $k$ itself is $\geq 5$. When $k = 2$ note that

$$X^2 = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2}, \quad p_1 + p_2 = 1,\ n_1 + n_2 = n.$$

Let $p_1 = p$, then $p_2 = 1 - p = q$ and $n_2 = n - n_1$. Then

$$\frac{(n_2 - np_2)^2}{np_2} = \frac{(n - n_1 - n(1-p))^2}{nq} = \frac{(n_1 - np_1)^2}{nq}.$$

Then

$$X^2 = \frac{(n_1 - np)^2}{n}\left[\frac{1}{p} + \frac{1}{q}\right] = \frac{(n_1 - np)^2}{npq} \tag{11.13}$$

which is the square of a standardized binomial random variable $n_1$, having a good normal approximation for $n$ as small as 20 provided $p$ is near $\frac{1}{2}$. Hence $X^2$ will be approximately a $\chi_1^2$ for $n \geq 20$, $np \geq 5$, $nq \geq 5$. For $k = 3$ and $k = 4$ also for large values of $n$, one can have a good approximation to chi-square provided no $p_i$ is near to zero or one. Students must compute the exact probabilities by using a multinomial law and compare with chi-square approximation to realize that strict conditions on $n$ and $np_i$ are needed to have a good approximation. The conditions $np_i \geq 5$, $k \geq 5$, $i = 1, \dots, k$ will be a sufficient condition for a good approximation, in general, but when $p_i = p_i(\theta)$ is unknown then there is no way of checking this condition unless we have some information beforehand about the behavior of $p_i(\theta)$ for all $\theta$.

### 11.3.3 Method of minimum Pearson's $X^2$ statistic or minimum chi-square method

As mentioned above, this method is applicable in situations where one has categorized data in hand. The principle is to minimize $X^2$ over $\theta$ and estimate $\theta$. Here, $\theta$ represents all the parameters involved in computing the exact probabilities in the various classes. Then

$$\min_{\theta \in \Omega} \sum_{j=1}^{n} \left[\frac{(n_i - np_i(\theta))^2}{np_i(\theta)}\right] \quad \Rightarrow \quad \theta = \hat{\theta}. \tag{11.14}$$

From (11.14), observe that minimization of $X^2$ for $k = 2$ and $p_1(\theta) = \theta$ gives the estimate as $\hat{\theta} = \frac{n_1}{n}$, the binomial proportion.

This Pearson's $X^2$ statistic is quite often misused in applications. The misuse comes by taking it as a chi-square without checking for the conditions on $n$ and $p_i$'s for the approximation to hold.

Estimators of parameters in a multinomial population can be done in a similar way as in the binomial case. If $(x_1,\dots,x_k)$ has a $(k-1)$-variate multinomial law then the probability function is the following:

$$f(x_1,\dots,x_k)=\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k},$$

where $p_i\geq 0$, $i=1,\dots,k$, $p_1+\cdots+p_k=1$, $n_i=0,1,\dots,n$, $n_1+\cdots+n_k=n$. Since the sum $x_1+\cdots+x_k$ is fixed as $n$, one of the variables can be written in terms of the others, that is, for example, $x_k=n-x_1-\cdots-x_{k-1}$ and $p_k=1-p_1-\cdots-p_{k-1}$. We know that

$$E(x_i)=np_i,\quad i=1,\dots,k\quad\Rightarrow\quad \hat{p}_i=\frac{x_i}{n},\quad i=1,\dots,k \tag{11.15}$$

are the moment estimates. For computing the maximum likelihood estimates (MLE), either consider one observation $(x_1,\dots,x_k)$ from the multinomial law or iid variables from the point Bernoulli multinomial trials. Take $\ln f(x_1,\dots,x_k)$ and use calculus, that is, differentiate partially with respect to $p_i$, $i=1,\dots,k$, equate to zero and solve. This will lead to the MLE as $\hat{p}_i=\frac{x_i}{n}$, $i=1,\dots,k$, the same as the moment estimates. [This derivation and illustration that the matrix of second-order derivatives, at this critical point $(p_1,\dots,p_k)=(\frac{x_1}{n},\dots,\frac{x_k}{n})$, is negative definite, are left to the students. When trying to show negative definiteness remember that there are only $k-1$ variables $x_1,\dots,x_{k-1}$ and only $k-1$ parameters $p_1,\dots,p_{k-1}$.] On the other hand, if we consider point Bernoulli multinomial trials, then we have iid variables and the $i$-th variable takes the value $(x_{1i},\dots,x_{k_i})$ where only one of $x_{ji}$'s is 1 and the remaining are zeros for each such $n$ trials so that $x_j=\sum_{i=1}^{n}x_{ji}$, $j=1,\dots,k$ so that the likelihood function becomes

$$L=p_1^{x_1}\cdots p_k^{x_k}$$

and note that the multinomial coefficient $\frac{n!}{x_1!\cdots x_k!}$ is absent. Now, proceed with the maximization of this $L$ and we will end up with the same estimates as above. Calculus can also be used here.

### 11.3.4 Minimum dispersion method

This method, introduced by this author in 1967, is based on the principle of minimization of a measure of "dispersion" or "scatter" or "distance". Let $x_1,\dots,x_n$ be a sample from a specified population with parameter $\theta$ to be estimated. Let $T=T(x_1,\dots,x_n)$ be an arbitrary estimator for the parameter $\theta$. Then $E|T-\theta|$ is a measure of distance between $T$ and $\theta$. Similarly, $\{E|T-\theta|^r\}^{\frac{1}{r}}$, $r\geq 1$ is a measure of distance between $T$ and $\theta$ or a measure of dispersion or scatter in $T$ from the point of location $\theta$. If there exists a $T$ such that a pre-selected measure of scatter in $T$ from the point of location $\theta$ is a minimum, then such a $T$ is the minimum dispersion estimator. In Decision Theory, $|T-\theta|$ is called a "loss function" (the terminology in Decision Theory is not that proper because $(T-\theta)$ is understood by a layman to be loss or gain in using $T$ to estimate $\theta$, and then taking the expected loss as risk also is not that logical) and the expected value of the loss function is called the "risk" function, which, in fact, will be a one to one function of the dispersion in $T$ from the point of location $\theta$. Some general results associated with squared distances will be considered later, and hence further discussion is postponed.

For the next properties to be discussed, we need to recall a few results from earlier chapters. These will be stated here as lemmas without proofs.

**Lemma 11.1.** *Let $x$ be a real random variable and $\alpha$ be an arbitrary constant. Then that value of $\alpha$ for which $E[x-\alpha]^2$ is a minimum is $\alpha = E(x)$ or $\min_\alpha E[x-\alpha]^2 \Rightarrow \alpha = E(x)$, and that value of $\beta$ for which $E|x-\beta|$ is a minimum is $\beta =$ the median of $x$, that is, $\min_\beta E|x-\beta| \Rightarrow \beta =$ median of $x$. That is,*

$$\min_\alpha E[x-\alpha]^2 \quad \Rightarrow \quad \alpha = E(x) \tag{11.16}$$

$$\min_\beta E|x-\beta| \quad \Rightarrow \quad \beta = \textit{median of } x. \tag{11.17}$$

**Lemma 11.2.** *Let $y$ and $x$ be real random variables and let $g(x)$ be an arbitrary function of $x$ at a given value of $x$. Then*

$$\min_g E[y-g(x)]^2\Big|_{x=given} \quad \Rightarrow \quad g(x) = E(y|x) \tag{11.18}$$

*or $g(x=b)$ is the conditional expectation of $y$ at the given value of $x=b$, which will minimize the squared distance between $y$ and an arbitrary function of $x$ at a given $x$.*

**Lemma 11.3.**

$$E(y) = E_x[E(y|x)] \tag{11.19}$$

*whenever all the expected values exist, where the inside expectation is taken in the conditional space of $y$ given $x$, and the outside expectation is taken in the marginal space $x$. Once the inside expectation is taken, then the given value of $x$ is replaced by the random variable $x$ and then the expectation with respect to $x$ is taken. This is the meaning of expectation of the conditional expectation.*

**Lemma 11.4.**

$$\mathrm{Var}(y) = E[\mathrm{Var}(y|x)] + \mathrm{Var}[E(y|x)] \tag{11.20}$$

*whenever all the variances exist. That is, the variance of $y$ is the sum of the expected value of the conditional variance of $y$ given $x$ and the variance of the conditional expectation of $y$ given $x$.*

### 11.3.5 Some general properties of point estimators

We have looked into the property of relative efficiency of estimators. We have also looked into minimum dispersion estimators and minimum risk estimators. Do such estimators really exist? We can obtain some general properties which will produce some bounds for the distance of the estimator from the parameter for which the estimator is constructed. One such property is in terms of what is known as *Fisher's information*. Consider a simple random sample from a population with density/probability function $f(x,\theta)$. Then the joint density/probability function, denoted by $L = L(x_1,\dots,x_n)$, is the following:

$$L = \prod_{j=1}^{n} f(x_j,\theta) \quad \Rightarrow \quad \ln L = \sum_{j=1}^{n} \ln f(x_j,\theta).$$

Let us differentiate partially with respect to $\theta$. For the time being, we will take $\theta$ as a real scalar parameter. Then we have

$$\frac{\partial}{\partial\theta}\ln L = \sum_{j=1}^{n} \frac{\partial}{\partial\theta}\ln f(x_j,\theta). \tag{11.21}$$

Since the total probability is 1, we have

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_n = 1 \tag{11.22}$$

when the variables are continuous. Replace the integrals by sums when discrete. We will illustrate the results for the continuous case. The steps will be parallel for the discrete case. Let us differentiate both sides of (11.22) with respect to $\theta$. The right side gives zero. Can we differentiate the left side inside the integral sign? If the support of $f(x,\theta)$ or the interval where $f(x,\theta)$ is non-zero contains the parameter $\theta$, such as a uniform density over $[0,\theta]$, then we cannot take the derivative inside the integral sign because the limits of integration will contain $\theta$ also. Hence the following procedure is not applicable in situations where the support of $f(x,\theta)$ depends on $\theta$. If we can differentiate inside the integral, then we have the following, by writing the multiple integral as $\int_X$ and the wedge product of differentials as $\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_n = \mathrm{d}X$, where $X' = (x_1,\dots,x_n)$, prime denoting the transpose:

$$\int_X \frac{\partial L}{\partial\theta}\mathrm{d}X = 0. \tag{11.23}$$

But note that $\frac{\partial L}{\partial\theta} = [\frac{\partial}{\partial\theta}\ln L]L$ so that we can write (11.23) as an expected value

$$\begin{aligned}\int_X \frac{\partial L}{\partial\theta}\mathrm{d}X = 0 \quad &\Rightarrow \quad \int_X\left[\frac{\partial}{\partial\theta}\ln L\right]L\mathrm{d}X = 0\\ &\Rightarrow \quad E\left[\frac{\partial}{\partial\theta}\ln L\right] = 0. \end{aligned}\tag{11.24}$$

Then from (11.21),

$$E\Big[\frac{\partial}{\partial\theta}\ln L\Big]=\sum_{j=1}^{n}E\Big[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\Big]=nE\Big[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\Big] \tag{11.25}$$

since $x_1,\dots,x_n$ are iid variables. Let $T=T(x_1,\dots,x_n)$ be an estimator for $\theta$. If $T$ is unbiased for $\theta$ then $E(T)=\theta$, otherwise $E(T)=\theta+b(\theta)$ where $b(\theta)$ is the bias in using $T$ to estimate $\theta$. Writing it as an integral and then differentiating both sides with respect to $\theta$ and differentiating inside the integral sign, we have the following:

$$\begin{aligned}&\int_X TL\mathrm{d}X=\theta+b(\theta)\quad\Rightarrow\quad\int_X T\frac{\partial L}{\partial\theta}\mathrm{d}X=1+b'(\theta)\\&\frac{\partial}{\partial\theta}(TL)=T\Big(\frac{\partial L}{\partial\theta}\Big)=T\Big[\frac{\partial}{\partial\theta}\ln L\Big]L.\end{aligned} \tag{11.26}$$

Note that $T$ does not contain $\theta$. Writing as expected values, we have

$$\begin{aligned}\int_X T\Big[\frac{\partial}{\partial\theta}\ln L\Big]L\mathrm{d}X=1+b'(\theta)\quad\Rightarrow\\E\Big[T\Big(\frac{\partial}{\partial\theta}\ln L\Big)\Big]=1+b'(\theta).\end{aligned}$$

For any two real random variables $u$ and $v$, we have the following results for covariances and correlations:

$$\begin{aligned}\mathrm{Cov}(u,v)&=E[(u-E(u))(v-E(v))]\\&=E[u(v-E(v))]=E[v(u-E(u))]\end{aligned}$$

because the second terms will be zeros due to the fact that for any random variable $u$, $E[u-E(u)]=0$. Further, since correlation, in absolute value, is $\le 1$, we have

$$[\mathrm{Cov}(u,v)]^2\le\mathrm{Var}(u)\,\mathrm{Var}(v)$$

which is also Cauchy–Schwarz inequality. Since $E[\frac{\partial}{\partial\theta}\ln L]=0$ from (11.24) and since $E[T(\frac{\partial}{\partial\theta}\ln L)]=\mathrm{Cov}(T,\frac{\partial}{\partial\theta}\ln L)$ we have the following result, which holds under the following regularity conditions: (i) The support of $f(x_j,\theta)$ is free of $\theta$; (ii) $f(x_j,\theta)$ is differentiable with respect to $\theta$; (iii) $\int_{x_j}\frac{\partial}{\partial\theta}f(x_j,\theta)\mathrm{d}x_j$ exists; (iv) $\mathrm{Var}(\frac{\partial}{\partial\theta}\ln f(x_j,\theta))$ is finite and positive for all $\theta\in\Omega$.

**Result 11.5** (Cramer–Rao inequality). *For the quantities defined above,*

$$\begin{aligned}\Big[\mathrm{Cov}\Big(T,\frac{\partial}{\partial\theta}\ln L\Big)\Big]^2=[1+b'(\theta)]^2\le\mathrm{Var}(T)\,\mathrm{Var}\Big(\frac{\partial}{\partial\theta}\ln L\Big)\quad\Rightarrow\\\mathrm{Var}(T)\ge\frac{[1+b'(\theta)]^2}{\mathrm{Var}(\frac{\partial}{\partial\theta}\ln L)}=\frac{[1+b'(\theta)]^2}{n\,\mathrm{Var}(\frac{\partial}{\partial\theta}\ln f(x_j,\theta))}\\=\frac{[1+b'(\theta)]^2}{E[\frac{\partial}{\partial\theta}\ln L]^2}=\frac{[1+b'(\theta)]^2}{nE[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)]^2}\end{aligned} \tag{11.27}$$

*which gives a lower bound for the variance of the estimator $T$.*

Note that since $x_1,\dots,x_n$ are iid variables and since $E[\frac{\partial}{\partial\theta}\ln L]=0$ and $E[\frac{\partial}{\partial\theta}f(x_j,\theta)]=0$ we have

$$\mathrm{Var}\Big(\frac{\partial}{\partial\theta}\ln L\Big)=E\Big[\frac{\partial}{\partial\theta}\ln L\Big]^2=nE\Big[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\Big]^2. \tag{11.28}$$

The result in (11.27) is known as *Cramer–Rao inequality*. If $T$ is unbiased for $\theta$, then $b(\theta)=0\Rightarrow b'(\theta)=0$ and then

$$\mathrm{Var}(T)\geq\frac{1}{I_n(\theta)}=\frac{1}{nI_1(\theta)} \tag{11.29}$$

where $I_n(\theta)=\mathrm{Var}(\frac{\partial}{\partial\theta}\ln L)$ which has the various representations given in (11.28). This $I_n(\theta)$ is called *Fisher's information about $\theta$ in the sample of size n* and $I_1(\theta)$ is Fisher's information in one observation. Hence the bounds for variance of $T$, given in (11.27) and (11.29) are called *information bound* for the variance of an estimator. Since $\mathrm{Var}(T)$ and $I_n(\theta)$ are inversely related, smaller the variance means larger the information content, which is also consistent with a layman's visualization of "information" about $\theta$ that is contained in the sample. Larger the variance means smaller the information and smaller the variance means larger the information content because in this case the estimator $T$ is concentrated around $\theta$ when the variance is small.

**Note 11.9.** "Information" in "Information Theory" is different from Fisher's information given in (11.29). The information in Information Theory is a measure of lack of uncertainty in a given scheme of events and the corresponding probabilities. This lack of "information" is also called "entropy" and this is the measure appearing in Communication Theory, Engineering and Physics problems. Students who wish to know more about Information Theory may look into the book [14], a copy is available in CMS library.

We can also obtain an alternative representation for $I_n(\theta)$. Let us consider the equation $E[\frac{\partial}{\partial\theta}\ln L]=0$ and let us differentiate both sides with respect to $\theta$.

$$\begin{aligned}
E\Big[\frac{\partial}{\partial\theta}\ln L\Big]=0\quad&\Rightarrow\quad\int_X\Big(\frac{\partial}{\partial\theta}\ln L\Big)L\mathrm{d}X=0\\
&\Rightarrow\quad\int_X\frac{\partial}{\partial\theta}\Big[\Big(\frac{\partial}{\partial\theta}\ln L\Big)L\Big]\mathrm{d}X=0\\
&\Rightarrow\quad\int_X\Big\{\Big(\frac{\partial^2}{\partial\theta^2}\ln L\Big)L+\Big(\frac{\partial}{\partial\theta}\ln L\Big)\frac{\partial L}{\partial\theta}\Big\}=0\\
&\Rightarrow\quad\int_X\Big\{\Big(\frac{\partial^2}{\partial\theta^2}\ln L\Big)L+\Big(\frac{\partial}{\partial\theta}\ln L\Big)^2L\Big\}\mathrm{d}X=0\\
&\Rightarrow\quad\int_X\Big(\frac{\partial}{\partial\theta}\ln L\Big)^2L\mathrm{d}X=-\int_X\Big(\frac{\partial^2}{\partial\theta^2}\ln L\Big)L\mathrm{d}X.
\end{aligned}$$

This shows that

$$\begin{aligned}\operatorname{Var}\Big(\frac{\partial}{\partial\theta}\ln L\Big) &= E\Big[\frac{\partial}{\partial\theta}\ln L\Big]^2 = -E\Big[\frac{\partial^2}{\partial\theta^2}\ln L\Big]\\ &= nE\Big[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\Big]^2 = -nE\Big[\frac{\partial^2}{\partial\theta^2}\ln f(x_j,\theta)\Big].\end{aligned}\tag{11.30}$$

This may be a more convenient formula when computing Fisher's information.

**Example 11.15.** Check whether the minimum variance bound or information bound is attained for the moment estimator of the parameter (1) $\theta$ in an exponential population, (2) $p$ in a binomial population.

**Solution 11.15.** (1) The exponential density is given by

$$f(x,\theta) = \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, \quad \theta > 0,\ x \ge 0$$

and zero elsewhere. Hence

$$\begin{aligned}\ln f(x_j,\theta) &= -\ln\theta - \frac{x_j}{\theta}\\ \Rightarrow\quad \frac{\partial}{\partial\theta}\ln f(x_j,\theta) &= -\frac{1}{\theta} + \frac{x_j}{\theta^2}\\ -\frac{\partial^2}{\partial\theta^2}\ln f(x_j,\theta) &= -\frac{1}{\theta^2} + \frac{2x_j}{\theta^3}.\end{aligned}$$

Hence

$$\begin{aligned}-nE\Big[\frac{\partial^2}{\partial\theta^2}\ln f(x_j,\theta)\Big] &= -\frac{n}{\theta^2} + \frac{2nE(x_j)}{\theta^3}\\ &= -\frac{n}{\theta^2} + \frac{2n\theta}{\theta^3}\\ &= \frac{n}{\theta^2} = nI_1(\theta) = I_n(\theta).\end{aligned}$$

The moment estimator of $\theta$ is $\bar{x}$ and

$$\operatorname{Var}(\bar{x}) = \frac{\operatorname{Var}(x)}{n} = \frac{\theta^2}{n} = \frac{1}{I_n(\theta)}.$$

Hence the information bound is attained or $\bar{x}$ is the minimum variance unbiased estimator (MVUE) for $\theta$ here.

(2) For the binomial case the moment estimator of $p$ is $\hat{p} = \frac{x}{n}$ and

$$\operatorname{Var}(\hat{p}) = \frac{\operatorname{Var}(x)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

The likelihood function for the Bernoulli population is

$$L = p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots .$$

[One can also take the binomial probability function as it is, taken as a sample of size 1 from a binomial population, namely $f(x) = \binom{n}{x}p^x(1-p)^{n-x}$.]

$$\begin{aligned}
\ln L &= x \ln p + (n-x)\ln(1-p) \\
\frac{\mathrm{d}}{\mathrm{d}p}\ln L &= \frac{x}{p} - \frac{(n-x)}{1-p}, \\
\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ln L &= -\frac{x}{p^2} - \frac{(n-x)}{(1-p)^2} \\
E\left[-\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ln L\right] &= \frac{E(x)}{p^2} + \frac{n - E(x)}{(1-p)^2} = \frac{n}{p(1-p)} \\
\Rightarrow \quad \mathrm{Var}(\hat{p}) &= \frac{1}{I_n(p)}.
\end{aligned}$$

Hence $\hat{p}$ here is the MVUE for $p$.

## Exercises 11.3

**11.3.1.** By using the method of moments estimate the parameters $\alpha$ and $\beta$ in (1) type-1 beta population with parameters $(\alpha, \beta)$; (2) type-2 beta population with the parameters $(\alpha, \beta)$.

**11.3.2.** Prove Lemmas 11.1 and 11.2.

**11.3.3.** Prove Lemmas 11.3 and 11.4.

**11.3.4.** Consider the populations (1) $f_1(x) = \alpha x^{\alpha-1}$, $0 \le x \le 1$, $\alpha > 0$, and zero elsewhere; (2) $f_2(x) = \beta(1-x)^{\beta-1}$, $0 \le x \le 1$, $\beta > 0$ and zero elsewhere. Construct two unbiased estimators each for the parameters $\alpha$ and $\beta$ in (1) and (2).

**11.3.5.** Show that the sample mean is unbiased for (1) $\lambda$ in a Poisson population $f_1(x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}$, $x = 0, 1, 2, \ldots, \lambda > 0$ and zero elsewhere; (2) $\theta$ in the exponential population $f_2(x) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}$, $x \ge 0$, $\theta > 0$ and zero elsewhere.

**11.3.6.** For $\theta$ in a uniform population over $[0, \theta]$ construct an estimate $T = c_1x_1 + c_2x_2 + c_3x_3$, where $x_1$, $x_2$, $x_3$ are iid, as uniform over $[0, \theta]$, such that $E(T) = \theta$. Find $c_1$, $c_2$, $c_3$ such that two unbiased estimators $T_1$ and $T_2$ are obtained where $T_1 = 2\bar{x}$, $\bar{x}$ = the sample mean. Compute $E[T_1 - \theta]^2$ and $E[T_2 - \theta]^2$. Which is relatively more efficient?

**11.3.7.** Consider a simple random sample of size $n$ from a Laplace density or double exponential density

$$f(x) = \frac{1}{2}\mathrm{e}^{-|x-\theta|}, \quad -\infty < x < \infty,\ -\infty < \theta < \infty.$$

Evaluate the MLE of $\theta$. Is it MVUE of $\theta$?

**11.3.8.** For the population in Exercise 11.2.6, let $T_1$ be the moment estimator and $T_2$ be the MLE of $\theta$. Check to see whether $T_1$ or $T_2$ is relatively more efficient.

**11.3.9.** In Exercise 11.2.6, is the moment estimator (1) consistent, (2) sufficient, for $\theta$?

**11.3.10.** In Exercise 11.2.6, is the maximum likelihood estimator (1) consistent, (2) sufficient, for $\theta$?

**11.3.11.** Consider a uniform population over $[a,b]$, $b > a$. Construct the moment estimators for $a$ and $b$.

**11.3.12.** In Exercise 11.3.4, is the moment estimator or MLE relatively more efficient for (1) $a$ when $b$ is known, (2) $b$ when $a$ is known.

**11.3.13.** In Exercise 11.3.4, are the moment estimators sufficient for the parameters in situations (1) and (2) there?

**11.3.14.** In Exercise 11.3.4, are the MLE sufficient for the parameters in situations (1) and (2) there?

## 11.4 Point estimation in the conditional space

### 11.4.1 Bayes' estimates

So far, we have been considering one given or pre-selected population having one or more parameters which are fixed but unknown constants. We were trying to give point estimates based on a simple random sample of size $n$ from this pre-selected population with fixed parameters. Now we consider the problem of estimating one variable by observing or preassigning another variable. There are different types of topics under this general procedure. General model building problems fall in this category of estimating or predicting one or more variables by observing or preassigning one or more other variables. We will start with a Bayesian type problem first.

Usual Bayesian analysis in the simplest situation is stated in terms of one variable, and one parameter having its own distribution. Let $x$ have a density/probability function for a fixed value of the parameter $\theta$. If $\theta$ is likely to have its own distribution, then we denote the density of $x$ as a conditional statement, $f(x|\theta)$ or the density/probability function of $x$ at a given $\theta$. If $\theta$ has a density/probability function of its own, denoted by $g(\theta)$, then the joint density/probability function of $x$ and $\theta$, denoted by $f(x,\theta)$, is the following:

$$f(x,\theta) = f(x|\theta)g(\theta).$$

We may use the following technical terms: $g(\theta)$ as the *prior* density /probability function of $\theta$ and $f(x|\theta)$ as the conditional density /probability function of $x$, at preassigned value of $\theta$. Then the unconditional density/probability function of $x$, denoted by $f_x(x)$, is available by integrating out (or summing up in the discrete case) the variable $\theta$ from the joint density/probability function. That is,

$$f_x(x) = \int_\theta f(x|\theta)g(\theta)\mathrm{d}\theta \quad \text{when } \theta \text{ is continuous} \tag{11.31}$$

$$= \sum_\theta f(x|\theta)g(\theta) \quad \text{when } \theta \text{ is discrete.} \tag{11.32}$$

Then the conditional density of $\theta$ at given value of $x$, denoted by $g(\theta|x)$, which may also be called the *posterior density/probability function of* $\theta$ as opposed to *prior density/probability function of* $\theta$, is the following:

$$g(\theta|x) = \frac{f(x,\theta)}{f_x(x)} = \frac{f(x|\theta)g(\theta)}{\int_\theta f(x|\theta)g(\theta)\mathrm{d}\theta} \tag{11.33}$$

and replace the integral by the sum in the discrete case.

If we are planning to estimate or predict $\theta$ by using $x$ or at a preassigned value of $x$, then from Lemma 11.2 we see that the "best" estimate or best predictor of $\theta$, given $x$, is the conditional expectation of $\theta$, given $x$. Here, we are asking the question: what is a very good estimate of $\theta$ once $x$ is observed or what is a good estimate of $\theta$ in the presence of a preassigned value of $x$? Then from Lemma 11.2 we have the answer as the 'best" estimator, best in the minimum mean square sense, is the conditional expectation of $\theta$, given $x$. This is the Bayes' estimate, given by

$$E(\theta|x) = \int_\theta \theta g(\theta|x)\mathrm{d}\theta \quad \text{if } \theta \text{ is continuous}$$

$$= \sum_\theta \theta g(\theta|x) \quad \text{if } \theta \text{ is discrete.} \tag{11.34}$$

**Example 11.16.** An experiment started with $n = 20$ rabbits. But rabbits die out one by one before the experiment is completed. Let $x$ be the number that survived and let $p$ be the true probability of survival for reach rabbit. Then $x$ is a binomial random variable with parameters $(p, n = 20)$. Suppose that in this particular experiment 15 rabbits survived at the completion of the experiment. This $p$ need not be the same for all experimental rabbits. Suppose that $p$ has a type-1 beta density with parameters $(\alpha = 3, \beta = 5)$. Compute the Bayes' estimate of $p$ in the light of the observation $x = 15$.

**Solution 11.16.** According to our notation,

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n,\ 0 < p < 1$$

and

$$g(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 \le p \le 1,\ \alpha > 0, \beta > 0.$$

The joint probability function

$$\begin{aligned} f(x,p) &= f(x|p)g(p) \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha+x-1}(1-p)^{\beta+n-x-1}, \quad x = 0,1,\dots,n \\ & 0 \le p \le 1,\ \alpha > 0,\ \beta > 0. \end{aligned}$$

The unconditional probability function of $x$ is given by

$$\begin{aligned} f_x(x) &= \int_0^1 f(x,p)\mathrm{d}p = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\mathrm{d}p \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}. \end{aligned}$$

Therefore,

$$\begin{aligned} g(p|x) &= \frac{f(x,p)}{f_x(x)} \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}p^{\alpha+x-1}(1-p)^{\beta+n-x-1}, \quad 0 \le p \le 1. \end{aligned}$$

Hence

$$\begin{aligned} E(p|x) &= \int_0^1 pg(p|x)\mathrm{d}p \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\int_0^1 p \times p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\mathrm{d}p \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\frac{\Gamma(\alpha+x+1)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n+1)} \\ &= \frac{\alpha+x}{\alpha+\beta+n}. \end{aligned}$$

This is the Bayes' estimator as a function of $x$. Since we have $\alpha = 3$, $\beta = 5$, $n = 20$ and $x$ is observed as 15, the Bayes' estimate of $p$, denoted by $E[p|x = 15]$, is given by

$$E[p|x = 15] = \frac{3+15}{3+5+20} = \frac{9}{14}.$$

The moment estimator and the maximum likelihood estimator of $p$ is $\hat{p} = \frac{x}{n}$ and the corresponding estimate is $\frac{15}{20} = \frac{9}{12}$. Bayes' estimate here is slightly reduced from $\frac{9}{12}$ to $\frac{9}{14}$ and the unconditional expected value of $p$ is

$$\frac{\alpha}{\alpha+\beta} = \frac{3}{3+5} = \frac{3}{8}.$$

**Example 11.17.** Let the conditional density of $x$ given $\theta$ be a gamma density of the type

$$f(x|\theta) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\theta x}, \quad x \geq 0,\ \alpha > 0,\ \theta > 0$$

and the prior density of $\theta$ be again a gamma of the form

$$g(\theta) = \frac{\gamma^{\beta}}{\Gamma(\beta)} \theta^{\beta-1} \mathrm{e}^{-\gamma\theta}, \quad \gamma > 0,\ \theta > 0,\ \beta > 0.$$

Compute the unconditional density of $x$.

**Solution 11.17.** The joint density of $x$ and $\theta$ is given by

$$f(x,\theta) = \frac{\gamma^{\beta}}{\Gamma(\beta)\Gamma(\alpha)} x^{\alpha-1} \theta^{\alpha+\beta-1} \mathrm{e}^{-\theta(x+\gamma)}$$

for $\theta > 0$, $x \geq 0$, $\gamma > 0$, $\alpha > 0$, $\beta > 0$. The unconditional density of $x$ is given by

$$\begin{aligned} f_x(x) &= \int_{\theta=0}^{\infty} f(x,\theta)\mathrm{d}\theta \\ &= \frac{\gamma^{\beta} x^{\alpha-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{\infty} \theta^{\alpha+\beta-1} \mathrm{e}^{-\theta(x+\gamma)} \mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \gamma^{\beta} x^{\alpha-1} (x+\gamma)^{-(\alpha+\beta)},\ x \geq 0,\ \gamma > 0, \alpha > 0,\ \beta > 0 \\ &= \frac{\Gamma(\alpha+\beta)}{\gamma^{\alpha}\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} \left(1 + \frac{x}{\gamma}\right)^{-(\alpha+\beta)} \end{aligned}$$

for $x \geq 0$, $\gamma > 0$, $\alpha > 0$, $\beta > 0$ and zero elsewhere.

**Note 11.10.** The last expression for $f_x(x)$ above is also known as *superstatistics* in physics. For $\alpha = 1$, $\gamma = \frac{1}{q-1}$, $\beta + 1 = \frac{1}{q-1}$, $q \geq 1$ it is Tsallis statistics, for $q \geq 1$, in physics in the area of non-extensive statistical mechanics. Lots of applications are there in various areas of physics and engineering.

### 11.4.2 Estimation in the conditional space: model building

Suppose that a farmer is watching the growth of his nutmeg tree from the time of germination, growth being measured in terms of its height $h$ and time $t$ being measured in units of weeks. At $t = 0$, the seed germinated and the height, $h = 0$. When $t = 1$, after one week, let the height be 10 cm. Then at $t = 1$, $h = 10$, height being measured in centimeters. The question is whether we can predict or estimate height $h$ by observing $t$ so that will we be able to give a “good” estimated value of $h$ at a preassigned value such as $t = 100$, that is, after 100 weeks what will be the height $h$? Let $g(t)$ be an arbitrary function of $t$ which is used to predict or estimate $h$. Then $E|h - g(t)|^2$ is the square of a distance between $h$ and $g(t)$, $E$ denoting the expected value. Suppose that this distance is minimized over all possible functions $g$ and then come up with that $g$ for which this distance is the minimum. Such an estimator can be called a good estimator of $h$. That is, $\min_g E|h - g(t)|^2 \Rightarrow g = ?$ at a preassigned value of $t$. This is already available from Lemma 11.2 and the answer is that $g(t) = E[h|t]$ or it is the conditional expectation of $h$, given $t$, which is the "best" estimator of $h$, best in the sense of minimizing the expected squared error. This conditional expectation or the function $g$ can be constructed in the following situations: (1) the joint density of $h$ and $t$ is available, (2) the joint density of $h$ and $t$ is not available but the conditional density of $h$, given $t$, is available. In general, **the problems of this type is to estimate or predict a variable $y$ at a preassigned value of $x$ where $x$ may contain one or more variables**. The best estimator or the best predictor, best in the minimum mean square sense, is the conditional expectation of $y$ given $x$.

**Example 11.18.** Construct the best estimator of $y$ at $x = \frac{1}{3}$ if $x$ and $y$ have the following joint density:

$$f(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-2-3x)^2}, \quad -\infty < y < \infty,\ 0 \le x \le 1$$

and zero elsewhere.

**Solution 11.18.** Here, $y$ can be easily integrated out.

$$\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-2-3x)^2}\mathrm{d}y = 1$$

from the total probability of a normal density with $\mu = 2 + 3x$ and $\sigma^2 = 1$. Hence the marginal density of $x$ is uniform over $[0,1]$. That is,

$$f_1(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Then, naturally, the conditional density of $y$ given $x$ is normal $N(\mu = 2 + 3x, \sigma^2 = 1)$. Therefore, the conditional expectation of $y$, given $x$, is $E[y|x] = 2 + 3x$, which is the best predictor or estimator of $y$ at any preassigned value of $x$. Hence the best predicted value or the best estimated value of $y$ at $x = \frac{1}{3}$ is $2 + 3(\frac{1}{3}) = 3$. Problems of this type will be taken up later in the chapter on regression problems and hence this method will not be elaborated here.

In Sections 11.3.1 to 11.3.4, we examined point estimation procedures of estimating a parameter or parametric function, a fixed quantity, in a density/probability function of $x$ by taking observations on $x$. Then in Sections 11.4.1 and 11.4.2, we examined two situations of estimating one variable by observing or preassigning another variables or other variables. Now we will examine how to estimate a density function itself, after examining some more properties of estimators.

### 11.4.3 Some properties of estimators

Some interesting properties can be obtained in the conditional space, connecting the properties of unbiasedness and sufficiency of estimators. Let $x_1, \dots, x_n$ be iid from the population designated by the density/probability function $f(x, \theta)$. Let $g(\theta)$ be a function of $\theta$. Let $u = u(x_1, \dots, x_n)$ be an unbiased estimator of $g(\theta)$. Let $T = T(x_1, \dots, x_n)$ be a sufficient statistic for $\theta$. Let the conditional expectation of $u$ given $T$ be denoted by $h(T)$, that is, $h(T) = E[u|T]$. From the unbiasedness, we have $E[u] = g(\theta)$. Now, going to the conditional space with the help of Lemma 11.3 we have

$$g(\theta) = E[u] = E[E(u|T)] = E[h(T)] \quad \Rightarrow \quad E[h(T)] = g(\theta) \tag{11.35}$$

or, $h(T)$ is also unbiased for $g(\theta)$. Thus we see that if there exists a sufficient statistic $T$ for $\theta$ and if there exists an unbiased estimator $u$ for a function of $\theta$, namely, $g(\theta)$, then the conditional expectation of $u$ for given values of the sufficient statistic $T$ is also unbiased for $g(\theta)$. Now, we will obtain an interesting result on the variance of any unbiased estimator for $g(\theta)$ and the variance of $h(T)$, a function of a sufficient statistic for $\theta$. From Lemma 11.4, we have

$$\begin{aligned} \mathrm{Var}(u) = E[u - g(\theta)]^2 &= \mathrm{Var}(E[u|T]) + E[\mathrm{Var}(E(u|T))] \\ &= \mathrm{Var}(h(T)) + \delta, \quad \delta \geq 0 \end{aligned}$$

where $\delta$ is the expected value of a variance, and variance of a real random variable, whether in the conditional space or in the unconditional space, is always non-negative. Hence what we have established is that the variance of an unbiased estimator for $g(\theta)$, if there exists an unbiased estimator, is greater than or equal to the variance of a function of a sufficient statistic, if there exists a sufficient statistic for $\theta$. That is,

$$\mathrm{Var}(u) \geq \mathrm{Var}(h(T)) \tag{11.36}$$

where $h(T) = E[u|T]$ is the conditional expectation of $u$ given $T$, which is a function of a sufficient statistic for $\theta$. The result in (11.36) is known as *Rao–Blackwell theorem*, named after C.R. Rao and David Blackwell who derived the inequality first. The beauty of the result is that if we are looking for the minimum variance bound for unbiased estimators then we need to look only in the class of functions of sufficient statistics, if there exists a sufficient statistic for the parameter $\theta$. We have seen earlier that if one sufficient statistic $T$ exists for a parameter $\theta$ then there exist many sufficient statistics for the same $\theta$. Let $T_1$ and $T_2$ be two sufficient statistics for $\theta$. Let $E[u|T_i] = h_i(T_i), i = 1, 2$. Should we take $h_1(T_1)$ or $h_2(T_2)$ if we are trying to improve the estimator $u$, in the sense of finding another unbiased estimator which has a smaller variance? Uniqueness for $h(T)$ cannot be achieved unless the estimator $T$ satisfies one more condition of *completeness*. Note that

$$g(\theta) = E(u) = E[h_1(T_1)] = E[h_2(T_2)] \quad \Rightarrow \quad E[h_1(T_1) - h_2(T_2)] = 0. \tag{11.37}$$

**Definition 11.10** (Complete statistics). Let $T$ be a statistic and let $k(T)$ be an arbitrary function of $T$. If $E[k(T)] = 0$ for all $\theta$ in the parameter space $\Omega$ implies that $k(T) \equiv 0$ with probability one then we say that $T$ is a complete statistic for the parameter $\theta$.

Observe that completeness is a property of the density/probability function of $T$ and it tells more about the structure of the density/probability function. If $T$ is a sufficient and complete statistic for $\theta$, then $E[u|T] = h(T)$ is unique. Thus, in a practical situation, if we try to improve an unbiased estimator $u$ for a parametric function $g(\theta)$ then look for a complete sufficient statistic $T$, if there exists such a $T$, then take $h(T) = E[u|T]$ which will give an improved estimator in the sense of having a smaller variance compared to the variance of $u$, and $h(T)$ is unique here also.

**Example 11.19.** Let $x_1$, $x_2$ be iid as exponential with parameter $\theta$. That is, with the density

$$f(x) = \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, \quad \theta > 0,\ x \geq 0$$

and zero elsewhere. Let $u_1 = 0.6x_1 + 0.4x_2$, $u_2 = x_1 + x_2$. Then we know that $u_1$ is unbiased for $\theta$ and $u_2$ is a sufficient statistic for $\theta$. Construct $h(u_2) = E[u_1|u_2]$ and show that it has smaller variance compared to the variance of $u_1$.

**Solution 11.19.** Let us transform $x_1$, $x_2$ to $u_1$, $u_2$. Then

$$\begin{aligned} u_1 &= 0.6x_1 + 0.4x_2 \quad \text{and} \quad u_2 = x_1 + x_2 \quad \Rightarrow \\ x_1 &= -2u_2 + 5u_1 \quad \text{and} \quad x_2 = 3u_2 - 5u_1 \end{aligned}$$

and the Jacobian is 5. Let the joint densities of $x_1$, $x_2$ and $u_1$, $u_2$ be denoted by $f(x_1, x_2)$ and $g(u_1, u_2)$, respectively. Then

$$f(x_1, x_2) = \frac{1}{\theta^2}\mathrm{e}^{-(x_1+x_2)/\theta}, \quad \theta > 0,\ x_1 \geq 0,\ x_2 \geq 0$$

and zero elsewhere, and

$$g(u_1, u_2) = \frac{5}{\theta^2}\mathrm{e}^{-\frac{u_2}{\theta}}$$

and zero elsewhere, where $\frac{5}{3}u_1 < u_2 < \frac{5}{2}u_1$ and $0 < u_1 < \infty$, or $\frac{2}{5}u_2 < u_1 < \frac{3}{5}u_2$ and $0 < u_2 < \infty$. Since $x_1$ and $x_2$ are iid exponential, the sum $u_2$ is a gamma with parameters $(\alpha = 2, \beta = \theta)$ or the density function of $u_2$, denoted by $f_2(u_2)$, is given by

$$f_2(u_2) = \frac{u_2}{\theta^2}\mathrm{e}^{-\frac{u_2}{\theta}}, \quad u_2 \geq 0,\ \theta > 0$$

and zero elsewhere. Hence the conditional density of $u_1$, given $u_2$, denoted by $g(u_1|u_2)$, is available as

$$g(u_1|u_2) = \frac{g(u_1,u_2)}{f_2(u_2)} = \frac{5}{u_2}, \quad \frac{2}{5}u_2 < u_1 < \frac{3}{5}u_2.$$

Hence the conditional expectation of $u_1$, given $u_2$, is the following:

$$\begin{aligned} E[u_1|u_2] &= \frac{5}{u_2}\int_{\frac{2}{5}u_2}^{\frac{3}{5}u_2} u_1 \mathrm{d}u_1 = \frac{5}{u_2}\left[\frac{u_1^2}{2}\right]_{\frac{2}{5}u_2}^{\frac{3}{5}u_2} \\ &= \frac{5}{2u_2}\left[\frac{9}{25}u_2^2 - \frac{4}{25}u_2^2\right] = \frac{u_2}{2}. \end{aligned}$$

Denoting this conditional expectation as $h(u_2) = \frac{u_2}{2}$ and treating it as a function of the random variable $u_2$ we have the variance of $h(u_2)$ as follows:

$$\mathrm{Var}(h(u_2)) = \frac{1}{4}\,\mathrm{Var}(x_1 + x_2) = \frac{1}{4}(\theta^2 + \theta^2) = 0.5\theta^2.$$

But

$$\mathrm{Var}(u_1) = \mathrm{Var}(0.6x_1 + 0.4x_2) = (0.6)^2\theta^2 + (0.4)^2\theta^2 = 0.52\theta^2.$$

This shows that $h(u_2)$ has a smaller variance compared to the variance of the unbiased estimator $u_1$. This illustrates the Rao–Blackwell theorem.

### 11.4.4 Some large sample properties of maximum likelihood estimators

Here, we will examine a few results which will show that the maximum likelihood estimator of a parameter $\theta$ possesses some interesting large sample (as the sample size becomes larger and larger) properties. Rigorous proofs of these results are beyond the scope of this book. We will give an outline of the derivations. In the following derivations, we will be using differentiability of the density/probability function, $f(x,\theta)$, differentiation with respect to a parameter $\theta$ inside the integrals or summations, etc., and hence the procedures are not applicable when the support of $f(x,\theta)$ (where $f(x,\theta)$ is non-zero) depends on the parameter $\theta$. These aspects should be kept in mind. The regularity conditions in Result 11.5 must hold.

Let $x_1,\dots,x_n$ be iid with density/probability function $f(x,\theta)$. The joint density/probability function, denoted by $L_n(X,\theta)$, $X' = (x_1,\dots,x_n)$, a prime denoting a transpose, is given by

$$L_n(X,\theta) = \prod_{j=1}^{n} f(x_j,\theta).$$

Then

$$\frac{\partial}{\partial\theta}\ln L_n(X,\theta) = 0 \tag{11.38}$$

is called the *likelihood equation*. Let a solution of the likelihood equation be denoted by $\hat{\theta}$. Let the true value of the parameter be denoted by $\theta_0$. Then from equation (11.24), we know that

$$E\left[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\bigg|_{\theta=\theta_0}\right] = 0. \tag{11.39}$$

From (11.38), we have

$$\frac{\partial}{\partial\theta}\ln L_n(X,\theta)\bigg|_{\theta=\hat{\theta}} = 0 \quad\Rightarrow\quad \sum_{j=1}^{n}\left[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\bigg|_{\theta=\hat{\theta}}\right] = 0.$$

But from the weak law of large numbers (see Chapter 9),

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\bigg|_{\theta=\hat{\theta}} \to E\left[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\bigg|_{\theta=\theta_0}\right]$$

as $n\to\infty$. But the right side expected value is already zero at the true parameter value $\theta_0$ by equation (11.24). Hence as $n\to\infty$, $\hat{\theta}$ goes to the true parameter value $\theta_0$ with probability one, or

$$\Pr\{|\hat{\theta}-\theta_0| < \epsilon\} \to 1 \quad \text{as } n\to\infty.$$

This shows that $\hat{\theta}$, a solution of the likelihood equation (11.38), is a consistent estimator of the true parameter value $\theta_0$.

**Result 11.6.** *Consider a density/probability function $f(x,\theta)$ where the support does not depend on $\theta$, and the regularity conditions of Result* 11.5 *hold, then the MLE for the parameter $\theta$ is a consistent estimator for $\theta$ when the sample size $n\to\infty$.*

This means that $\hat{\theta}$ is in the neighborhood of the true parameter value $\theta_0$. Let us expand $0 = \frac{\partial}{\partial\theta}\ln L_n(X,\theta)|_{\theta=\hat{\theta}}$ in the neighborhood of the true parameter value $\theta_0$ to the second-order terms. Then

$$0 = \frac{\partial}{\partial\theta}\ln L_n(X,\theta)\bigg|_{\theta=\theta_0} + (\hat{\theta}-\theta_0)\frac{\partial^2}{\partial\theta^2}\ln L_n(X,\theta)\bigg|_{\theta=\theta_0}$$

$$+\frac{(\hat{\theta}-\theta_0)^2}{2}\frac{\partial^3}{\partial\theta^3}\ln L_n(X,\theta)\Big|_{\theta=\theta_1} \tag{11.40}$$

where $|\hat{\theta}-\theta_1|<|\hat{\theta}-\theta_0|$. From (11.40), by multiplying both sides by $\sqrt{n}$ and rearranging terms we have the following:

$$\sqrt{n}(\hat{\theta}-\theta_0)=\frac{-\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ln L_n(X,\theta)|_{\theta=\theta_0}}{\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L_n(X,\theta)|_{\theta=\theta_0}+\frac{1}{n}\frac{(\hat{\theta}-\theta_0)}{2}\frac{\partial^3}{\partial\theta^3}\ln L_n(X,\theta)|_{\theta=\theta_1}} \tag{11.41}$$

The second term in the denominator of (11.41) goes to zero because $\hat{\theta}\to\theta_0$ as $n\to\infty$ and the third derivative of $\ln L_n(X,\theta)$ is assumed to be bounded. Then the first term in the denominator is such that

$$\begin{aligned}\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L_n(X,\theta)\Big|_{\theta=\theta_0}&=\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln f(x_j,\theta)\Big|_{\theta=\theta_0}\\&\to-\mathrm{Var}\left[\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\right]\Big|_{\theta=\theta_0}\end{aligned}$$

by (11.30), which is the information bound $I_1(\theta_0)$. Hence

$$\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L_n(X,\theta)\Big|_{\theta=\theta_0}\to-I_1(\theta_0)$$

where $I_1(\theta_0)$ is assumed to be positive. Hence we may rewrite (11.41) as follows:

$$\sqrt{I_1(\theta_0)}\sqrt{n}(\hat{\theta}-\theta_0)\approx\frac{\sqrt{n}}{\sqrt{I_1(\theta_0)}}\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\Big|_{\theta=\theta_0}$$

where $\frac{\partial}{\partial\theta}\ln f(x_j,\theta)$ has expected value zero and variance $I_1(\theta_0)$. Further, $f(x_j,\theta)$ for $j=1,\dots,n$ are iid variables. Hence by the central limit theorem

$$\frac{\sqrt{n}}{\sqrt{I(\theta_0)}}\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\to N(0,1),\quad\text{as } n\to\infty$$

where $N(0,1)$ is the standard normal, or we may write

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\Big|_{\theta=\theta_0}\to N(0,I_1(\theta_0))$$

which shows that the left side

$$\sqrt{I_1(\theta_0)}\sqrt{n}(\hat{\theta}-\theta_0)\to N(0,1). \tag{11.41a}$$

Since $I_1(\theta_0)$ is free of $n$, this is also the same as saying

$$\sqrt{n}(\hat{\theta}-\theta_0)\to N\left(0,\frac{1}{I_1(\theta_0)}\right) \tag{11.41b}$$

which also shows that $\sqrt{n}\hat{\theta}$ attains its minimum variance bound as $n\to\infty$ or $\hat{\theta}$ is relatively most efficient for $\theta_0$ when $n\to\infty$. Thus, we have the following result.

**Result 11.7.** *When the regularity conditions of Result* 11.5 *hold, the MLE $\hat{\theta}$ of the true parameter value $\theta_0$ is at least asymptotically (as $n \to \infty$) the most efficient and*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\Big(0, \frac{1}{I_1(\theta_0)}\Big) \quad \textit{as } n \to \infty.$$

**Note 11.11.** Equations (11.41a) and (11.41b) are very often misinterpreted in statistical literature. Hence the student must be very careful in using and interpreting (11.41a) and (11.41b). Misinterpretation comes from assuming that $\hat{\theta}$ is approximately normal for large values of the sample size $n$, in the light of (11.41b) or (11.41a), which is incorrect. When $n$ becomes larger and larger the density/probability function of $\hat{\theta}$ may come closer and closer to a degenerate density and not to an approximate normal density. For each $n$, as well as when $n \to \infty$, $\hat{\theta}$, may have its own distribution. For example, if $\theta$ is the mean value in the exponential population then $\hat{\theta} = \bar{x}$, the sample mean, but $\bar{x}$ has a gamma distribution for all $n$, and not an approximate normal distribution. A certain weighted and relocated $\hat{\theta}$, as shown in (11.41a) and (11.41b), has approximate normal distribution as $n$ becomes larger and larger, and finally when $n \to \infty$ a normal distribution.

**Example 11.20.** Illustrate the properties of the maximum likelihood estimator of $\theta$ in the exponential population with density

$$f(x,\theta) = \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, \quad x \geq 0,\ \theta > 0$$

and zero elsewhere.

**Solution 11.20.** Let $x_1, \dots, x_n$ be iid with the density as above. Then the joint density of $X' = (x_1, \dots, x_n)$, prime denoting the transpose, is given by

$$L_n(X,\theta) = \frac{1}{\theta^n}\exp\Big\{-\frac{1}{\theta}(x_1 + \dots + x_n)\Big\},$$

and

$$\ln L_n(X,\theta) = -n\ln\theta - \frac{1}{\theta}(x_1 + \dots + x_n)$$

$$\frac{\partial}{\partial\theta}\ln L_n(X,\theta) = 0 \quad \Rightarrow \quad -\frac{n}{\theta} + \frac{1}{\theta^2}(x_1 + \dots + x_n) = 0$$

$$\Rightarrow \quad \hat{\theta} = \bar{x}.$$

Thus, $\hat{\theta}$ is a solution of the likelihood equation. We know from exponential population that $E(\bar{x}) = \theta_0$ and $\mathrm{Var}(\bar{x}) = \frac{\theta_0^2}{n}$, where $\theta_0$ is the true parameter value, for all $n$. Hence $\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\bar{x} - \theta_0)$. But by the central limit theorem

$$\frac{\bar{x} - \theta_0}{\sqrt{\mathrm{Var}(\bar{x})}} = \frac{\sqrt{n}(\bar{x} - \theta_0)}{\theta_0} \to N(0,1) \quad \text{as } n \to \infty. \tag{a}$$

This also shows that $\sqrt{n}(\bar{x}-\theta_0) \to N(0,\theta_0^2)$ as $n \to \infty$ since $\theta_0^2$ is free of $n$. Now,

$$\frac{\partial^2}{\partial\theta^2}\ln f(x_j,\theta) = \frac{\partial}{\partial\theta}\Big(-\frac{1}{\theta}+\frac{x_j}{\theta^2}\Big)$$
$$= \frac{1}{\theta^2}-\frac{2x_j}{\theta^3}.$$

Hence

$$-E\Big[\frac{\partial^2}{\partial\theta^2}\ln f(x_j,\theta)\Big] = -\frac{1}{\theta^2}+\frac{2E(x_j)}{\theta^3}$$
$$= \frac{1}{\theta^2} = I_1(\theta). \qquad \text{(b)}$$

Therefore, we may write the result in (a) as

$$\sqrt{n}(\hat{\theta}-\theta_0) \to N\Big(0,\frac{1}{I_1(\theta_0)}\Big) \quad \text{as } n \to \infty \qquad \text{(c)}$$

which illustrates the result on asymptotic (as $n \to \infty$) efficiency and normality of the MLE of $\theta$ here.

## Exercises 11.4

**11.4.1.** Derive the Bayes' estimator of the parameter $\lambda$ in a Poisson population if $\lambda$ has a prior (1) exponential distribution with known scale parameter, (2) gamma distribution with known scale and shape parameters.

**11.4.2.** If the conditional density of $x$, given $\theta$, is given by

$$f(x|\theta) = c_1 x^{\alpha-1}\mathrm{e}^{-\theta^\delta x^\gamma}$$

for $\alpha > 0$, $\theta > 0$, $\gamma > 0$, $\delta > 0$, $x \geq 0$ and zero elsewhere and $\theta$ has a prior density of the form

$$g(\theta) = c_2\theta^{\epsilon-1}\mathrm{e}^{-\eta\theta^\delta}$$

for $\epsilon > 0$, $\eta > 0$, $\delta > 0$, $\theta > 0$ and zero elsewhere, (1) evaluate the normalizing constants $c_1$ and $c_2$, (2) evaluate the Bayes' estimate of $\theta$ if $\epsilon$, $\eta$ and $\delta$ are known.

**11.4.3.** Write down your answer in Exercise 11.4.2 for the following special cases: (1) $\delta = 1$; (2) $\delta = 1$, $\alpha = 1$; (3) $\delta = 1$, $\alpha = 1$, $\epsilon = 1$; (4) $\gamma = 1$, $\delta = 1$.

**11.4.4.** Derive the best estimator, best in the minimum mean square sense, of $y$ at preassigned values of $x$, if the joint density of $x$ and $y$ is given by the following: (1)

$$f(x,y) = \frac{\mathrm{e}^{-\frac{y}{2+3x}}}{2+3x}, \quad y \geq 0,\ 0 \leq x \leq 1$$

and zero elsewhere; (2)

$$f(x,y) = \frac{6x^5}{2+3x} e^{-\frac{y}{2+3x}}, \quad y \geq 0,\ 0 \leq x \leq 1$$

and zero elsewhere.

**11.4.5.** Evaluate the best estimator of $y$ at (1) $x = \frac{1}{2}$; (2) $x = \frac{1}{5}$ in Exercise 11.4.4.

**11.4.6.** Evaluate the best predictor or the best estimator of $y$ at preassigned values of $x$ if the conditional density of $y$, given $x$, is the following:

$$g(y|x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y - 2 - 3x - 5x^2)^2\right\}, \quad -\infty < y < \infty$$

and evaluate the best estimate of $y$ when (1) $x = 0$; (2) $x = 1$.

**11.4.7.** Check for asymptotic (as $n \to \infty$) unbiasedness, efficiency, normality and consistency of the maximum likelihood estimator of the parameter (1) $\lambda$ in a Poisson population; (2) $p$ in a Bernoulli population; (3) $\mu$ in $N(\mu,\sigma^2)$ with $\sigma^2$ known; (4) $\sigma^2$ in $N(\mu,\sigma^2)$ where $\mu$ is known; (5) $\alpha$ in a type-1 beta with $\beta = 1$; (6) $\beta$ in a type-1 beta when $\alpha = 1$. Assume that a simple random sample of size $n$ is available in each case.

**11.4.8.** Check for the asymptotic normality of a relocated and re-scaled MLE of $\theta$ in a uniform population over $[0,\theta]$, assuming that a simple random sample of size $n$ is available.

**11.4.9.** Is the MLE in Exercise 11.4.8 consistent for $\theta$?

**11.4.10.** Verify (a) Cramer–Rao inequality, (b) Rao–Blackwell theorem with reference to the MLE of the parameter (1) $p$ in a Bernoulli population; (2) $\lambda$ in a Poisson population; (3) $\theta$ in an exponential population, by taking suitable sufficient statistics whenever necessary. Assume that a simple random sample of size $n$ is available.

## 11.5 Density estimation

Here, we will consider a few situations where one can uniquely determine a density/probability function from some known characteristics. When such characteristic properties are not available, then we will try to estimate the density from data points.

### 11.5.1 Unique determination of the density/probability function

Suppose that for a real positive continuous scalar random variable $x$ the density is unknown but its $h$-th moment is available as

$$E(x^h) = C\frac{\Gamma(\alpha+h)}{\Gamma(\alpha+\beta+h)}, \quad \alpha > 0, \beta > 0 \tag{11.42}$$

and $C$ is such that when $h = 0$ the right side is one. If (11.42) is available for an arbitrary $h$, including complex values of $h$, then we know that a type-1 beta random variable has the $h$-th moment of the type in (11.42). We can identify the density of $x$ as

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1,\ \alpha > 0,\ \beta > 0$$

and zero elsewhere. From the structure of the moment if one cannot see the density right away, then one may go through the inverse Mellin transform formula

$$f(x) = x^{-1}\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} [E(x^h)]x^{-h}\mathrm{d}h, \quad i = \sqrt{-1} \tag{11.43}$$

and $c$ in the contour is such that $c > -\alpha$. In general, if for some positive real scalar random variable, $E(x^h)$ is available, for an arbitrary $h$, then we may go through the formula in (11.43) to obtain the density $f(x)$. The conditions under which $f(x)$ is uniquely determined are the conditions for the existence of the inverse Mellin transform. The discussion of the conditions is beyond the scope of this book. (Details are available in the book [2].) Hence a practical procedure is to search in the list of $h$-th moments of known variables and identify the variable if the $h$-th moment is in the class of moments known to you.

If the characteristic function of a real scalar random variable is available, then we may go through the inverse Fourier transform and obtain the corresponding density function. If the Laplace transform of a positive real scalar random variable, $M_x(-t)$ where $M_x(t)$ is the moment generating function (mgf), is available then we may go through the inverse Laplace transform and obtain the density.

Instead of the transform of the density, such as Mellin transform (arbitrary moments for positive random variables), Laplace transform (mgf with $t$ replaced by $-t$ for positive random variables), Fourier transform (characteristic function) or other transforms, suppose that some properties of the random variable are available. If such properties are unique properties of some specific random variables, then from the properties one can reach the random variable through mathematical techniques of integral equations, functional equations, differential equations, algebraic manipulations, etc. This area is known as characterizations of distributions. (An insight into this area is available from the book [11].) If the properties are not characteristic properties, still we can come up with a class of functions having those properties, and thus we can narrow down the set of functions where the underlying density function belongs. These are some of the procedures for uniquely determining the densities from known characteristics.

### 11.5.2 Estimation of densities

Suppose that such characteristics as described in Section 11.5.1 are not available but only an observed sample (a set of numbers) is available. Can we identify or at least estimate the underlying density function, if there is one? Observe that infinitely many distributions can give rise to the data at hand, and hence unique determination of the underlying density is not possible. This point should be kept in mind when looking at any method of density estimation from observations.

One method is to take the sample distribution (cumulative relative frequencies) function as a representative of the population distribution function (cumulative probability function) $F(x)$.

**Definition 11.11** (Sample distribution function). Let $x_1, \dots, x_n$ be the $n$ observations. Let

$$S_n(x) = \frac{\text{number of observations less than or equal to } x}{n} \tag{11.44}$$

for $-\infty < x < \infty$. Then $S_n(x)$ is called the *sample distribution function or empirical distribution function* based on $n$ observations.

**Example 11.21.** Construct the sample distribution function if the following is a set of observations from some population: $-3, 2, 1, 5$.

**Solution 11.21.** For $-\infty < x < -3$, $S_n(x) = 0$ because there are no observations there. $S_n(x) = \frac{1}{4}$ at $x = -3$ and it remains the same until $x = 1$. Then $S_n(x) = \frac{2}{4}$ at $x = 1$ and in the interval $1 \le x < 2$, and so on. That is,

$$S_n(x) = \begin{cases} 0, & -\infty < x < -3 \\ \frac{1}{4}, & -3 \le x < 1 \\ \frac{2}{4}, & 1 \le x < 2 \\ \frac{3}{4}, & 2 \le x < 5 \\ 1, & x \ge 5 \end{cases}$$

Note that it is a step function. [The student is asked to draw the graph to see that the graph is looking like steps.] We will examine some basic properties of this $S_n(x)$.

Let $u = nS_n(x)$ = the number of observations less than or equal to $x$. Let $p$ be the true probability of finding an observation less than or equal to $x$. Then $p = F(x)$ = the population distribution function of the underlying population. Then $u$ is distributed as a binomial random variables with parameters $(p = F(x), n)$. Then from the binomial probability law

$$\begin{aligned} E[nS_n(x)] &= E(u) = np = nF(x) \\ &\Rightarrow \quad E[S_n(x)] = p = F(x) \end{aligned} \tag{11.45}$$

and

$$\operatorname{Var}(nS_n(x)) = \operatorname{Var}(u) = np(1-p) = nF(x)(1-F(x))$$

$$\Rightarrow \quad \mathrm{Var}(S_n(x)) = \frac{F(x)(1-F(x))}{n}. \tag{11.46}$$

Note that $\mathrm{Var}(S_n(x)) \to 0$ as $n \to \infty$ and $E[S_n(x)] = F(x)$ for all $x$ and $n$. From the weak law of large numbers or from Chebyshev inequality, we have stochastic convergence of $S_n(x)$ to the true distribution function $F(x)$ or

$$\lim_{n\to\infty} \Pr\{|S_n(x) - F(x)| < \epsilon\} = 1$$

for $\epsilon > 0$, however small it may be. Thus, we can say that $S_n(x)$ is a representative of the true distribution function $F(x)$. Then, when $F(x)$ is differentiable, we have the density $f(x)$, given by

$$f(x) = \lim_{\delta\to 0} \frac{F(x+\delta) - F(x)}{\delta}$$

and hence we may make the approximation

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}.$$

Hence let

$$f_n^*(x) = \frac{S_n(x+h_n) - S_n(x-h_n)}{2h_n} \tag{11.47}$$

where $h_n$ is any positive sequence of real numbers converging to zero. This $f_n^*(x)$ can be taken as an estimate of the true density $f(x)$.

## Exercises 11.5

**11.5.1.** Determine the density of the non-negative random variable $x$ where $x$ has the $h$-th moment, for arbitrary $h$, of the form:

$$(i) \quad E(x^h) = \frac{\Gamma(1+h)}{\Gamma(2+h)}, \quad (ii) \quad E(x^h) = \Gamma(1+h)\Gamma(1-h).$$

**11.5.2.** Determine the density of $x$ if the characteristic function is $\phi(t) = \mathrm{e}^{-t^2}$.

**11.5.3.** Determine the density of a non-negative random variable $x$ if the Laplace transform of the density $f(x)$ is $L_f(t) = (1+t)^{-\frac{1}{2}}, 1+t > 0$.

**11.5.4.** Let $f(x)$ be the density of a continuous real scalar random variable $x$. Then Shannon's *entropy* is given by $S = -c\int_x f(x)\ln f(x)\mathrm{d}x$, where $c$ is a constant. By using calculus of variation, or otherwise, determine that $f$ for which $S$ is maximum, subject to the condition that $E(x) = \int_x xf(x)\mathrm{d}x = d < \infty$.

**11.5.5.** Someone is throwing a dart at a target on a plane board. Let the point of hit be $(x,y)$ under a rectangular coordinate system on the board. Let the density function of $(x,y)$ be $f(x,y)$. Let the Euclidean distance of the point of hit from the origin of the rectangular coordinate system be $r = \sqrt{x^2 + y^2}$. Under the assumption that $f(x,y) = g(r)$ where $g$ is some unknown function, and assuming that $x$ and $y$ are independently distributed, derive the densities of $x$ and $y$ and show that $x$ and $y$ are identically normally distributed.

# 12 Interval estimation

## 12.1 Introduction

In Chapter 11, we looked into point estimation in the sense of giving single values or points as estimates for well-defined parameters in a pre-selected population density/probability function. If $p$ is the probability that someone contesting an election will win and if we give an estimate as $p = 0.7$, then we are saying that there is exactly 70% chance of winning. From a layman's point of view, such an exact number may not be that reasonable. If we say that the chance is between 60 and 75%, it may be more acceptable to a layman. If the waiting time in a queue at a check-out counter in a grocery store is exponentially distributed with expected waiting time $\theta$ minutes, time being measured in minutes, and if we give an estimate of $\theta$ as between 5 and 10 minutes it may be more reasonable than giving a single number such as the expected waiting time is exactly 6 minutes. If we give an estimate of the expected life-time of individuals in a certain community of people as between 80 and 90 years, it may be more acceptable rather than saying that the expected life time exactly 83 years. Thus, when the unknown parameter $\theta$ has a continuous parameter space $\Omega$ it may be more reasonable to come up with an interval so that we can say that the unknown parameter $\theta$ is somewhere on this interval. We will examine such interval estimation problems here.

## 12.2 Interval estimation problems

In order to explain the various technical terms in this area, it is better to examine a simple problem and then define various terms appearing there, in the light of the illustrations.

**Example 12.1.** Let $x_1, \dots, x_n$ be iid variables from an exponential population with density

$$f(x,\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}, \quad x \geq 0,\ \theta > 0$$

and zero elsewhere. Compute the densities of (1) $u = x_1 + \dots + x_n$; (2) $v = \frac{u}{\theta}$ and then evaluate $a$ and $b$ such that $\Pr\{a \leq v \leq b\} = 0.95$.

**Solution 12.1.** The moment generating function (mgf) of $x$ is known and it is $M_x(t) = (1-\theta t)^{-1}, 1-\theta t > 0$. Since $x_1, \dots, x_n$ are iid, the mgf of $u = x_1 + \dots + x_n$ is $M_u(t) = (1-\theta t)^{-n}$, $1-\theta t > 0$ or $u$ has a gamma distribution with parameters $(\alpha = n, \beta = \theta)$. The mgf of $v$ is available from $M_u(t)$ as $M_v(t) = (1-t)^{-n}, 1-t > 0$. In other words, $v$ has a gamma density with the parameters $(\alpha = n, \beta = 1)$ or it is free of all parameters since $n$ is known. Let the density of $v$ be denoted by $g(v)$. Then all sorts of probability statements can be made

https://doi.org/10.1515/9783110562545-012

on the variable $v$. Suppose that we wish to find an $a$ such that $\Pr\{v \le a\} = 0.025$ then we have

$$\int_0^a \frac{v^{n-1}}{\Gamma(n)} \mathrm{e}^{-v} \mathrm{d}v = 0.025.$$

We can either integrate by parts or use incomplete gamma function tables to obtain the exact value of $a$ since $n$ is known. Similarly, we can find a $b$ such that

$$\Pr\{x \ge b\} = 0.025 \quad \Rightarrow \quad \int_b^\infty \frac{v^{n-1}}{\Gamma(n)} \mathrm{e}^{-v} \mathrm{d}v = 0.025.$$

This $b$ is also available either integrating by parts or from the incomplete gamma function tables. Then the probability coverage over the interval $[a, b]$ is 0.95 or

$$\Pr\{a \le v \le b\} = 0.95.$$

We are successful in finding $a$ and $b$ because the distribution of $v$ is free of all parameters. If the density of $v$ contained some parameters, then we could not have found $a$ and $b$ because those points would have been functions of the parameters involved. Hence the success of our procedure depends upon finding a quantity such as $v$ here, which is a function of the sample values $x_1, \ldots, x_n$ and the parameter (or parameters) under consideration, but whose distribution is free of all parameters. Such quantities are called *pivotal quantities*.

**Definition 12.1** (Pivotal quantities). A function of the sample values $x_1, \ldots, x_n$ and the parameters under consideration but whose distribution is free of all parameters is called a pivotal quantity.

Let us examine Example 12.1 once again. We have a probability statement

$$\Pr\{a \le v \le b\} = 0.95.$$

Let us examine the mathematical inequalities here.

$$\begin{aligned} a \le v \le b \quad &\Rightarrow \quad a \le \frac{(x_1 + \cdots + x_n)}{\theta} \le b \\ &\Rightarrow \quad \frac{1}{b} \le \frac{\theta}{(x_1 + \cdots + x_n)} \le \frac{1}{a} \\ &\Rightarrow \quad \frac{(x_1 + \cdots + x_n)}{b} \le \theta \le \frac{(x_1 + \cdots + x_n)}{a}. \end{aligned}$$

Since these inequalities are mathematically identical, we must have the probability statements over these intervals identical. That is,

$$\Pr\left\{a \le \frac{(x_1 + \cdots + x_n)}{\theta} \le b\right\} = \Pr\left\{\frac{(x_1 + \cdots + x_n)}{b} \le \theta \le \frac{(x_1 + \cdots + x_n)}{a}\right\}. \tag{12.1}$$

Thus, we have converted a probability statement over $v$ into a probability statement over $\theta$. What is the difference between these two probability statements? The first one says that the probability that the random variable falls on the fixed interval $[a,b]$ is 0.95. In the second statement, $\theta$ is not a random variable but a fixed but unknown parameter and the random variables are at the end points of the interval or here the interval is random, not $\theta$. Hence the probability statement over $\theta$ is to be interpreted as the probability for the random interval $[\frac{u}{b},\frac{u}{a}]$ covers the unknown $\theta$ is 0.95.

In this example, we have cut off 0.025 area at the right tail and 0.025 area at the left tail so that the total area cut off is $0.025 + 0.025 = 0.05$. If we had cut off an area $\frac{\alpha}{2}$ each at both the tails then the total area cut off is $\alpha$ and the area in the middle if $1-\alpha$. In our Example 12.1, $\alpha = 0.05$ and $1-\alpha = 0.95$. We will introduce some standard notations which will come in handy later on.

**Notation 12.1.** Let $y$ be a random variable whose density $f(y)$ is free of all parameters. Then we can compute a point $b$ such that from that point onward to the right the area cut off is a specified number, say $\alpha$. Then this $b$ is usually denoted as $y_\alpha$ or the value of $y$ from there onward to the right the area under the density curve or probability function is $\alpha$ or

$$\Pr\{y \geq y_\alpha\} = \alpha. \tag{12.2}$$

Then from Notation 12.1 if $a$ is a point below which of the left tail area is $\alpha$ then the point $a$ should be denoted as $y_{1-\alpha}$ or the point from where onward to the right the area under the curve is $1-\alpha$ or the left tail area is $\alpha$. In Example 12.1 if we wanted to compute $a$ and $b$ so that equal areas $\frac{\alpha}{2}$ is cut off at the right and left tails, then the first part of equation (12.1) could have been written as

$$\Pr\{v_{1-\frac{\alpha}{2}} \leq v \leq v_{\frac{\alpha}{2}}\} = 1-\alpha.$$

**Definition 12.2** (Confidence intervals). Let $x_1,\dots,x_n$ be a sample from the population $f(x|\theta)$ where $\theta$ is the parameter. Suppose that it is possible to construct two functions of the sample values $\phi_1(x_1,\dots,x_n)$ and $\phi_2(x_1,\dots,x_n)$ so that the probability for the random interval $[\phi_1,\phi_2]$ covers the unknown parameter $\theta$ is $1-\alpha$ for a given $\alpha$. That is,

$$\Pr\{\phi_1(x_1,\dots,x_n) \leq \theta \leq \phi_2(x_1,\dots,x_n)\} = 1-\alpha$$

for all $\theta$ in the parameter space $\Omega$. Then $1-\alpha$ is called the *confidence coefficient*, the interval $[\phi_1,\phi_2]$ is called a $100(1-\alpha)\%$ *confidence interval for* $\theta$, $\phi_1$ is called the *lower confidence limit*, $\phi_2$ is called the *upper confidence limit* and $\phi_2 - \phi_1$ the length of the confidence interval.

When a random interval $[\phi_1,\phi_2]$ is given we are placing $100(1-\alpha)\%$ confidence on our interval saying that this interval will cover the true parameter value $\theta$ with probability $1-\alpha$. The meaning is that if we construct the same interval by using samples of the same size $n$ then in the long run $100(1-\alpha)\%$ of the intervals will contain the true parameter $\theta$. If one interval is constructed, then that interval need not contain the true parameter $\theta$, the chance that this interval contains the true parameter $\theta$ is $1-\alpha$. In our Example 12.1, we were placing 95% confidence in the interval $[\frac{(x_1+\dots+x_n)}{v_{0.025}}, \frac{(x_1+\dots+x_n)}{v_{0.975}}]$ to contain the unknown parameter $\theta$.

From Example 12.1 and the discussions above, it is clear that we will be successful in coming up with a $100(1-\alpha)\%$ confidence interval for a given parameter $\theta$ if we have the following:

(i) A pivotal quantity $Q$, that is, a quantity containing the sample values and the parameter $\theta$ but whose distribution is free of all parameters. [Note that there may be many pivotal quantities in a given situation.]

(ii) $Q$ enables us to convert a probability statement on $Q$ into a mathematically equivalent statement on $\theta$.

How many such $100(1-\alpha)\%$ confidence intervals can be constructed for a given $\theta$, if one such interval can be constructed? The answer is: infinitely many. From our Example 12.1, it is seen that instead of cutting off 0.025 or in general $\frac{\alpha}{2}$ at both ends we could have cut off $\alpha$ at the right tail, or $\alpha$ at the left tail or any $\alpha_1$ at the left tail and $\alpha_2$ at the right tail so that $\alpha_1 + \alpha_2 = \alpha$. In our example, $v_\alpha \le v < \infty$ would have produced an interval of infinite length. Such an interval may not be of much use because it is of infinite length, but our aim is to give an interval which covers the unknown $\theta$ with a given confidence coefficient $1-\alpha$, and if we say that an interval of infinite length will cover the unknown parameter then such a statement may not have much significance. Hence a very desirable property is that the expected length of the interval is as short as possible.

**Definition 12.3** (Central intervals). Confidence intervals, obtained by cutting off equal areas $\frac{\alpha}{2}$ at both the tails of the distribution of the pivotal quantity so that we obtain a $100(1-\alpha)\%$ confidence interval, are called *central intervals*.

It can be shown that if the pivotal quantity has a symmetric distribution then the central interval is usually the shortest in expected value. Observe also that when the length, which is the upper confidence limit minus the lower confidence limit, is taken, it may be free of all variables. In this case, the length and the expected length are one and the same.

## 12.3 Confidence interval for parameters in an exponential population

We have already given one example for setting up confidence interval for the parameter $\theta$ in the exponential population

$$f(x|\theta) = \frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, \quad x \geq 0,\ \theta > 0$$

and zero elsewhere. Our pivotal quantity was $u = \frac{(x_1+\cdots+x_n)}{\theta}$ where $u$ has a gamma distribution with the parameters $(\alpha = n, \beta = 1)$ where $n$ is the sample size, which is known. Hence there is no parameter and, therefore, probabilities can be read from incomplete gamma tables or can be obtained by integration by parts. Then a $100(1-\alpha)\%$ confidence interval for $\theta$ in an exponential population is given by

$$\left[\frac{(x_1+\cdots+x_n)}{u_{\frac{\alpha}{2}}}, \frac{(x_1+\cdots+x_n)}{u_{1-\frac{\alpha}{2}}}\right] = \text{a } 100(1-\alpha)\% \text{ confidence interval}$$

where

$$\begin{aligned}\int_0^{u_{1-\frac{\alpha}{2}}} g(u)\mathrm{d}u &= \frac{\alpha}{2}\\ \int_{u_{\frac{\alpha}{2}}}^{\infty} g(u)\mathrm{d}u &= \frac{\alpha}{2}\end{aligned} \tag{12.3}$$

and

$$g(u) = \frac{u^{n-1}}{\Gamma(n)}\mathrm{e}^{-u}, \quad u \geq 0.$$

**Example 12.2.** Construct a $100(1-\alpha)\%$ confidence interval for the location parameter $\gamma$ in an exponential population, where the scale parameter $\theta$ is known, say $\theta = 1$. Assume that a simple random sample of size $n$ is available.

**Solution 12.2.** The density function is given by

$$f(x|\gamma) = \mathrm{e}^{-(x-\gamma)}, \quad x \geq \gamma$$

and zero elsewhere. Let us consider the MLE of $\gamma$ which is the smallest order statistic $x_{n:1} = y_1$. Then the density of $y_1$ is available as

$$\begin{aligned}g(y_1|\gamma) &= -\frac{\mathrm{d}}{\mathrm{d}z}[\Pr\{x_j \geq z\}]^n\Big|_{z=y_1}\\ &= n\mathrm{e}^{-n(y_1-\gamma)}, \quad y_1 \geq \gamma\end{aligned}$$

and zero elsewhere. Let $u = y_1 - \gamma$. Then $u$ has the density, denoted by $g_1(u)$, as follows:

$$g_1(u) = n\mathrm{e}^{-nu}, \quad u \geq 0$$

and zero elsewhere. Then we can read off $u_{\frac{\alpha}{2}}$ and $u_{1-\frac{\alpha}{2}}$ for any given $\alpha$ from this density. That is,

$$\int_0^{u_{1-\frac{\alpha}{2}}} n\mathrm{e}^{-nu}\mathrm{d}u = \frac{\alpha}{2} \quad \Rightarrow \quad 1 - \mathrm{e}^{-nu_{1-\frac{\alpha}{2}}} = \frac{\alpha}{2}$$

$$\Rightarrow \quad u_{1-\frac{\alpha}{2}} = -\frac{1}{n}\ln\left(1-\frac{\alpha}{2}\right) \tag{a}$$

$$\int_{u_{\frac{\alpha}{2}}}^{\infty} ne^{-nu}\mathrm{d}u = \frac{\alpha}{2} \quad \Rightarrow \quad \mathrm{e}^{-nu_{\frac{\alpha}{2}}} = \frac{\alpha}{2}$$

$$\Rightarrow \quad u_{\frac{\alpha}{2}} = -\frac{1}{n}\ln\left(\frac{\alpha}{2}\right). \tag{b}$$

Now, we have the probability statement

$$\Pr\{u_{1-\frac{\alpha}{2}} \le y_1 - \gamma \le u_{\frac{\alpha}{2}}\} = 1-\alpha.$$

That is,

$$\Pr\{y_1 - u_{\frac{\alpha}{2}} \le \gamma \le y_1 - u_{1-\frac{\alpha}{2}}\} = 1-\alpha.$$

Hence a 100(1 – $\alpha$)% confidence interval for $\gamma$ is given by

$$[y_1 - u_{\frac{\alpha}{2}}, y_1 - u_{1-\frac{\alpha}{2}}]. \tag{12.4}$$

For example, for an observed sample 2, 8, 5 of size 3, a 95% confidence interval for gamma is given by the following:

$$\alpha = 0.05 \quad \Rightarrow \quad \frac{\alpha}{2} = 0.025.$$

$$u_{\frac{\alpha}{2}} = -\frac{1}{n}\ln\left(\frac{\alpha}{2}\right) = -\frac{1}{3}\ln(0.025).$$

$$u_{1-\frac{\alpha}{2}} = -\frac{1}{3}\ln(0.975).$$

An observed value of $y_1 = 2$. Hence a 95% confidence interval for $\gamma$ is $[2 + \frac{1}{3}\ln(0.025), 2 + \frac{1}{3}\ln(0.975)]$.

**Note 12.1.** If both scale parameter $\theta$ and location parameter $\gamma$ are present, then we need simultaneous confidence intervals or a confidence region for the point $(\theta, \gamma)$. Confidence region will be considered later.

**Note 12.2.** In Example 12.2, we have taken the pivotal quantity as the smallest order statistic $y_1 = x_{n:1}$. We could have constructed confidence interval by using a single observation or sum of observations or the sample mean.

## 12.4 Confidence interval for the parameters in a uniform density

Consider $x_1, \ldots, x_n$, iid from a one parameter uniform density

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta$$

and zero elsewhere. Let us construct a 100(1 – $\alpha$)% confidence interval for $\theta$. Assume that a simple random sample of size $n$ is available. The largest order statistic seems to be a convenient starting point since it is the MLE of $\theta$. Let $y_n = x_{n:n}$ be the largest order statistic. Then $y_n$ has the density

$$g(y_n|\theta) = \frac{\mathrm{d}}{\mathrm{d}z}[\Pr\{x_j \le z\}]^n\Big|_{z=y_n} = \frac{n}{\theta^n}y_n^{n-1}, \quad 0 \le y_n \le \theta.$$

Let us take the pivotal quantity as $u = \frac{y_n}{\theta}$. The density of $u$, denoted by $g_1(u)$ is given by

$$g_1(u) = nu^{n-1}, \quad 0 \le u \le 1$$

and zero elsewhere. Hence

$$\int_0^{u_{1-\frac{\alpha}{2}}} nu^{n-1}\mathrm{d}u = \frac{\alpha}{2} \quad \Rightarrow \quad u_{1-\frac{\alpha}{2}} = \left[\frac{\alpha}{2}\right]^{\frac{1}{n}}$$

and

$$\int_{u_{\frac{\alpha}{2}}}^1 nu^{n-1}\mathrm{d}u = \frac{\alpha}{2} \quad \Rightarrow \quad u_{\frac{\alpha}{2}} = \left[1-\frac{\alpha}{2}\right]^{\frac{1}{n}}.$$

Therefore,

$$\begin{aligned}&\Pr\{u_{1-\frac{\alpha}{2}} \le u \le u_{\frac{\alpha}{2}}\} = 1-\alpha\\ &\quad\Rightarrow \quad \Pr\left\{\left[\frac{\alpha}{2}\right]^{\frac{1}{n}} \le \frac{y_n}{\theta} \le \left[1-\frac{\alpha}{2}\right]^{\frac{1}{n}}\right\} = 1-\alpha\\ &\quad\Rightarrow \quad \Pr\left\{\frac{y_n}{(1-\frac{\alpha}{2})^{\frac{1}{n}}} \le \theta \le \frac{y_n}{(\frac{\alpha}{2})^{\frac{1}{n}}}\right\} = 1-\alpha.\end{aligned}$$

Hence a $100(1-\alpha)\%$ confidence interval for $\theta$ in this case is

$$\left[\frac{y_n}{(1-\frac{\alpha}{2})^{\frac{1}{n}}}, \frac{y_n}{(\frac{\alpha}{2})^{\frac{1}{n}}}\right]. \tag{12.5}$$

For example, for an observed sample 8, 2, 5 from this one parameter uniform population a 90% confidence interval for $\theta$ is given by $[\frac{8}{(0.95)^{\frac{1}{3}}}, \frac{8}{(0.05)^{\frac{1}{3}}}]$.

**Note 12.3.** If the uniform population is over $[a, b]$, $b > a$, then by using the largest and smallest order statistics one can construct confidence intervals for $b$ when $a$ is known, for $a$ when $b$ is known. Simultaneous intervals for $a$ and $b$ will be discussed later.

## 12.5 Confidence intervals in discrete distributions

Here, we will consider a general procedure of setting up confidence intervals for the Bernoulli parameter $p$ and the Poisson parameter $\lambda$. In discrete cases, such as a binomial, cutting off tail probability equal to $\frac{\alpha}{2}$ each may not be possible because the probability masses are at individually distinct points. When we add up the tail probabilities we may not get exact values $\frac{\alpha}{2}$, for example, 0.025. When we add up a few points, the sum of the probabilities may be less than 0.025 and when we add up the next probability the total may exceed 0.025. Hence in discrete situations we take the tail probabilities as $\le\frac{\alpha}{2}$ so that the middle probability will be $\ge 1-\alpha$. Take the nearest point so that the tail probability is closest to $\frac{\alpha}{2}$ but less than or equal to $\frac{\alpha}{2}$.

### 12.5.1 Confidence interval for the Bernoulli parameter *p*

We can set up confidence intervals for the Bernoulli parameter $p$ by taking $n$ observations from a Bernoulli population or one observation from a binomial population. The binomial population has the probability function

$$f(x,p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 < p < 1,\ x = 0,1,\dots,n$$

and zero elsewhere. We can assume $n$ to be known. We will see that we cannot find a pivotal quantity $Q$ so that the probability function of $Q$ is free of $p$. For a binomial random variable $x$, we can make a statement

$$\Pr\{x \le x_{1-\frac{\alpha}{2}}\} \le \frac{\alpha}{2}, \tag{12.6}$$

that is, the left tail probability is less than or equal to $\frac{\alpha}{2}$ for any given $\alpha$ if $p$ is known. But since $x$ is not a pivotal quantity $x_{1-\frac{\alpha}{2}}$ will be a function of $p$, that is $x_{1-\frac{\alpha}{2}}(p)$. For a given $p$, we can compute $x_{1-\frac{\alpha}{2}}$ for any given $\alpha$. For a given $p$, we can compute two points $x_1(p)$ and $x_2(p)$ such that

$$\Pr\{x_1(p) \le x \le x_2(p)\} \ge 1-\alpha \tag{a}$$

or we can select $x_1(p)$ and $x_2(p)$, for a given $p$, such that

$$\Pr\{x \le x_1(p)\} \le \frac{\alpha}{2} \tag{b}$$

and

$$\Pr\{x \ge x_2(p)\} \le \frac{\alpha}{2}. \tag{c}$$

For every given $p$, the points $x_1(p)$ and $x_2(p)$ are available. If we plot $x = x_1(p)$ and $x = x_2(p)$, against $p$ then we may get the graphs as shown in Figure 12.1. Let the observed value of $x$ be $x_0$. If the line $x = x_0$ cuts the bands $x_1(p)$ and $x_2(p)$, then the inverse images will be $p_1$ and $p_2$ as shown in Figure 12.1. The cut on $x_1(p)$ will give $p_2$ and that on $x_2(p)$ will give $p_1$ or a $100(1-\alpha)\%$ confidence interval for $p$ is $[p_1,p_2]$. Note that the region below the line $x = x_0$ is characterized by the probability $\frac{\alpha}{2}$ and similarly the region above the line $x = x_0$ is characterized by $\frac{\alpha}{2}$. Hence the practical procedure is the following: Consider equation (b) with $x_1(p) = x_0$ and search through the binomial table for a $p$, then the solution in (b) will give $p_2$. Take equation (c) with $x_2(p) = x_0$ and search through the binomial tables for a $p$, then the solution in (c) gives $p_1$.

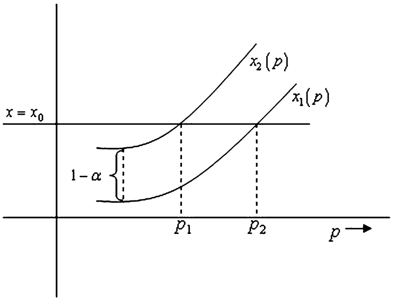


**Figure 12.1:** Lower and upper confidence bands.

Note that in some situations the line $x = x_0$ may not cut one or both of the curves $x_1(p)$ and $x_2(p)$. We may have situations where $p_1$ and $p_2$ cannot be found or $p_1$ may be 0 or $p_1$ may be 1.

Let the observed value of $x$ be $x_0$, for example suppose that we observed 3 successes in $n = 10$ trials. Then our $x_0 = 3$. We can take $x_1(p) = x_{1-\frac{\alpha}{2}}(p) = x_0$ and search for that $p$, say $p_2$, which will satisfy the inequality

$$\sum_{x=0}^{x_0} \binom{n}{x} p_2^x (1-p_2)^{n-x} \le \frac{\alpha}{2}. \tag{12.7}$$

This will give one value of $p$, namely, $p_2$ for which (12.6) holds. Now consider the upper tail probability. Consider the inequality

$$\Pr\{x \ge x_{\frac{\alpha}{2}}(p)\} \le \frac{\alpha}{2}. \tag{12.8}$$

Again let us take $x_2(p) = x_{\frac{\alpha}{2}} = x_0$ and search for $p$ for which (12.8) holds. Call it $p_1$. That is,

$$\sum_{x=x_0}^{n} \binom{n}{x} p_1^x (1-p_1)^{n-x} \le \frac{\alpha}{2}. \tag{12.9}$$

Then

$$\Pr\{p_1 \le p \le p_2\} \le 1 - \alpha \tag{12.10}$$

is the required $100(1-\alpha)\%$ confidence interval for $p$.

**Example 12.3.** If 10 Bernoulli trials gave 3 successes, compute a 95% confidence interval for the probability of success $p$. Note that for the same $p$ both (12.7) and (12.9) cannot hold simultaneously.

**Solution 12.3.** Consider the inequality

$$\sum_{x=0}^{3}\binom{10}{x}p_2^x(1-p_2)^{10-x} \leq 0.025.$$

Look through a binomial table for $n = 10$ and all values of $p$. From tables, we see that for $p = 0.5$ the sum is 0.1710 which indicates that the value of $p_2$ is bigger than 0.5. Most of the tables are given only for $p$ up to 0.5. The reason being that for $p > 0.5$ we can still use the same table. By putting $y = n - x$ and writing

$$\begin{aligned}\sum_{x=0}^{3}\binom{10}{x}p^x(1-p)^{10-x} &= \sum_{x=0}^{3}\binom{10}{y}q^y(1-q)^{n-y} \\ &= \sum_{y=7}^{10}\binom{10}{y}q^y(1-q)^{10-y} \leq 0.025\end{aligned}$$

where $q = 1 - p$. Now looking through the binomial tables we see that $q = 0.4$. Hence $p_2 = 1 - q = 0.6$. Now we consider the inequality

$$\sum_{x=3}^{10}\binom{10}{x}p_1^x(1-p_1)^{10-x} \leq 0.025,$$

which is the same as saying

$$\sum_{x=0}^{2}\binom{10}{x}p_1^x(1-p_1)^{10-x} \geq 0.975.$$

Now, looking through the binomial table for $n = 10$ and all $p$ we see that $p_1 = 0.05$. Hence the required 95% confidence interval for $p$ is $[p_1, p_2] = [0.05, 0.60]$. We have 95% confidence on this interval.

**Note 12.4.** We can use this exact procedure of this section to construct confidence interval for the parameter $\theta$ of a one-parameter distribution whether we have a pivotal quantity or not. Take any convenient statistic $T$ for which the distribution can be derived. This distribution will contain $\theta$. Let $T_0$ be the observed value of $T$. Consider the inequalities

$$\Pr\{T \leq T_0\} \leq \frac{\alpha}{2} \tag{a}$$

and

$$\Pr\{T \geq T_0\} \leq \frac{\alpha}{2}. \tag{b}$$

If the inequalities have solutions, note that both cannot be satisfied by the same $\theta$ value, then the solution of (a) gives $\theta_2$ and the solution of (b) gives $\theta_1$ and then $[\theta_1,\theta_2]$ is a $100(1-\alpha)\%$ confidence interval for $\theta$. As an exercise, the student is advised to use this exact procedure to construct confidence interval for $\theta$ in an exponential population. Use the sample sum as $T$.

This exact procedure can be adopted for getting confidence intervals for the Poisson parameter $\lambda$. In this case, make use of the property that the sample sum is again a Poisson with the parameter $n\lambda$. This is left as an exercise to the student.

## Exercises 12.2–12.5

**12.5.1.** Construct a 95% confidence interval for the location parameter $\gamma$ in an exponential population in Example 12.2 by using (1) $\bar{x}$ the sample mean of a sample of size $n$; (2) the sample sum for a sample of size 2; (3) one observation from the population.

**12.5.2.** By using the observed sample $3,8,4,5$ from an exponential population,

$$f(x|\theta,\gamma)=\frac{1}{\theta}\mathrm{e}^{-(x-\gamma)},\quad x\geq\gamma,\ \theta>0$$

and zero elsewhere, construct a 95% confidence interval for (1): $\theta$ if $\gamma=2$; (2): $\gamma$ if $\theta=4$.

**12.5.3.** Consider a uniform population over $[a,b]$, $b>a$. Assume that the observed sample $2,8,3$ is available from this population. Construct a 95% confidence interval for (1) $a$ when $b=8$; (2) $b$ when $a=1$, by using order statistics.

**12.5.4.** Consider the same uniform population in Exercise 12.5.3 with $a=0$. Assume that a sample of size 2 is available. (1) Compute the density of the sample sum $y=x_1+x_2$; (2) by using $y$ construct a 95% confidence interval for $b$ if the observed sample is $2,6$.

**12.5.5.** Construct a 90% confidence interval for the Bernoulli parameter $p$ if 2 successes are obtained in (1) 10 trials; (2) eight trials.

**12.5.6.** Consider a Poisson population with parameter $\lambda$. Construct a 90% confidence interval for $\lambda$ if $3,7,4$ is an observed sample.

## 12.6 Confidence intervals for parameters in $N(\mu,\sigma^2)$

First, we will consider a simple problem of constructing a confidence interval for the mean value $\mu$ in a normal population when the population variance is known. Then we will consider intervals for $\mu$ when $\sigma^2$ is not known. Then we will look at intervals for $\sigma^2$. In the following situations, we will be constructing the central intervals for convenience. These central intervals will be the shortest when the pivotal quantities have symmetric distributions. In the case of confidence intervals for the population variance, the pivotal quantity taken is a chi-square variable, which does not have a symmetric distribution, and hence the central interval cannot be expected to be the shortest, but for convenience we will consider the central intervals in all situations.

### 12.6.1 Confidence intervals for $\mu$

**Case 1** (Population variance $\sigma^2$ is known)**.** Here, we can take a pivotal quantity as the standardized sample mean

$$z = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0,1)$$

which is free of all parameters when $\sigma$ is known. Hence we can read off $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ so that

$$\Pr\{z_{1-\frac{\alpha}{2}} \le z \le z_{\frac{\alpha}{2}}\} = 1-\alpha.$$

Since a standard normal density is symmetric at $z = 0$, we have $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$. Let us examine the mathematical inequalities.

$$\begin{aligned} -z_{\frac{\alpha}{2}} \le z \le z_{\frac{\alpha}{2}} \quad &\Rightarrow \quad -z_{\frac{\alpha}{2}} \le \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \le z_{\frac{\alpha}{2}} \\ &\Rightarrow \quad -z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \\ &\Rightarrow \quad \bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

and hence

$$\begin{aligned} \Pr\{-z_{\frac{\alpha}{2}} \le z \le z_{\frac{\alpha}{2}}\} &= \Pr\left\{\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} \\ &= 1-\alpha. \end{aligned}$$

Hence a $100(1-\alpha)\%$ confidence interval for $\mu$, when $\sigma^2$ is known, is given by

$$\left[\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]. \tag{12.11}$$

The following Figure 12.2 gives an illustration of the construction of the central confidence interval for $\mu$ in a normal population with $\sigma^2$ known.

**Example 12.4.** Construct a 95% confidence interval for $\mu$ in a $N(\mu, \sigma^2 = 4)$ from the following observed sample: $-5, 0, 2, 15$.

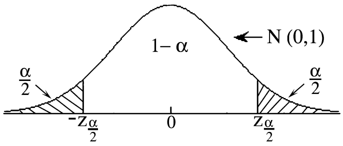


**Figure 12.2:** Confidence interval for $\mu$ in a $N(\mu,\sigma^2)$, $\sigma^2$ known.

**Solution 12.4.** Here, the sample mean $\bar{x} = (-5+0+2+15)/4 = 3$. $1-\alpha = 0.95$ means $\frac{\alpha}{2} = 0.025$. From a standard normal table, we have $z_{0.025} = 1.96$ approximately. $\sigma^2$ is given to be 4, and hence $\sigma = 2$. Therefore, from (12.6), one 95% confidence interval for $\mu$ is given by

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = \left[3 - 1.96\left(\frac{2}{2}\right), 3 + 1.96\left(\frac{2}{2}\right)\right]$$
$$= [1.04, 4.96].$$

We have 95% confidence that the unknown $\mu$ is on this interval.

Note that the length of the interval in this case is

$$\left[\bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] - \left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 2z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

which is free of all variables, and hence it is equal to its expected value, or the expected length of the interval in this case is $2z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 2(1.96) = 3.92$ in Example 12.4.

**Example 12.5.** For a binomial random variable $x$, it is known that for large $n$ ($n \geq 20$, $np \geq 5$, $n(1-p) \geq 5$) the standardized binomial variable is approximately a standard normal. By using this approximation set up an approximate $100(1-\alpha)\%$ confidence interval for $p$ the probability of success.

**Solution 12.5.** We will construct a central interval. We have

$$\frac{x - np}{\sqrt{np(1-p)}} \approx z, \quad z \sim N(0,1).$$

From a standard normal table, we can obtain $z_{\frac{\alpha}{2}}$ so that an approximate probability is the following:

$$\Pr\left\{-z_{\frac{\alpha}{2}} \leq \frac{x - np}{\sqrt{np(1-p)}} \leq z_{\frac{\alpha}{2}}\right\} \approx 1 - \alpha.$$

The inequality can be written as

$$\frac{(x-np)^2}{np(1-p)} \leq z_{\frac{\alpha}{2}}^2.$$

Opening this up as a quadratic equation in $p$, when the equality holds, and then solving for $p$, one has

$$p = \frac{(x + \frac{1}{2}z_{\frac{\alpha}{2}}^2) \mp \sqrt{(x + \frac{1}{2}z_{\frac{\alpha}{2}}^2)^2 - x^2(1 + \frac{1}{n}z_{\frac{\alpha}{2}}^2)}}{n(1 + \frac{1}{n}z_{\frac{\alpha}{2}}^2)}. \tag{12.12}$$

These two roots are the lower and upper $100(1-\alpha)\%$ central confidence limits for $p$ approximately. For example, for $n = 20$, $x = 8$, $\alpha = 0.05$ we have $z_{0.025} = 1.96$. Substituting these values in (12.12) we obtain the approximate roots as 0.22 and 0.61. Hence an approximate 95% central confidence interval for the binomial parameter $p$ in this case is $[0.22, 0.61]$. [Simplifications of the computations are left to the student.]

**Case 2** (Confidence intervals for $\mu$ when $\sigma^2$ is unknown). In this case, we cannot take the standardized normal variable as our pivotal quantity because, even though the distribution of the standardized normal is free of all parameters, we have a $\sigma$ present in the standardized variable, which acts as a *nuisance parameter* here.

**Definition 12.4** (Nuisance parameters). These are parameters which are not relevant for the problem under consideration but which are going to be present in the computations.

Hence our aim is to come up with a pivotal quantity involving the sample values and $\mu$ only and whose distribution is free of all parameters. We have such a quantity here, which is the Student-$t$ variable. Consider the following pivotal quantity, which has a Student-$t$ distribution:

$$\frac{\sqrt{n}(\bar{x} - \mu)}{s_1} \sim t_{n-1}, \quad s_1^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1} \tag{12.13}$$

where $s_1^2$ is an unbiased estimator for the population variance $\sigma^2$. Note that a Student-$t$ distribution is symmetric around $t = 0$. Hence we can expect the central interval being the shortest interval in expected value. For constructing a central $100(1-\alpha)\%$ confidence interval for $\mu$ read off the upper tail point $t_{n-1,\frac{\alpha}{2}}$ such that

$$\Pr\{t_{n-1} \ge t_{n-1,\frac{\alpha}{2}}\} = \frac{\alpha}{2}.$$

Then we can make the probability statement

$$\Pr\{-t_{n-1,\frac{\alpha}{2}} \le t_{n-1} \le t_{n-1,\frac{\alpha}{2}}\} = 1 - \alpha. \tag{12.14}$$

Substituting for $t_{n-1}$ and converting the inequalities into inequalities over $\mu$, we have the following:

$$\Pr\left\{\bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}}\right\} = 1 - \alpha \tag{12.15}$$

which gives a central $100(1-\alpha)\%$ confidence interval for $\mu$. Figure 12.3 gives the illustration of the percentage points.

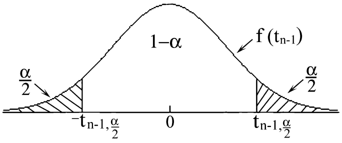


**Figure 12.3:** Percentage points from a Student-$t$ density.

The interval is of length $2t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}}$, which contains the variable $s_1$, and hence it is a random quantity. We can compute the expected value of this length by using the fact that

$$\frac{(n-1)s_1^2}{\sigma^2} \sim \chi_{n-1}^2$$

where $\chi_{n-1}^2$ is a chi-square variable with $(n-1)$ degrees of freedom.

**Example 12.6.** Construct a 99% confidence interval for $\mu$ in a normal population with unknown variance, by using the observed sample 1, 0, 5 from this normal population.

**Solution 12.6.** The sample mean $\bar{x} = (1+0+5)/3 = 2$. An observed value of $s_1^2$ is given by

$$s_1^2 = \frac{1}{2}[(1-2)^2 + (0-2)^2 + (5-2)^2] = 7$$
$$\Rightarrow \quad s_1 = \sqrt{7} = 2.6457513.$$

Now, our $\alpha = 0.01 \Rightarrow \frac{\alpha}{2} = 0.005$. From a Student-$t$ table for $n-1=2$ degrees of freedom, $t_{2,0.005} = 9.925$. Hence a 99% central confidence interval for $\mu$ here is given by

$$\left[2 - 9.925\frac{\sqrt{7}}{\sqrt{3}}, 2 + 9.925\frac{\sqrt{7}}{\sqrt{3}}\right] \approx [-13.16, 17.16].$$

**Note 12.5.** In some books, the students may find the statement that when the sample size $n \geq 30$ one can get a good normal approximation for Student-$t$, and hence take $z_\alpha$ from a standard normal table instead of $t_{\nu,\alpha}$ from the Student-$t$ table with $\nu$ degrees of freedom, for $\nu \geq 30$. The student may look into the exact percentage points from the Student-$t$ table to see that even for the degrees of freedom $\nu = 120$ the upper tail areas of the standard normal and Student-$t$ do not agree with each other. Hence taking $z_\alpha$ instead of $t_{\nu,\alpha}$ for $\nu \geq 30$ is not a proper procedure.

### 12.6.2 Confidence intervals for $\sigma^2$ in $N(\mu, \sigma^2)$

Here, we can consider two situations. (1) $\mu$ is known, (2) $\mu$ is not known, and we wish to construct confidence intervals for $\sigma^2$ in $N(\mu, \sigma^2)$. Convenient pivotal quantities are the following: When $\mu$ is known we can use

$$\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\sim\chi_n^2 \quad\text{and}\quad \sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}\sim\chi_{n-1}^2.$$

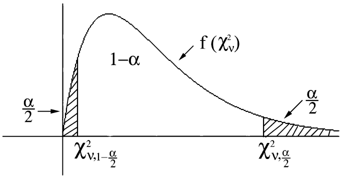
Then from a chi-square density we have

$$\Pr\left\{\chi^2_{n,1-\frac{\alpha}{2}}\le\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{\sigma^2}\le\chi^2_{n,\frac{\alpha}{2}}\right\}=1-\alpha \tag{12.16}$$

and

$$\Pr\left\{\chi^2_{n-1,1,1-\frac{\alpha}{2}}\le\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}\le\chi^2_{n-1,\frac{\alpha}{2}}\right\}=1-\alpha. \tag{12.17}$$

The percentage points are marked in Figure 12.4.



$\frac{\alpha}{2}$
1−α
$f(\chi^2_\nu)$
$\frac{\alpha}{2}$
$\chi^2_{\nu,1-\frac{\alpha}{2}}$
$\chi^2_{\nu,\frac{\alpha}{2}}$

**Figure 12.4:** Percentage points from a chi-square density.

Note that (12.16) can be rewritten as

$$\Pr\left\{\frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}\le\sigma^2\le\frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}\right\}=1-\alpha.$$

A similar probability statement can be obtained by rewriting (12.17). Therefore, a $100(1-\alpha)\%$ central confidence interval for $\sigma^2$ is the following:

$$\left[\frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}},\frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}\right];\quad\left[\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\chi^2_{n-1,\frac{\alpha}{2}}},\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]. \tag{12.18}$$

Note that a $\chi^2$ distribution is not symmetric and hence we cannot expect to get the shortest interval by taking the central intervals. The central intervals are taken only for convenience. When $\mu$ is unknown, then we cannot use $\frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\sigma^2}\sim\chi_n^2$ because the nuisance parameter $\mu$ is present. We can use the pivotal quantity

$$\sum_{j=1}^{n}\frac{(x_j-\bar{x})^2}{\sigma^2}\sim\chi^2_{n-1}$$

and construct a $100(1-\alpha)\%$ central confidence interval, and it is the second one given in (12.18). When $\mu$ is known, we can also use the standardized normal

$$\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\sim N(0,1)$$

as a pivotal quantity to construct confidence interval for $\sigma$, thereby the confidence interval for $\sigma^2$. Note that if $[T_1,T_2]$ is a $100(1-\alpha)\%$ confidence interval for $\theta$ then

$[g(T_1), g(T_2)]$ is a $100(1-\alpha)\%$ confidence interval for $g(\theta)$ when $\theta$ to $g(\theta)$ is a one to one function.

**Example 12.7.** If $-2, 1, 7$ is an observed sample from a $N(\mu,\sigma^2)$, construct a 95% percent confidence interval for $\sigma^2$ when (1) $\mu = 1$, (2) $\mu$ is unknown.

**Solution 12.7.** $\bar{x} = \frac{(-2+1+7)}{3} = 2$, $\sum_{j=1}^{3}(x_j-\bar{x})^2 = (-2-2)^2+(1-2)^2+(7-2)^2 = 42$. $\sum_{j=1}^{3}(x_j-\mu)^2 = (-2-1)^1+(1-1)^2+(7-1)^2 = 45$. $1-\alpha = 0.95 \Rightarrow \frac{\alpha}{2} = 0.025$. From a chi-square table $\chi^2_{n,\frac{\alpha}{2}} = \chi^2_{3,0.025} = 9.35$, $\chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{2,0.025} = 7.38$, $\chi^2_{n,1-\frac{\alpha}{2}} = \chi^2_{3,0.975} = 0.216$, $\chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{2,0.975} = 0.0506$. (2) Then when $\mu$ is unknown a 95% central confidence interval for $\sigma^2$ is given by

$$\left[\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right] = \left[\frac{42}{7.38}, \frac{42}{0.0506}\right]$$
$$= [5.69, 830.04].$$

(1) When $\mu = 1$, we can use the above interval as well as the following interval:

$$\left[\frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}, \frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}\right] = \left[\frac{45}{9.35}, \frac{45}{0.216}\right]$$
$$= [4.81, 208.33].$$

Note that when the information about $\mu$ is used the interval is shorter.

**Note 12.6.** The student may be wondering whether it is possible to construct confidence intervals for $\sigma$, once confidence interval for $\sigma^2$ is established. Then take the corresponding square roots. If $[\phi_1(x_1,\dots,x_n), \phi_2(x_1,\dots,x_n)]$ is a $100(1-\alpha)\%$ confidence interval for $\theta$, then $[h(\phi_1), h(\phi_2)]$ is a $100(1-\alpha)\%$ confidence interval for $h(\theta)$ as long as $\theta$ to $h(\theta)$ is a one to one function.

## Exercises 12.6

**12.6.1.** Consider a $100(1-\alpha)\%$ confidence interval for $\mu$ in a $N(\mu,\sigma^2)$ where $\sigma^2$ is known, by using the standardized sample mean. Construct the interval so that the left tail area left out is $\alpha_1$ and the right tail area left out is $\alpha_2$ so that $\alpha_1+\alpha_2 = \alpha$. Show that the length of the interval is shortest when $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

**12.6.2.** Let $x_1,\dots,x_n$ be iid as $N(\mu,\sigma^2)$ where $\sigma^2$ is known. Construct a $100(1-\alpha)\%$ central confidence interval for $\mu$ by using the statistic $c_1x_1+\dots+c_nx_n$ where $c_1,\dots,c_n$ are known constants. Illustrate the result for $c_1 = 2$, $c_2 = -3$, $c_3 = 5$ and based on the observed sample $2, 1, -5$.

**12.6.3.** Construct (1) a 90%, (2) a 95%, (3) a 99% central confidence interval for $\mu$ in Exercise 12.6.1 with $\sigma^2 = 2$ and based on the observed sample $-1, 2, 5, 7$.

**12.6.4.** Do the same Exercise 12.6.3 if $\sigma^2$ is unknown.

**12.6.5.** Compute the expected length in the central interval for the parameter $\mu$ in a $N(\mu, \sigma^2)$, where $\sigma^2$ is unknown, and based on a Student-$t$ statistic.

**12.6.6.** Compute the expected length as in Exercise 12.6.5 if the interval is obtained by cutting off the areas $\alpha_1$ at the left tail and $\alpha_2$ at the right tail. Show that the expected length is least when $\alpha_1 = \alpha_2$.

**12.6.7.** Construct a 95% central confidence interval for $\mu$ in a $N(\mu, \sigma^2)$, when $\sigma^2$ is unknown, by using the statistic $u = 2x_1 + x_2 - 5x_3$, and based on the observed sample $5, -2, 6$.

**12.6.8.** By using the standard normal approximation for a standardized binomial variable construct a 90% confidence interval (central) for $p$ the probability of success if (1) 7 successes are obtained in 20 trials; (2) 12 successes are obtained in 22 trials.

**12.6.9.** The grades obtained by students in a statistics course are assumed to be normally distributed with mean value $\mu$ and variance $\sigma^2$. Construct a 95% confidence interval for $\sigma^2$ when (1) $\mu = 80$, (2) $\mu$ is unknown, based on the following observed sample: $75, 85, 90, 90$; (a) Consider central intervals, (b) Consider cutting off 0.5 at the right tail.

**12.6.10.** Show that for the problem of constructing confidence interval for $\sigma^2$ in a $N(\mu, \sigma^2)$, based on a pivotal quantity having a chi-square distribution, the central interval is not the shortest in expected length when the degrees of freedom is small.

## 12.7 Confidence intervals for linear functions of mean values

Here, we are mainly interested in situations of the following types: (1) A new drug is administered to lower blood pressure in human beings. A random sample of $n$ individuals is taken. Let $x_j$ be the blood pressure before administering the drug and $y_j$ be the blood pressure after administering the drug on the $j$-th individual, for $j = 1, \dots, n$. Then we have paired values $(x_j, y_j)$, $j = 1, \dots, n$. Our aim may be to estimate the expected difference, namely $\mu_2 - \mu_1$, $\mu_2 = E(y_j)$, $\mu_1 = E(x_j)$ and test a hypothesis that $(x_j, y_j)$, $j = 1, \dots, n$ are identically distributed. But obviously, $y$ = the blood pressure after administering the drug depends on $x$ = the blood pressure before administering the drug. Here, $x$ and $y$ are dependent variables and may have a joint distribution. (2) A sample of $n_1$ test plots are planted with corn variety 1 and a sample of $n_2$ test plots are planted with corn variety 2. Let $x_1, \dots, x_{n_1}$ be the observations on the yield $x$ of corn variety 1 and let $y_1, \dots, y_{n_2}$ be the observations on the yield $y$ of corn variety 2. Let the test plots be homogeneous in all respects. Let $E(x) = \mu_1$ and $E(y) = \mu_2$. Someone may have a claim that the expected yield of variety 2 is 3 times that of variety 1. Then our aim may be to estimate $\mu_2 - 3\mu_1$. If someone has the claim that variety 2 is better than variety 1, then our aim may be to estimate $\mu_2 - \mu_1$. In this example, without loss of generality, we may assume that $x$ and $y$ are independently distributed. (3) A random sample of $n_1$ students of the same background are subjected to method 1 of teaching (consisting of lectures followed by one final examination), and a random sample of $n_2$ students of the same background, as of the first set of students, are subjected to method 2 of teaching (may be consisting of each lecture followed by problem sessions and three cumulative tests). Our aim may be to claim that method 2 is superior to method 1. Let $\mu_2 = E(y)$, $\mu_1 = E(x)$ where $x$ and $y$ represent the grades under method 1 and method 2, respectively. Then we may want to estimate $\mu_2 - \mu_1$. Here also, it can be assumed that $x$ and $y$ are independently distributed. (3) Suppose that a farmer has planted 5 varieties of paddy (rice). Let the yield per test plot of the 5 varieties be denoted by $x_1, \dots, x_5$ with $\mu_i = E(x_i)$, $i = 1, \dots, 5$. The market prices of these varieties are respectively Rs 20, Rs 25, Rs 30, Rs 32, Rs 38 per kilogram. Then the farmer's interest may be to estimate the money value, that is, $20\mu_1 + 25\mu_2 + 30\mu_3 + 32\mu_4 + 38\mu_5$. Variety $i$ may be planted in $n_i$ test plots so that the yields are $x_{ij}$, $j = 1, \dots, n_i$, $i = 1, \dots, 5$, where $x_{ij}$ is the yield of the $j$-th test plot under variety $i$.

Problems of the above types are of interest in this section. We will consider only situations involving two variables. The procedure is exactly parallel when more variables are involved. In the two variables case also, we will look at situations where the variables are dependent in the sense of having a joint distribution, and situations where the variables are assumed to be statistically independently distributed in the sense of holding product probability property will be considered later.

### 12.7.1 Confidence intervals for mean values when the variables are dependent

When we have paired variables $(x, y)$, where $x$ and $y$ are dependent, then one way of handling the situation is to consider $u = y - x$, in situations such as blood pressure before administering the drug $(x)$ and blood pressure after administering the drug $(y)$, if we wish to estimate $\mu_2 - \mu_1 = E(y) - E(x)$. If we wish to estimate a linear function $a\mu_1 + b\mu_2$, then consider the function $u = ax + by$. For example, $a = -1$ and $b = 1$ gives $\mu_2 - \mu_1$. When $(x, y)$ has a bivariate normal distribution then it can be proved that every linear function is univariate normal. That means, $u \sim N(\tilde{\mu}, \tilde{\sigma}^2)$ where $\tilde{\mu} = a\mu_1 + b\mu_2$ and $\tilde{\sigma}^2 = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\,\mathrm{Cov}(x, y)$, $\sigma_1^2 = \mathrm{Var}(x)$, $\sigma_2^2 = \mathrm{Var}(y)$. Now, construct confidence intervals for the mean value of $u$ in situations where (1) $\mathrm{Var}(u)$ is known, (2) $\mathrm{Var}(u)$ is unknown, and confidence intervals for $\mathrm{Var}(u)$ for the cases when (1) $E(u)$ is known, (2) $E(u)$ is unknown, by using the procedures in Section 12.5. Note that we need not know about the individual parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ and $\mathrm{Cov}(x, y)$ in this procedure.

**Note 12.7.** Many books may proceed with the assumption that $x$ and $y$ are independently distributed, in situations like blood pressure example, claiming that the effect of the drug is washed out after two hours or dependency is gone after two hours. Assuming statistical independence in such situations is not a proper procedure. When paired values are available we can handle by using $u$ as described above, which is a correct procedure when the joint distribution is normal. If the joint distribution is not normal, then we may evaluate the distribution of a linear function first and then use a linear function to construct confidence intervals for linear functions for mean values.

**Example 12.8.** The following are the paired observations on $(x,y) = (1,4), (4,8), (3,6), (2,7)$ where $x$ is the amount of a special animal feed and $y$ is the gain in weight. It is conjectured that $y$ is approximately $3x + 1$. Construct a 95% confidence interval for (1) $E(u) = E[y - (3x+1)] = \mu_2 - 3\mu_1 - 1$, $E(y) = \mu_2$, $E(x) = \mu_1$, (2) variance of $u$, assuming that $(x,y)$ has a bivariate normal distribution.

**Solution 12.8.** Let $u = y - 3x - 1$. Then the observations on $u$ are the following:

$$u_1 = 4 - 2(1) - 1 = 1, \quad u_2 = 8 - 2(4) - 1 = -1, \quad u_3 = 6 - 2(3) - 1 = -1,$$

$$u_4 = 7 - 2(2) - 1 = 2, \quad \bar{u} = \frac{1}{4}(1 - 1 - 1 + 2) = \frac{1}{4}$$

$$s_1^2 = \sum_{j=1}^{n} \frac{(u_j - \bar{u})^2}{n-1};$$

$$\text{Observed value} = \frac{1}{3}\left[\left(1 - \frac{1}{4}\right)^2 + \left(-1 - \frac{1}{4}\right)^2 + \left(-1 - \frac{1}{4}\right)^2 + \left(2 - \frac{1}{4}\right)^2\right] = \frac{108}{16 \times 3}.$$

$$\frac{\sqrt{n}[\bar{u} - E(\bar{u})]}{s_1} \sim t_{n-1} = t_3 \tag{12.19}$$

is Student-$t$ with 3 degrees of freedom. [Since all linear functions of normal variables (correlated or not) are normally distributed, $u$ is $N(\mu, \sigma^2)$ where $\mu = E(u)$, $\sigma^2 = \mathrm{Var}(u)$.] $t_{n-1,\frac{\alpha}{2}} = t_{3,0.025} = 3.182$ from Student-$t$ tables (see the illustration in Figure 12.3). Hence a 95% central confidence interval for $E(u) = \mu_2 - 3\mu_1 - 1$ is the following:

$$\left[\bar{u} - t_{n-1,\frac{\alpha}{2}} \frac{s_1}{\sqrt{n}}, \bar{u} + t_{n-1,\frac{\alpha}{2}} \frac{s_1}{\sqrt{n}}\right] = \left[\frac{1}{4} - 3.182 \frac{\sqrt{108}}{4(\sqrt{12})}, \frac{1}{4} + 3.182 \frac{\sqrt{108}}{4(\sqrt{12})}\right]$$
$$= [-2.14, 2.64].$$

For constructing a 95% confidence interval for $\mathrm{Var}(u)$, one can take the pivotal quantity as

$$\sum_{j=1}^{n} \frac{(u_j - \bar{u})^2}{\sigma^2} \sim \chi_{n-1}^2 = \chi_3^2; \quad \chi_{3,0.025}^2 = 9.35, \quad \chi_{3,0.975}^2 = 0.216.$$

See the illustration of the percentage points in Figure 12.4. Then a 95% central confidence interval is given by the following:

$$\left[\frac{\sum_{j=1}^{n}(u_j-\bar{u})^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{\sum_{j=1}^{n}(u_j-\bar{u})^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right] = \left[\frac{108}{16(9.35)}, \frac{108}{16(0.216)}\right]$$
$$= [0.72, 31.25].$$

**Note 12.8.** Note that in the paired variable $(x,y)$ case if our interest is to construct a confidence interval for $\mu_2-\mu_1$ then take $u=y-x$ and proceed as above. Whatever be the linear function of $\mu_1$ and $\mu_2$, for which a confidence interval is needed, take the corresponding linear function of $x$ and $y$ as $u$ and then proceed. Do not assume statistical independence of $x$ and $y$ unless there is theoretical justification to do so.

### 12.7.2 Confidence intervals for linear functions of mean values when there is statistical independence

If $x$ and $y$ are statistically independently distributed with $E(x)=\mu_1$, $E(y)=\mu_2$, $\mathrm{Var}(x)=\sigma_1^2$, $\mathrm{Var}(y)=\sigma_2^2$ and if simple random samples of sizes $n_1$ and $n_2$ are available from $x$ and $y$, then how can we set up confidence intervals for $a\mu_1+b\mu_2+c$ where $a,b,c$ are known constants? Let $x_1,\dots,x_{n_1}$ and $y_1,\dots,y_{n_2}$ be the samples from $x$ and $y$, respectively. If $x$ and $y$ are normally distributed then the problem is easy, otherwise one has to work out the distribution of the linear function first and then proceed. Let us assume that $x\sim N(\mu_1,\sigma_1^2)$, $y\sim N(\mu_2,\sigma_2^2)$ and be independently distributed. Let

$$\bar{x}=\frac{\sum_{j=1}^{n_1}x_j}{n_1},\quad \bar{y}=\frac{\sum_{j=1}^{n_2}y_j}{n_2},\quad v_1^2=\sum_{j=1}^{n_1}(x_j-\bar{x})^2,\quad v_2^2=\sum_{j=1}^{n_2}(y_j-\bar{y})^2 \tag{12.20}$$

and $u=a\bar{x}+b\bar{y}+c$. Then $u\sim N(\mu,\sigma^2)$, where

$$\begin{aligned}\mu&=E(u)=aE[\bar{x}]+bE[\bar{y}]+c=a\mu_1+b\mu_2+c\\ \sigma^2&=\mathrm{Var}(u)=\mathrm{Var}(a\bar{x}+b\bar{y}+c)=\mathrm{Var}(a\bar{x}+b\bar{y})\\ &=a^2\,\mathrm{Var}(\bar{x})+b^2\,\mathrm{Var}(\bar{y})\end{aligned}$$

since $\bar{x}$ and $\bar{y}$ are independently distributed

$$\sigma^2=a^2\frac{\sigma_1^2}{n_1}+b^2\frac{\sigma_2^2}{n_2}.$$

Our interest here is to set up confidence intervals for $a\mu_1+b\mu_2+c$. A usual situation may be to set up confidence intervals for $\mu_2-\mu_1$. In that case, $c=0$, $b=1$, $a=-1$. Various situations are possible.

**Case 1** ($\sigma_1^2$ and $\sigma_2^2$ are known)**.** In this case, we can take the pivotal quantity as the standardized $u$. That is,

$$\frac{u-E(u)}{\sqrt{\mathrm{Var}(u)}}=\frac{u-(a\mu_1+b\mu_2+c)}{\sqrt{a^2\frac{\sigma_1^2}{n_1}+b^2\frac{\sigma_2^2}{n_2}}}\sim N(0,1). \tag{12.21}$$

Hence a $100(1-\alpha)\%$ central confidence interval for $a\mu_1 + b\mu_2 + c$ is the following:

$$\left[u - z_{\frac{\alpha}{2}}\sqrt{a^2\frac{\sigma_1^2}{n_1} + b^2\frac{\sigma_2^2}{n_2}}, u + z_{\frac{\alpha}{2}}\sqrt{a^2\frac{\sigma_1^2}{n_1} + b^2\frac{\sigma_2^2}{n_2}}\right] \tag{12.22}$$

where $z_{\frac{\alpha}{2}}$ is illustrated in Figure 12.2.

**Case 2** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ = unknown)**.** In this case, the population variances are given to be equal but it is unknown. In that case, we can use a Student-$t$ statistic. Note from (12.20) that $E[v_1^2] = (n_1 - 1)\sigma_1^2$ and $E[v_2^2] = (n_2 - 1)\sigma_2^2$, and hence when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then $E[v_1^2 + v_2^2] = (n_1 + n_2 - 2)\sigma^2$ or

$$E[v^2] = E\left[\frac{(\sum_{j=1}^{n_1}(x_j - \bar{x})^2 + \sum_{j=1}^{n_2}(y_j - \bar{y})^2)}{n_1 + n_2 - 2}\right] = \sigma^2. \tag{12.23}$$

Hence $\hat{\sigma}^2 = v^2$ can be taken as an unbiased estimator of $\sigma^2$. In the standardized normal variable if we replace $\sigma^2$ by $\hat{\sigma}^2$ , then we should get a Student-$t$ with $n_1 + n_2 - 2$ degrees of freedom because the corresponding chi-square has $n_1 + n_2 - 2$ degrees of freedom. Hence the pivotal quantity that we will use is the following:

$$\frac{(a\bar{x} + b\bar{y} + c) - (a\mu_1 + b\mu_2 + c)}{\hat{\sigma}\sqrt{\frac{a^2}{n_1} + \frac{b^2}{n_2}}} = \frac{(a\bar{x} + b\bar{y} + c) - (a\mu_1 + b\mu_2 + c)}{v\sqrt{\frac{a^2}{n_1} + \frac{b^2}{n_2}}}$$

$$\sim t_{n_1+n_2-2} \tag{12.24}$$

where $v$ is defined in (12.23). Now a $100(1-\alpha)\%$ central confidence interval for $a\mu_1 + b\mu_2 + c$ is given by

$$\left[(a\bar{x} + b\bar{y} + c) \mp t_{n_1+n_2-2,\frac{\alpha}{2}} v\sqrt{\frac{a^2}{n_1} + \frac{b^2}{n_2}}\right]. \tag{12.25}$$

The percentage point $t_{n_1+n_2-2,\frac{\alpha}{2}}$ is available from Figure 12.3 and $v$ is available from (12.23). If the confidence interval for $\mu_2 - \mu_1$ is needed, then put $c = 0$, $b = 1$, $a = -1$ in (12.25).

**Case 3** ($\sigma_1^2$ and $\sigma_2^2$ are unknown but $n_1 \geq 30$, $n_2 \geq 30$)**.** In this case, one may use the following approximation to standard normal for setting up confidence intervals.

$$\frac{(a\bar{x} + b\bar{y} + c) - (a\mu_1 + b\mu_2 + c)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1) \tag{12.26}$$

approximately, where $s_1^2 = \sum_{j=1}^{n_1}\frac{(x_j-\bar{x})^2}{n_1}$, $s_2^2 = \sum_{j=1}^{n_2}\frac{(y_j-\bar{y})^2}{n_2}$ are the sample variances. When $n_1$ and $n_2$ are large, dividing by $n_i$ or $n_i - 1$ for $i = 1, 2$ will not make a difference. Then the approximate $100(1-\alpha)\%$ central confidence interval for $a\mu_1 + b\mu_2 + c$ is given by

$$(a\bar{x} + b\bar{y} + c) \mp z_{\frac{\alpha}{2}} \sqrt{\frac{a^2 s_1^2}{n_1} + \frac{b^2 s_2^2}{n_2}} \tag{12.27}$$

where the percentage point $z_{\frac{\alpha}{2}}$ is available from the standard normal density in Figure 12.2.

### 12.7.3 Confidence intervals for the ratio of variances

Here again, we consider two independently distributed normal variables $x \sim N(\mu_1, \sigma_1^2)$ and $y \sim N(\mu_2, \sigma_2^2)$ and simple random samples of sizes $n_1$ and $n_2$ from $x$ and $y$, respectively. We would like to construct a $100(1-\alpha)\%$ confidence interval for $\theta = \frac{\sigma_1^2}{\sigma_2^2}$. We will make use of the property that

$$\begin{aligned}
\frac{\sum_{j=1}^{n_1}(x_j - \bar{x})^2}{\sigma_1^2} &\sim \chi^2_{n_1-1} \\
\frac{\sum_{j=1}^{n_2}(y_j - \bar{y})^2}{\sigma_2^2} &\sim \chi^2_{n_2-1} \\
u\left(\frac{1}{\theta}\right) &= \frac{[\sum_{j=1}^{n_1}(x_j - \bar{x})^2/(n_1-1)]}{[\sum_{j=1}^{n_2}(y_j - \bar{y})^2/(n_2-1)]}\left(\frac{1}{\theta}\right) \\
&\sim F_{n_1-1,n_2-1}.
\end{aligned} \tag{12.28}$$

From this, one can make the following probability statement:

$$\Pr\left\{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} \le u\left(\frac{1}{\theta}\right) \le F_{n_1-1,n_2-1,\frac{\alpha}{2}}\right\} = 1-\alpha.$$

Rewriting this as a statement on $\theta$, we have

$$\Pr\left\{\frac{u}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}} \le \theta \le \frac{u}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}\right\} = 1-\alpha \tag{12.29}$$

where the percentage points $F_{n_1-1,n_2-1,\frac{\alpha}{2}}$ and $F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ are given in Figure 12.5, and

$$u = \frac{[\sum_{j=1}^{n_1}(x_j - \bar{x})^2/(n_1-1)]}{[\sum_{j=1}^{n_2}(y_j - \bar{y})^2/(n_2-1)]} \sim \theta F_{n_1-1,n_2-1}. \tag{12.30}$$

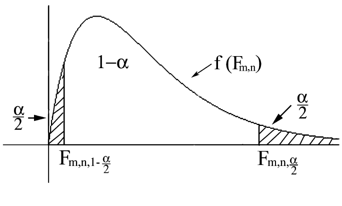


**Figure 12.5:** Percentage points from a $F$-density.

**Note 12.9.** If confidence intervals for $a\frac{\sigma_1^2}{\sigma_2^2} = a\theta$, where $a$ is a constant, is needed then multiply and divide $u$ in (12.28) by $a$, absorb the denominator $a$ with $\theta$ and proceed to get the confidence intervals from (12.29). Also note that only the central interval is considered in (12.29).

**Note 12.10.** Since $F$-random variable has the property that $F_{m,n} = \frac{1}{F_{n,m}}$ we can convert the lower percentage point $F_{m,n,1-\alpha/2}$ to an upper percentage point on $F_{n,m,\alpha/2}$. That is,

$$F_{m,n,1-\frac{\alpha}{2}} = \frac{1}{F_{n,m,\frac{\alpha}{2}}}. \tag{12.31}$$

Hence usually the lower percentage points are not given in $F$-tables.

**Example 12.9.** Nine test plots of variety 1 and 5 test plots of variety 2 of tapioca gave the following summary data: $s_1^2 = 10\,\text{kg}$ and $s_2^2 = 5\,\text{kg}$, where $s_1^2$ and $s_2^2$ are the sample variances. The yield $x$ under variety 1 is assumed to be distributed as $N(\mu_1, \sigma_1^2)$ and the yield $y$ of variety 2 is assumed to be distributed as $N(\mu_2, \sigma_2^2)$ and independently of $x$. Construct a 90% confidence interval for $3\frac{\sigma_1^2}{\sigma_2^2}$.

**Solution 12.9.** We want to construct a 90% confidence interval and hence in our notation, $\alpha = 0.10$, $\frac{\alpha}{2} = 0.05$. The parameter of interest is $3\theta = 3\frac{\sigma_1^2}{\sigma_2^2}$. Construct interval for $\theta$ and then multiply by 3. Hence the required statistic, in observed value, is

$$\begin{aligned} u &= \frac{[\sum_{j=1}^{n_1}(x_j - \bar{x})^2/(n_1 - 1)]}{[\sum_{j=1}^{n_2}(y_j - \bar{y})^2/(n_2 - 1)]} \\ &= \frac{[9s_1^2/(8)]}{[5s_2^2/(4)]} \sim F_{8,4} \quad \text{and in observed value} \\ &= \left[\frac{(9)(10)}{8}\right] \Big/ \left[\frac{(5)(5)}{4}\right] = \frac{9}{5}. \end{aligned}$$

From $F$-tables, we have $F_{8,4,0.05} = 6.04$ and $F_{4,8,0.05} = 3.84$. Hence a 90% central confidence interval for $3\theta$ is given by

$$\begin{aligned} \left[\frac{27}{5(F_{8,4,0.05})}, \frac{27}{5(F_{8,4,0.95})}\right] &= \left[\frac{27}{5(F_{8,4,0.05})}, \frac{27(F_{4,8,0.05})}{5}\right] \\ &= \left[\frac{27}{5(6.04)}, \frac{27(3.84)}{5}\right] = [0.89, 20.74]. \end{aligned}$$

**Note 12.11** (Confidence regions). In a population such as gamma (real scalar random variable), there are usually two parameters, the scale parameter $\beta$, $\beta > 0$ and the shape parameter $\alpha$, $\alpha > 0$. If relocation of the variable is involved, then there is an additional location parameter $\gamma$, $-\infty < \gamma < \infty$. In a real scalar normal population $N(\mu,\sigma^2)$, there are two parameters $\mu$, $-\infty < \mu < \infty$ and $\sigma^2$, $0 < \sigma^2 < \infty$. The parameter spaces in the 3-parameter gamma density is

$$\Omega = \{(\alpha,\beta,\gamma) \mid 0 < \alpha < \infty, 0 < \beta < \infty, -\infty < \gamma < \infty\}.$$

In the normal case, the parameter space is $\Omega = \{(\mu,\sigma^2) \mid -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$. Let $\theta = (\theta_1,\ldots,\theta_s)$ represent the set of all parameters in a real scalar population. In the above gamma case, $\theta = (\alpha,\beta,\gamma)$, $s = 3$ and in the above normal case $\theta = (\mu,\sigma^2)$, $s = 2$. We may be able to come up with a collection of one or more functions of the sample values $x_1,\ldots,x_n$ and some of the parameters from $\theta$, say, $P = (P_1,\ldots,P_r)$ such that the joint distribution of $P$ is free of all parameters in $\theta$. Then we will be able to make a statement of the type

$$\Pr\{P \in R_1\} = 1 - \alpha \tag{12.32}$$

for a given $\alpha$, where $R_1$ is a subspace of $R^r = R \times R \times \cdots \times R$ where $R$ is the real line. If we can convert this statement into a statement of the form

$$\Pr\{S_1 \text{ covers } \theta\} = 1 - \alpha \tag{12.33}$$

where $S_1$ is a subspace of the sample space $S$, then $S_1$ is the confidence region for $\theta$. Since computations of confidence regions will be more involved, we will not be discussing this topic further.

## Exercises 12.7

**12.7.1.** In a weight reduction experiment, a random sample of 5 individuals underwent a certain dieting program. The weight of a randomly selected person, before the program started, is $x$ and when the program is finished it is $y$. $(x,y)$ is assumed to have a bivariate normal distribution. The following are the observations on $(x,y)$: $(80,80), (90,85), (100,80), (60,55), (65,70)$. Construct a 95% central confidence interval for (a) $\mu_1 - \mu_2$, when (1) variance of $x - y$ is 4, (2) when the variance of $x - y$ is unknown; (b) $0.2\mu_1 - \mu_2$ when (1) variance of $u = 0.2x - y$ is known to be 5, (2) variance of $u$ is unknown.

**12.7.2.** Two methods of teaching are experimented on sets of $n_1 = 10$ and $n_2 = 15$ students. These students are assumed to have the same backgrounds and are independently selected. If $x$ and $y$ are the grades of randomly selected students under the two methods, respectively, and if $x \sim N(\mu_1,\sigma_1^2)$ and $y \sim N(\mu_2,\sigma_2^2)$ construct 90% confidence intervals for (a) $\mu_1 - 2\mu_2$ when (1) $\sigma_1^2 = 2$, $\sigma_2^2 = 5$, (2) $\sigma_1^2 = \sigma_2^2$ but unknown; (b) $2\sigma_1^2/\sigma_2^2$ when (1) $\mu_1 = -10$, $\mu_2 = 5$, (2) $\mu_1, \mu_2$ are unknown. The following summary statistics are given, with the usual notations: $\bar{x} = 90$, $\bar{y} = 80$, $s_1^2 = 25$, $s_2^2 = 10$.

**12.7.3.** Consider the same problem as in Exercise 12.6.2 with $n_1 = 40$, $n_2 = 50$ but $\sigma_1^2$ and $\sigma_2^2$ are unknown. Construct a 95% confidence interval for $\mu_1 - \mu_2$, by using the same summary data as in Exercise 12.7.2.

**12.7.4.** Prove that $F_{m,n,1-\alpha} = \frac{1}{F_{n,m,\alpha}}$.

**12.7.5.** Let $x_1, \dots, x_n$ be iid variables from some population (discrete or continuous) with mean value $\mu$ and variance $\sigma^2 < \infty$. Use the result that

$$\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0, 1)$$

approximately for large $n$, and set up a $100(1-\alpha)\%$ confidence interval for $\mu$ when $\sigma^2$ is known.

**12.7.6.** The temperature reading $x$ at location 1 and $y$ at location 2 gave the following data. A simple random sample of size $n_1 = 5$ on $x$ gave $\bar{x} = 20c$ and $s_1^2 = 5c$, and a random sample of $n_2 = 8$ on $y$ gave $\bar{y} = 30c$ and $s_2^2 = 8c$. If $x \sim N(\mu_1, \sigma_1^2)$ and $y \sim N(\mu_2, \sigma_2^2)$ and independently distributed then construct a 90% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$.

# 13 Tests of statistical hypotheses

## 13.1 Introduction

People, organizations, companies, business firms, etc. make all sorts of claims. A business establishment producing a new exercise routine may claim that if someone goes through that routine the expected weight reduction will be 10 kilograms. If $\theta$ is the expected reduction of weight, then the claim here is $\theta = 10$. A coaching centre may claim that if a student goes through their coaching scheme, then the expected grade in the national test will be more than 90%. If $\mu$ is the expected grade under their coaching scheme, then the claim is that $\mu > 90$. A bird watcher may claim that birds on the average lay more eggs in Tamilnadu compared to Kerala. If the expected number of eggs per bird nest in Tamilnadu and Kerala are respectively $\mu_1$ and $\mu_2$, then the claim is $\mu_1 > \mu_2$. A tourist resort operator in Kerala may claim that the true proportion of tourists from outside Kerala visiting his resort is 0.9. If the probability of finding a tourist from outside Kerala in this resort is $p$, then the claim is that $p = 0.9$. An economist may claim that the incomes in community 1 is more spread out compared to the income in community 2. If the spreads are denoted by the standard deviations $\sigma_1$ and $\sigma_2$, then the claim is that $\sigma_1 > \sigma_2$. An educationist may claim that the grades obtained by students in a particular course follow a bell curve. If the typical grade is denoted by $x$, then the claim is that $x$ is normally distributed if a traveler claims that the travel time needed to cover 5 kilometers in Ernakulam during peak traffic time is longer than the time needed in Trivandrum. Here, the claim is that one random variable is bigger than another random variable over a certain interval. In all the above examples, we are talking about quantitative characteristics. If a sociologist claims that the tendency for the destruction of public properties by students and membership in a certain political party are associated, then the claim is about the association between two qualitative characteristics. If an engineer claims that in a certain production process the occurrence of defective items (items which do not have quality specifications) is a random phenomenon and not following any specific pattern then we want to test the randomness of this event. If the villagers claim that snake bite occurring in a certain village follows a certain pattern, then we may want to test for that pattern or the negation that there is no pattern or the event is a random event. If a biologist claims that the chance of finding a Rosewood tree in Kerala forest is much less than that in Karnataka forest, then the claim may be of the form $p_1 < p_2$, where $p_1$ and $p_2$ are the respective probabilities. If a physicist claims that every particle attracts every other particle, then we classify this hypothesis as a conjecture if "attraction" is properly defined. This conjecture can be disproved if two particles are found not having attraction. If a religious preacher claims that the only way to go to heaven is through his religion, then we will not classify this as a hypothesis because there are several undefined or not precisely defined items such as "heaven" and the

https://doi.org/10.1515/9783110562545-013

method of reaching there, etc. There is no way of collecting data and verifying the claim.

We are looking at hypotheses where one can take data on some observable random variables and test the claims, or test the hypothesis by using some statistical procedures or we are looking for some verifiable or testable types of claims.

We have looked into various types of claims or hypotheses above. Out of these, the one about heaven has un-identifiable terms, and hence we do not count it as a hypothesis, and we cannot collect data either, to verify or test the claim. We will classify hypotheses into two types: *statistical* and *non-statistical*. A statistical hypothesis has something to say about the behavior of one or more random variables. If the hypothesis is well-defined but no random phenomenon is involved, then we call such hypotheses as non-statistical hypotheses. Many of the physical laws or mathematical conjectures are non-statistical hypotheses.

Statistical hypothesis
A hypothesis ↗↘
Non-statistical hypothesis

In this chapter, we are concerned with statistical hypotheses only. Out of the many statistical hypotheses described above, we have noted that some of them are dealing with parameters of well-defined distributions and others are of the type of testing hypotheses on qualitative characteristics, some are about randomness of phenomena, some are about certain patterns, etc. If a hypothesis is about the parameter(s) of well-defined distributions, then we call them *parametric hypotheses*. All other statistical hypotheses will be called *non-parametric hypotheses*.

Parametric hypothesis
A statistical hypothesis ↗↘
Non-parametric hypothesis

First, we will concentrate on parametric statistical hypotheses and later we will deal with non-parametric hypotheses.

## 13.2 Testing a parametric statistical hypothesis

We may test a parametric hypothesis of the type that the expected waiting time in a service station for servicing a car is greater than or equal to 40 minutes, then this hypothesis is of the type $\theta \geq 40$ where $\theta$ is expected waiting time, the waiting time may have an exponential distribution with expected value $\theta$. Then the hypothesis is on a parameter of a well-defined distribution. For testing a statistical hypothesis, we may take some data from that population, then use some test criteria and make a decision either to reject or not to reject that hypothesis. If the decision is to reject or not to reject, then we have a two-decision problem. If our decision is of the form, reject, not to reject, take more observations because a decision cannot be reached with the available observations, then it is a three-decision problem. Thus we may have a multiple decision problem in any given parametric statistical hypothesis. First, we will consider a two-decision situation. Here also, there are several possibilities. We have a hypothesis that is being tested and the natural alternate against which the hypothesis is tested. If we are testing the hypothesis that $\theta \geq 10$, then we are naturally testing it against the alternate that $\theta < 10$. If we test the hypothesis $\theta = 20$, then we are testing it against its natural alternate $\theta \neq 20$.

**Definition 13.1** (Null and alternate hypotheses). A hypothesis that is being tested is called the *null hypothesis* and it is usually denoted by $H_0$. The alternate hypothesis, against which the null hypothesis $H_0$ is tested, is called the *alternate hypothesis* and it is usually denoted by $H_1$ or $H_A$. We will use the notation $H_1$.

A parametric hypothesis ↗ A null parametric hypothesis $H_0$
A parametric hypothesis ↘ An alternate parametric hypothesis $H_1$

The term "null" came due to historical reasons. Originally, the claims that were tested were of the type that there is significant difference between two quantitative measurements such as the yield of corn without using fertilizers and with the use of fertilizers. The hypothesis is usually formulated as there is no significant difference (hypothesis of the type $H_0 : \mu_1 = \mu_2$) between expected yields and is tested against the hypothesis that the difference is significant (hypothesis of the type $\mu_1 \neq \mu_2$). Nowadays the term "null" is used to denote the hypothesis that is being tested whatever be the nature of the hypothesis.

We may also have the possibility that once the hypothesis (null or alternate) is imposed on the population the whole population may be fully known, in the sense of no unknown parameters remaining in it or the population may not be fully known. If the population is exponential with parameter $\theta$ and if the hypothesis is $H_0 : \theta = 20$, then when the hypothesis is imposed on the density there are no more parameters left and the density is fully known. In this case, we call $H_0$ a "simple hypothesis". The alternate in this case is $H_1 : \theta \neq 20$. Then under this alternate, there are still a lot of values possible for $\theta$, and hence the population is not determined. In such a case, we call it a "composite" hypothesis.

**Definition 13.2** (Simple and composite hypotheses). Once the hypothesis is imposed on the population if the population is fully known, then the hypothesis is called a simple hypothesis and if some unknown quantities are still left or the population is not fully known, then that hypothesis is called a composite hypothesis.

A simple hypothesis
A parametric hypothesis ↗↘
A composite hypothesis

Thus, a null parametric hypothesis can be simple or composite and similarly an alternate parametric hypothesis can be simple or composite. For example, let us take a normal population $N(\mu, \sigma^2)$. There are two parameters $\mu$ and $\sigma^2$ here. Let us consider the following null and alternate hypotheses:

$H_0 : \mu = 0, \sigma^2 = 1$ (simple), alternate $H_1 : \mu \neq 0, \sigma^2 \neq 1$ (composite);
$H_0 : \mu = 0$ (composite), alternate $H_1 : \mu \neq 0$ (composite), $\sigma^2$ is unknown;
$H_0 : \mu \leq 5, \sigma^2 = 1$ (composite), alternate $H_1 : \mu > 5, \sigma^2 = 1$ (composite);
$H_0 : \mu = 0, \sigma^2 \leq 4$ (composite), alternate $H_1 : \mu = 0, \sigma^2 > 4$ (composite).

The simplest testing problem that can be handled is a simple null hypothesis versus a simple alternate. But before starting the testing procedure we will examine more details. When a decision is taken, after testing $H_0$, either to reject or not to reject $H_0$, we have the following possibilities. The hypothesis itself may be correct or not correct. Our decision may be correct or wrong. If a testing procedure gave a decision not to reject the hypothesis that does not mean that the hypothesis is correct. If the expected waiting time in a queue, $\theta$, is hypothesized as $H_0 : \theta = 10$ minutes, this does not mean that in fact the expected waiting time is 10 minutes. Our hypothesis may be different from the reality of the situation. We have one of the following four possibilities in a given testing situation:

Hypothesis $H_0$
↙↘

| | | $H_0$ is true | $H_0$ is not true |
|---|---|---|---|
| Decision ↗↘ | Reject $H_0$ | Type-I error | Correct decision |
| | Not reject $H_0$ | Correct decision | Type-II error |

There are two situations of correct decision and two situations of wrong decision. The error committed in rejecting the null hypothesis $H_0$ when it is in fact true is called type-I error and the error of not rejecting when $H_0$ itself is not true is called type-II error. The probabilities of committing these two errors are denoted by $\alpha$ and $\beta$. That is,

$$\alpha = \Pr\{\text{reject } H_0 | H_0 \text{ is true}\}, \tag{13.1}$$

$$\beta = \Pr\{\text{not reject } H_0 | H_0 \text{ is not true}\}, \tag{13.2}$$

where a vertical bar indicates "given". As an example, consider the following: Suppose that we want to test the hypothesis $H_0 : \mu = 2$ in a normal population $N(\mu, 1)$. Here,

$\sigma^2$ is known to be one. Suppose that by using some procedure we have come up with the following test criterion: Take one observation from this normal population. Reject the null hypothesis if the observation is bigger than 3.75, otherwise not to reject $H_0$. Here, $\alpha$ is the probability of rejecting $H_0$ when it is true or the probability of rejecting $H_0 : \mu = 2$ when in fact the normal population is $N(\mu = 2, \sigma^2 = 1)$. Hence this probability is given by the following integral:

$$\begin{aligned}\alpha &= \int_{3.75}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x-2)^2} \mathrm{d}x \\ &= \int_{1.75}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y, \quad y = (x-\mu) = (x-2) \\ &= 0.04\end{aligned}$$

from standard normal tables. Then $\beta$ is the probability of not rejecting when $\mu \neq 2$. We do not reject when $x < 3.75$ as per our criterion. Then

$$\begin{aligned}\beta &= \int_{-\infty}^{3.75} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^2} \mathrm{d}x \\ &= \int_{-\infty}^{3.75-\mu} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y, \quad y = x - \mu\end{aligned}$$

which gives

$$1 - \beta = \int_{3.75-\mu}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y.$$

Note that $\beta$ as well as $1 - \beta$ can be read from standard normal tables for every given $\mu$. In general, $\beta$ is a function of $\mu$ or $\beta = \beta(\mu)$. For $\mu = 2$, then $1 - \beta = \alpha$ here.

We have seen that once a test criterion is given one can compute $\alpha$ = the probability of type-I error and $\beta$ = the probability of type-II error. But how to come up with a test criterion? If we can make $\alpha = 0$ and $\beta = 0$ and come up with a criterion then that is the best possible one. But we can see that we have a random situation or our problem is not a deterministic type, and hence making $\alpha$ and $\beta$ or one of them zero is not possible. Then is it possible to minimize both $\alpha$ and $\beta$ simultaneously and come up with a criterion? It can be seen that simultaneous minimization is not possible. This may be easily seen from the above illustrative example of $N(\mu, 1)$ and testing hypothesis $H_0 : \mu = 2$. Take a simple situation of a simple $H_0$ versus a simple $H_1$. Suppose that packets are filled with beans by an automatic filling machine. The machine is set for filling 2 kilograms (kg) per packet or the expected weight $\mu = 2\,\text{kg}$. If the machine is filling without cutting and chopping, then weight need not be exactly 2 kg but will be around 2 kg per packet. Suppose that the machine is operating in Pala where there is current fluctuation and stoppage almost every minute. At one of the sudden fluctuations, the machine setting went off and the machine started filling 2.1 kg instead of 2 kg. A shopkeeper bought some of these packets. She wants to know whether the packets belong to the 2 kg set or 2.1 kg set. These are the only possibilities. If the weight distribution is $N(\mu, 1)$, then $\mu$ can take only two values 2 or 2.1. If $H_0$ is $\mu = 2$, then naturally $H_1$ here is $\mu = 2.1$. Take our illustrative example for this simple $H_0$ versus simple $H_1$ case. Our criterion was to take one observation and if it is bigger than 3.75 reject the $H_0$. We can reduce $\alpha$ by shifting our rejection point or *critical point* to the right then we see that automatically $\beta$ is increased, and vice versa. Hence it is clear that simultaneous minimization of $\alpha$ and $\beta$ is not possible. Then what is the next best thing to do? We prefix either $\alpha$ or $\beta$ and minimize the other and then come up with a criterion. This procedure is possible. Then the question is which one to be pre-fixed? Which one is usually a more serious error? Suppose that a new drug is being introduced into the market for preventing a heart attack. The manufacturer's claim is that the drug can prevent a heart attack if administered within one hour of the appearance of symptoms of a heart attack. By a testing procedure, suppose that this claim is rejected and the drug is rejected. Suppose that the claim was correct. The net result is a loss of money for developing that drug. Suppose that the claim was not correct and the testing procedure did not reject the drug. The net result is that a lot of lives are lost. Hence usually a type-II error is more serious than a type-I error. Therefore, what is done is to prefix $\alpha$ and minimize $\beta$ or prefix $\alpha$ and maximize $1 - \beta$.

Let us see what is happening when we have a test criterion. If we have one observation $x_1$, then our sample space is the $x$-axis or the real line. If the test criterion says to reject $H_0$ if $x_1 \geq 5$, then the sample space $S$ is split into two regions $C \subset S$, where $C = \{x_1 \mid x_1 \geq 5\}$ and the complementary region $\bar{C} \subset S$ where $H_0$ is not rejected. If we have two observations $(x_1, x_2)$, then we have the plane. If the test criterion says to reject $H_0$ if the sample mean is greater than or equal to 1, then the rejection region in the sample space $S$ is $C = \{(x_1, x_2) \mid x_1 + x_2 \geq 2\} \subset S$. Figure 13.1 shows the illustration in (a), (b), and in (c) the general Venn diagrammatic representation of the sample space $S$ and the rejection region $C$ is given.

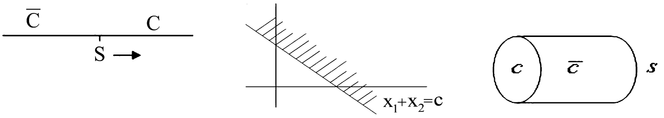


**Figure 13.1:** Critical regions.

**Definition 13.3** (Critical region and size and power of the critical region). The region $C$, $C \subset S$ where $S$ is the sample space, where the null hypothesis $H_0$ is rejected, is called the critical region. The probability of rejecting the null hypothesis when it is true is $\alpha$, that is, the sample point falls in the critical region when $H_0$ is true, or

$$\alpha = \Pr\{x \epsilon C|H_0\} \tag{13.3}$$

is called the size of the critical region or size of the test or size of the test criterion, where $x = (x_1, \dots, x_n)$ represents the sample point. Then $1 - \beta$ = the probability of rejecting $H_0$ when the alternative $H_1$ is true is called the power of the critical region $C$ or power of the test.

If $H_0$ is that $\theta \epsilon w$, $w \subset \Omega$, that is, $\theta$ belongs to the subset $w$ of the parameter space, then $H_1$ is that $\theta \epsilon \bar{w}$, where $\bar{w}$ is the complement of $w$ in $\Omega$. In this case, both $\alpha$ and $\beta$ will be functions of $\theta$. For example, for the normal population $N(\mu, \sigma^2 = 1)$, if the null hypothesis $H_0$ is $\mu \leq 5$ then $H_1$ is $\mu > 5$. In both of the cases, there are plenty of $\mu$ values present, and hence $\alpha = \alpha(\mu)$ and $\beta = \beta(\mu)$. We may write the above details in symbols as follows:

$$\begin{aligned}\alpha = \alpha(\theta) &= \Pr\{x \epsilon C|H_0\}, \quad x = (x_1, \dots, x_n), \quad H_0 : \theta \epsilon w \subset \Omega \\ &= \text{size of the critical region } C \qquad (13.4)\\ 1 - \beta = 1 - \beta(\theta) &= \Pr\{x \epsilon C|H_1\}, \quad H_1 : \theta \epsilon \bar{w} \subset \Omega \\ &= \text{power of the critical region } C. \qquad (13.5)\end{aligned}$$

**Definition 13.4** (Most powerful (MP) and uniformly most powerful (UMP) tests or critical regions). If $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ are simple, that is, the populations are fully known once the hypotheses are implemented, then there are only two points $\theta_0$ and $\theta_1$ in the parameter space. If $C$ is the critical region (or test) of size $\alpha$, which has more power compared to any other critical region (or test) of the same size $\alpha$, then $C$ is called the most powerful critical region (or test). If $H_0$ or $H_1$ or both are composite, then if $C$ is the critical region of size $\alpha(\theta)$ and has more power compared to any other critical region of the same size $\alpha(\theta)$, and for all values of $\theta \epsilon \bar{w} \subset \Omega$ then $C$ is called uniformly most powerful critical region (or test).

**Definition 13.5** (Power curve). If the power $1 - \beta(\theta)$ is drawn against $\theta$, assuming that $\theta$ is a scalar parameter, then the resulting curve is called the power curve.

In Figure 13.2, power curves are drawn for three critical regions $C, D, E$, all having the same size $\alpha$. For $\theta > \theta_0$, both $C$ and $D$ have more power compared to $E$. For $\theta < \theta_0$, both $C$ and $E$ have more power compared to $D$. For all $\theta \neq \theta_0$, $C$ has more power compared

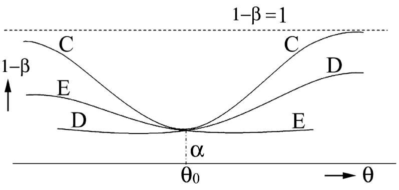


**Figure 13.2:** Power curves.

to $D$ and $E$. Thus $C$ is uniformly more powerful compared to $D$ and $E$. If there is a $C$, which is uniformly more powerful than any other critical region (or test) of the same size $\alpha$, then $C$ is uniformly the most powerful critical region (or test).

Thus, for constructing a test criterion our procedure will be the following: If it is the case of simple $H_0$ versus simple $H_1$, then look for the most powerful test (MPT). If $H_0$ or $H_1$ or both are composite, then look for uniformly most powerful test (UMPT). First, we will consider the situation of simple $H_0$ versus simple $H_1$. An example of this type was already discussed, of automatic filling of bags with beans, where the machine was set for the expected weight of the packet 2 kg but due to a machine setting changing to 2.1 kg unknowingly, due to power surge, the expected weight changes to 2.1 for 2.0. There are only two parameter values here 2.0 and 2.1. In the simple $H_0$ versus simple $H_1$, there is a small result which will give a procedure of constructing a test criterion for the most powerful test (MPT). This is called Neyman–Pearson lemma.

**Result 13.1** (Neyman–Pearson lemma). *Let $x = (x_1, \dots, x_n)$ be a sample point and let $L = L(x_1, \dots, x_n, \theta)$ = joint density or probability function of the sample point. Let $L_0$ be $L$ under $H_0$ and $L_1$ be $L$ under $H_1$. Let $H_0$ and $H_1$ be simple ($H_0$ will be of the form $\theta = \theta_0$ (given), and $H_1$ is of the form $\theta = \theta_1$ (given)). Let the population support be free of the parameter(s) $\theta$. Then the most powerful test (MPT) or the most powerful critical region $C$ is given by the rule*

$$\begin{aligned} &\frac{L_0}{L_1} \le k \quad \textit{inside } C \textit{ for some constant } k > 0 \\ &\frac{L_0}{L_1} > k \quad \textit{outside } C \end{aligned}$$

*then $C$ is the most powerful critical region.*

**Proof.** Let the size of the critical region $C$ be $\alpha$. Let $D$ be any other critical region of the same size $\alpha$. Then

$$\alpha = \Pr\{x \epsilon C|H_0\} = \Pr\{x \epsilon D|H_0\}.$$

We will give the proof in the continuous case, and for the discrete case the steps are parallel. In the discrete case, the size of the critical region is to be taken as $\le \alpha$ because when adding up the probabilities at individually distinct points we may not hit the exact value $\alpha$. Then take the closest value but $\le \alpha$. Then

$$\alpha = \int_C L_0 \mathrm{d}x = \int_{C\cap\bar{D}} L_0 \mathrm{d}x + \int_{C\cap D} L_0 \mathrm{d}x; \quad \alpha = \int_D L_0 \mathrm{d}x = \int_{C\cap D} L_0 \mathrm{d}x + \int_{\bar{C}\cap D} L_0 \mathrm{d}x, \qquad \text{(a)}$$

as shown in Figure 13.3, where $x = (x_1, \dots, x_n)$ and $\mathrm{d}x = \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n$, and $\int_x$ standing for the multiple integral.

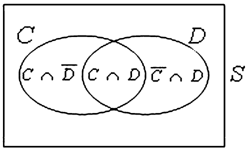


**Figure 13.3:** Illustration of Neyman-Pearson Lemma.

From equation (a), we have

$$\int_{C\cap\bar{D}} L_0 \mathrm{d}x = \int_{\bar{C}\cap D} L_0 \mathrm{d}x. \qquad \text{(b)}$$

Let $C$ be the critical region satisfying the condition $\frac{L_0}{L_1} \le k$ inside $C$. Consider the power of $C$, denoted by $p_1$. Then

$$p_1 = 1-\beta = \int_C L_1 \mathrm{d}x = \int_{C\cap\bar{D}} L_1 \mathrm{d}x + \int_{C\cap D} L_1 \mathrm{d}x.$$

But $C \cap \bar{D} \subset C$, and hence, inside $C$, $\frac{L_0}{L_1} \le k$ or $L_1 \ge \frac{L_0}{k}$. Therefore, we can write

$$\int_{C\cap\bar{D}} L_1 \mathrm{d}x \ge \int_{C\cap\bar{D}} \frac{L_0}{k} \mathrm{d}x.$$

Then

$$\begin{aligned} p_1 &\ge \int_{C\cap\bar{D}} \frac{L_0}{k} \mathrm{d}x + \int_{C\cap D} L_1 \mathrm{d}x \\ &= \int_{\bar{C}\cap D} \frac{L_0}{k} \mathrm{d}x + \int_{C\cap D} L_1 \mathrm{d}x \quad \text{from (b).} \end{aligned}$$

But $\bar{C} \cap D$ is outside $C$ and, therefore, $\frac{L_0}{L_1} > k$. Therefore, substituting this, we have

$$p_1 \ge \int_{\bar{C}\cap D} L_1 \mathrm{d}x + \int_{C\cap D} L_1 \mathrm{d}x = \int_D L_1 \mathrm{d}x = \text{power of } D.$$

Therefore, $C$ is the most powerful critical region.

**Example 13.1.** Consider a real scalar exponential population with parameter $\theta$ and a simple random sample $x_1, \dots, x_n$ of size $n$. Let $H_0$ be $\theta = 5$ or $\theta = \theta_0$ (given) and $H_1$ be $\theta = 10$ or $\theta = \theta_1$ (given). Assume that the parameter space $\Omega$ consists of $\theta_0$ and $\theta_1$ only. Construct the MPT or the most powerful critical region.

**Solution 13.1.**

$$L = \frac{1}{\theta^n} \mathrm{e}^{-\frac{1}{\theta}(x_1+\dots+x_n)}.$$

Consider the inequality

$$\frac{L_0}{L_1} \le k \quad \Rightarrow \quad \frac{\theta_1^n}{\theta_0^n} \mathrm{e}^{-(\frac{1}{\theta_0} - \frac{1}{\theta_1})(x_1+\dots+x_n)} \le k.$$

Taking the natural logarithms on both sides, we have

$$
\begin{aligned}
n\Big(\ln\frac{\theta_1}{\theta_0}\Big) - \Big(\frac{1}{\theta_0}-\frac{1}{\theta_1}\Big)(x_1+\cdots+x_n) &\le \ln k \quad \Rightarrow \\
-\Big(\frac{1}{\theta_0}-\frac{1}{\theta_1}\Big)(x_1+\cdots+x_n) &\le \ln k - n(\ln\theta_1-\ln\theta_0) = k_1 \quad \Rightarrow \\
\Big(\frac{1}{\theta_0}-\frac{1}{\theta_1}\Big)(x_1+\cdots+x_n) &\ge -k_1.
\end{aligned}
$$

Let $\theta_1 > \theta_0$. In this case $(\frac{1}{\theta_0}-\frac{1}{\theta_1}) > 0$ and then dividing both sides the inequality remains the same or we have $x_1+\cdots+x_n \ge k_2$ for some $k_2$. But we know the distribution of $u = x_1+\cdots+x_n$, which is a gamma with the parameters $(\alpha = n, \beta = \theta)$. But we have $\theta=\theta_0$ or $\theta=\theta_1$. In both of the cases, the gamma density is fully known, and hence we can compute percentage points. Let

$$
\begin{aligned}
\alpha &= \int_C L_0 \mathrm{d}x = \Pr\{x_1+\cdots+x_n \ge k_2 | \theta=\theta_0\} \\
&= \int_{u_\alpha}^{\infty} \frac{u^{n-1}}{\Gamma(n)\theta_0^n} \mathrm{e}^{-u/\theta_0} \mathrm{d}u.
\end{aligned}
$$

For a prefixed $\alpha$, compute $u_\alpha$ from a gamma density (known) and then the test criterion says: Reject $H_0$ if the observed $u = x_1+\cdots+x_n \ge u_\alpha$ as shown in Figure 13.4. This is the most powerful test. Note that if $\theta_1 < \theta_0$ then the MPT would have been the following: Reject $H_0$ if $u \le u_{1-\alpha}$. [This is left as an exercise to the student.]

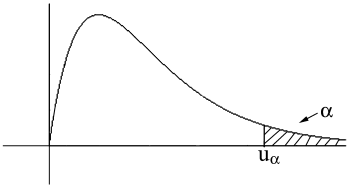


**Figure 13.4:** Percentage points for gamma density.

## Exercises 13.2

**13.2.1.** The ball-bearing for an ambassador car is manufactured by the car manufacturer's outfit at Kolkata. Identical duplicates are manufactured by some outfit in Punjab. It is found that the true percentage of defective ones produced by the Kolkata firm is 10% and that of the Punjab firm is 15%. A spare parts dealer has the stock of the original and duplicate spare parts. A garage bought 10 ball-bearing and 3 were found to be defective; test the hypothesis at the 5% level of rejection that the garage's lot were duplicates.

**13.2.2.** On a particular stretch of a highway, the expected number of monthly traffic accidents is 5, and if traffic policemen control the traffic then the expected number is 3. Assume that the number of traffic accidents there is Poisson distributed. Randomly selected 4 months gave the data 0, 3, 2, 3 accidents. Test the hypothesis, at the 5% level of rejection, that the traffic policemen were present on that stretch of the highway.

**13.2.3.** In the out-patient section of a small clinic, only one of two doctors Dr X and Dr Y will be present to attend to the out-patients on any given day. If Dr X is present, the expected waiting time in the queue is 30 minutes, and if Dr Y is present, then the expected waiting time is 40 minutes. Randomly selected 5 out-patients' waiting times on a particular day gave the data 50, 30, 40, 45, 25 minutes. Test the hypothesis at a 10% level of rejection, that Dr Y was present on that day, assuming an exponential distribution for the waiting time.

**13.2.4.** A margin-free shop has packets of a particular brand of potatoes marked as 5 kg packets. These packets are packed by automatic packing machines, for the expected weight of 5 kg, without cutting and chopping of potatoes. Sometimes the machine setting slips to 4.7 kg unknowingly. Four different housewives independently bought these bags of potatoes on a particular day and found to have the exact weights 3.6, 4.8, 4.8, 5.0 kg. Test the hypothesis, at a 5% level of rejection, that the machine setting was 4.7 kg when those packets were packed. Assume that the weight is approximately normally distributed $N(\mu, \sigma^2)$ with $\sigma^2 = 0.04$ kg.

### 13.2.1 The likelihood ratio criterion or the $\lambda$-criterion

The Neyman–Pearson lemma leads to a more general procedure called the lambda criterion or the likelihood ratio test criterion. Consider a simple random sample $X = (x_1, \dots, x_n)$ from some population $f(x, \theta)$. Then the joint density/probability function is $L(X, \theta) = \prod_{j=1}^{n} f(x_j, \theta)$. We can maximize this $L(X, \theta)$ over all $\theta \epsilon \Omega$, the parameter space. If there are several maxima, then take the supremum. Let the null hypothesis $H_0$ be $\theta \epsilon \omega \subset \Omega$. Under $H_0$, the parameter space is restricted to $\omega$. Then $L(X, \theta)|_{\theta \epsilon \omega}$ will be denoted by $L_0$. Then the likelihood ratio criterion or $\lambda$-criterion is defined, over the support of $f(x, \theta)$, as follows:

$$\lambda = \frac{\sup_{\theta \in \omega} L|H_0}{\sup_{\theta \in \Omega} L}. \tag{13.6}$$

Suppose that $H_0$ was in fact true. In this case, $\omega \equiv \Omega$ and in this case $\lambda \equiv 1$. In general, $0 < \lambda \leq 1$. Suppose that in a testing situation $\lambda$ is observed to be very small, close to zero. Then, from a layman's point of view, something is wrong with the null hypothesis because if the hypothesis is true, then $\lambda = 1$ and if it is nearly ok, then we could expect $\lambda$ to be close to 1. If $\lambda$ is observed at the other end, then we must reject our null hypothesis. Thus, we reject for small values of $\lambda$. Let $\lambda_\alpha$ be a point near zero. Then the criterion is the following:

Reject the null hypothesis if $\lambda \le \lambda_\alpha$ such that

$$\Pr\{\lambda \le \lambda_\alpha | H_0\} = \alpha. \tag{13.7}$$

This is known as the $\lambda$-criterion or the likelihood ratio criterion for testing the null hypothesis $H_0$.

**Example 13.2.** By using the likelihood ratio criterion, develop a test criterion for testing $H_0 : \mu \le \mu_0$ (given) for $\mu$ in a $N(\mu, \sigma^2)$ where $\sigma^2$ is known.

**Solution 13.2.** Let $X = (x_1, \dots, x_n)$ be a simple random sample from a $N(\mu, \sigma^2)$ where $\sigma^2$ is known. Then

$$\begin{aligned} L(X, \mu) &= \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x_j-\mu)^2} \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{n}(x_j - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right\}. \end{aligned} \tag{a}$$

The maximum likelihood estimator of $\mu$ over the whole parameter space $\Omega$ is $\bar{x}$. Here, $\sigma^2$ is known. Hence

$$\begin{aligned} \sup_{\mu \in \Omega} L(X, \mu) &= L\Big|_{\mu=\bar{x}} \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \bar{x})^2\right\}. \end{aligned}$$

The joint density of $X$ is given in (a). Now we want to maximize it under the null hypothesis $H_0 : \mu \le \mu_0$. From (a)

$$\max L \quad \Rightarrow \quad \max \ln L \quad \Rightarrow \quad \min n(\bar{x} - \mu)^2\big|_{\mu \le \mu_0}.$$

If the observed $\bar{x}$ is less than $\mu_0$, then it is an admissible value for $\mu$ since $\mu \le \mu_0$, and hence the maximum likelihood estimate is $\bar{x}$ itself. Then $\lambda = 1$ and we never reject $H_0$. Hence the rejection can come only when the observed $\bar{x} \ge \mu_0$. In this case, $(\bar{x} - \mu)^2$ can be made a minimum by assigning the maximum possible value for $\mu$, which is $\mu_0$ because $\mu \le \mu_0$. Hence the maximum likelihood estimate for $\mu$ under $H_0$ is $\mu_0$ and we reject only for large values of $\bar{x}$. Substituting these and simplifying, we have

$$\lambda = \exp\left\{-\frac{1}{2\sigma^2} n(\bar{x} - \mu_0)^2\right\}$$

which is a one to one function of $n(\bar{x} - \mu_0)^2$ or one to one function of $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \sim N(0,1)$ remembering that $\bar{x} > \mu_0$ and we reject for large values of $\bar{x}$ or for large values of

$z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}$. The probability coverage over this rejection region must be $\alpha$ for a prefixed $\alpha$. Then from $N(0,1)$ tables, we have

$$\Pr\{z \geq z_\alpha\} = \alpha. \tag{b}$$

Hence the test statistic here is $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}$ and the test criterion is:

Reject $H_0$ if $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \geq z_\alpha$ where $z_\alpha$ is given in (b).

**Note 13.1.** Was it possible to construct a test criterion if the null hypothesis was $H_0 : \mu < \mu_0$ in the open interval? From the procedure above, it may be noted that the maximum likelihood estimator (MLE) exists only when the boundary point $\mu_0$ is included. If $\mu_0$ is not included, then we could not have constructed $\lambda$. In that case, we could have tested a hypothesis of the type $\mu \geq \mu_0$ against $\mu < \mu_0$.

**Note 13.2.** If we had a hypothesis of the form $H_0 : \mu = \mu_0$ (given), $H_1 : \mu > \mu_0$ and if we had proceeded to evaluate the likelihood ratio criterion $\lambda$, then we would have ended up with the same criterion as for the hypotheses $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$. But hypotheses of the type $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ can be created only if we know beforehand that $\mu$ can only take values $\mu_0$ or higher. You may find misinterpretations in some books in the name of "one-sided tests". Such procedures of one-sided statements are logically incorrect if $\mu$ can also logically take values less than $\mu_0$. Similar comments hold for the case $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$.

**Note 13.3.** Hypotheses are to be formulated before the data are collected. Hypotheses have to come from theoretical considerations or claims made by manufacturers, business firms, etc. or proclamations made by public figures, etc. After formulating the hypotheses, data are to be collected to properly represent the populations assumed under the hypotheses. If hypotheses are formulated by looking at the data in hand, then it will result in the misuses of statistical procedures. Suppose that on four random occasions the information about a habitual gambler's net gain or loss is collected. All the four occasions resulted in huge gains. If you formulate a hypothesis that this gambler will always win, then such a claim may not be rejected by using any testing procedure based on the data in hand. Naturally, your conclusions will be logically absurd. Suppose that a sociologist has collected the data on annual incomes of families in Kerala. She checked five families at random. Her observations were 1 000, 2 000, 5 000, 5 000, 2 000 rupees. If she creates a hypothesis that the expected family income in Kerala can never be greater than 5 000 rupees or another hypothesis that the expected range of annual incomes will be between 1 000 and 5 000 rupees, both will result in absurd conclusions if the testing is done by using the data in hand. Hence hypotheses should not be formulated by looking at the data in hand.

**Example 13.3.** The grades obtained by the students in a particular course are assumed to be normally distributed, $N(\mu,\sigma^2)$, with $\sigma^2 = 16$. A randomly selected set of four students gave the grades as 60, 70, 80, 60. Test the hypothesis $H_0 : \mu \le 65$ against $H_1 : \mu > 65$, at the 2.5% level of rejection.

**Solution 13.3.** The observed sample mean $\bar{x} = \frac{1}{4}(60 + 70 + 80 + 60) = 67.5$. Here, $\mu_0 = 65$, $n = 4$. From a standard normal table, the 2.5% percentage point $z_{0.025} = 1.96$. Hence

$$\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \ge z_\alpha \quad \Rightarrow \quad \mu_0 + z_\alpha\frac{\sigma}{\sqrt{n}} = 65 + 1.96\left(\frac{4}{2}\right) = 68.92.$$

But the observed $\bar{x} = 67.5$, and hence the hypothesis is not rejected.

**Remark 13.1** ("Acceptance of a hypothesis"). If the testing procedure did not reject the null hypothesis $H_0$, the hypothesis being tested, can we "accept" the null hypothesis? This is a point of misuse of statistical techniques of testing of hypotheses. If you examine the procedures of constructing a test criterion, we see that it is done by minimizing the probability of type-II error and by prefixing the probability of type-I error and the whole procedure deals with rejecting $H_0$ and not for anything else. If the hypothesis $H_0$ is not rejected, then the procedure does not say anything about the decision to be made. When we do or do not reject our own hypothesis by using our own testing procedure we are not making a mathematical statement that such and such a thing is true.

In our example, the construction of the test statistic had at least the logical foundation of the likelihood ratio principle. Very often statisticians take a pivotal quantity on their own, not coming from any principle or procedure, and claim that under their own hypothesis they know the distribution of their pivotal quantity, and their own testing procedure did not reject their hypothesis, and hence they are accepting the hypothesis. The logical fallacy of such a procedure and argument is very clear. Suppose that a farmer's land bordering a forest area has troubles from wild animals. His vegetable gardens are usually destroyed. His claim is that wild buffaloes are the culprits. All buffaloes have four legs. Suppose that he selected the property "animals having four legs" as the property based on which his hypothesis will be tested. Note that this property of having four legs is not a characteristic property of buffaloes. Most of the statistical tests are not based on characterizing properties. (When a test statistic is taken, there is no unique determination of the whole distribution.) The farmer checked on four different nights and counted the number of legs of the culprits. If on a majority of the nights he found two-legged wild fowls, two-legged drunkards (human beings) destroying his vegetable garden. Then he can safely reject the hypothesis that wild buffaloes are destroying his vegetable garden. But suppose that on all nights he found 4-legged animals destroying his garden. He cannot accept the hypothesis that wild buffaloes are destroying his garden because 4-legged animals could be wild pigs (boars), porcupines, elephants, etc., including buffaloes.

## 13.3 Testing hypotheses on the parameters of a normal population $N(\mu,\sigma^2)$

### 13.3.1 Testing hypotheses on $\mu$ in $N(\mu,\sigma^2)$ when $\sigma^2$ is known

We have already looked into a problem of testing hypotheses on the mean value $\mu$ in a real scalar normal population $N(\mu,\sigma^2)$ when $\sigma^2$ is known, as an illustrative example. By using the likelihood ratio principle and $\lambda$-criterion we ended up with a criterion: reject $H_0$ if $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \ge z_\alpha$ for $H_0 : \mu \le \mu_0$ (given), $H_1 : \mu > \mu_0$. This is a *test at level $\alpha$* or the size of the critical region is $\alpha$ or test at *level of rejection $\alpha$*. If we had the hypotheses $H_0 : \mu \ge \mu_0$, $H_1 : \mu < \mu_0$, where $\sigma^2$ is known, then we would have ended up with the criterion: reject $H_0$ if $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \le -z_\alpha$. Similarly, for the hypotheses $H_0 : \mu = \mu_0$, $H_1 : \mu \ne \mu_0$ then the criterion would have been to reject $H_0$ if $|\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}| \ge z_{\frac{\alpha}{2}}$. These two cases are left to the student as exercises. Also, it was pointed out that if the hypotheses were $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ the criterion would have been the same as in the case of $H_0 : \mu \le \mu_0$, $H_1 : \mu > \mu_0$. Similarly, for the case $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ the criterion would have been the same as for the case $H_0 : \mu \ge \mu_0$, $H_1 : \mu < \mu_0$. But, such one-sided statements can be made only if one knows beforehand that $\mu$ cannot take the values less than $\mu_0$ or greater than $\mu_0$, as the case may be. The results of the tests of hypotheses on $\mu$ when $\sigma^2$ is known can be summarized as follows:

**Case (1).** Population $N(\mu,\sigma^2)$, $\sigma^2$ known. $H_0 : \mu \le \mu_0$, $H_1 : \mu > \mu_0$.
Test statistic: $z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \sim N(0,1)$.
Criterion: reject $H_0$ if the observed $z \ge z_\alpha$ or $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \ge z_\alpha$ or $\bar{x} \ge \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.

**Case (2).** Population $N(\mu,\sigma^2)$, $\sigma^2$ known. $H_0 : \mu \ge \mu_0$, $H_1 : \mu < \mu_0$.
Test statistic: $z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \sim N(0,1)$.
Criterion: reject $H_0$ if $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \le -z_\alpha$ or $\bar{x} \le \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$.

**Case (3).** Population $N(\mu,\sigma^2)$, $\sigma^2$ known. $H_0 : \mu = \mu_0$, $H_1 : \mu \ne \mu_0$.
Test statistic: $z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \sim N(0,1)$.
Criterion: reject $H_0$ if $|\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}| \ge z_{\alpha/2}$ or $\bar{x} \ge \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ or $\bar{x} \le \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

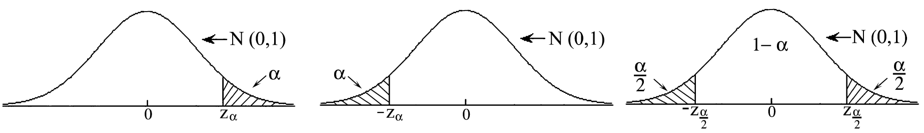


**Figure 13.5:** Illustration of the criteria for $H_0 : \mu \le \mu_0$; $H_0 : \mu \ge \mu_0$; $H_0 : \mu = \mu_0$.

The rejection region or *critical region* is $z_\alpha \le z < \infty$ for Case (1) or this region can be described as $\bar{x} \ge \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ and $\alpha$ is the size of the critical region or *level of the test*. The critical point is $z_\alpha$ in terms of $z$ or $\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ in terms of $\bar{x}$.

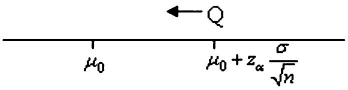


**Figure 13.6:** Critical point and critical region.

**Remark 13.2.** From Figure 13.6, note that the probability coverage over $\bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ is $\alpha$ for every $n$. Now, assume that $n$ is becoming larger and larger. Then the point $Q$ starts moving towards $\mu_0$ or the range for $\bar{x}$ to fall in the rejection region becomes larger and larger, with the same prefixed $\alpha$. Finally, when $n \to \infty$, by the weak law of large numbers, $\bar{x}$ goes to the true value $\mu$ or to $\mu_0$ if the hypothesized $\mu_0$ is the true value. For $n$ becoming larger and larger, we keep on rejecting $H_0$. By our procedure, $\bar{x}$ must fall above $\mu_0$. Hence critics of testing procedures say that we can always reject the null hypothesis by taking large enough sample size.

**Example 13.4.** The temperature at Pala during the month of August seems to hover around 28 °C. Someone wishes to test the hypothesis that the expected temperature on any given day in August in Pala area is less than 28 °, assuming that the temperature distribution is $N(\mu, \sigma^2)$, with $\sigma^2 = 4$. The following is the data on temperature reading on randomly selected 4 days in August: 30, 31, 28, 25. Test at 2.5% level of rejection.

**Solution 13.4.** The hypothesis of the type less than 28 °C cannot be tested because it is in the open interval. We can test $H_0 : \mu \geq 28$ against $H_1 : \mu < 28$. Hence we formulate the hypothesis to be tested in this format. The observed sample mean $\bar{x} = \frac{1}{4}(30 + 31 + 28 + 25) = 28.5$. $z_\alpha = z_{0.025} = 1.96$. $\mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = 28 - 1.96(\frac{2}{2}) = 26.04$. But the observed value of $\bar{x} = 28.5 > 26.04$, and hence the hypothesis $H_0 : \mu \geq 28$ is not rejected at the 2.5% level.

### 13.3.2 Tests of hypotheses on $\mu$ in $N(\mu, \sigma^2)$ when $\sigma^2$ is unknown

Here, since $\sigma^2$ is unknown we need to estimate $\sigma^2$ also. These estimates in the whole parameter space are $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$. Hence, substituting these, we have the maximum of the likelihood function, given by

$$\max_{\mu,\sigma^2} L = \frac{1}{(s^2 2\pi)^{\frac{n}{2}}} \mathrm{e}^{-n/2}. \tag{13.8}$$

Let us try to test the null hypothesis $\mu \geq \mu_0$ (given), against $H_1 : \mu < \mu_0$. If the observed $\bar{x}$ falls in the interval $\mu_0 \leq \bar{x} < \infty$, then it in the admissible range of $\mu$, under $H_0$, and hence the MLE of $\mu$, under the null hypothesis $H_0$, is $\bar{x}$ and then $\lambda \equiv 1$, and hence we do not reject $H_0$ in this case. Hence the question of rejection comes only when the observed $\bar{x}$ falls below $\mu_0$. But

$$\frac{\partial}{\partial \mu} \ln L = 0 \quad \Rightarrow \quad \mu - \bar{x} = 0.$$

Hence if $\mu$ cannot take the value $\bar{x}$, then we assign the closest possible value to $\bar{x}$ for $\mu$, which is $\mu_0$ under $H_0 : \mu \leq \mu_0$ as well as for $H_0 : \mu \geq \mu_0$. [Note that this value can be assigned only because $\mu_0$ is an admissible value under $H_0$ or $H_0$ contains that point. Hence in the open interval $\mu > \mu_0$ testing is not possible by using this procedure.] Then the MLE of $\sigma^2$, under $H_0$, is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_0)^2.$$

Substituting these MLE's for $\mu$ and $\sigma^2$, under $H_0$, we have

$$\begin{aligned}\lambda^{2/n} &= \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\sum_{j=1}^{n}(x_j-\mu_0)^2}\\ &= \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{[\sum_{j=1}^{n}(x_j-\bar{x})^2 + n(\bar{x}-\mu_0)^2]}\\ &= \frac{1}{[1+\frac{n(\bar{x}-\mu_0)^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}]}, \quad \bar{x} < \mu_0.\end{aligned}$$

This is a one to one function of the Student-$t$ statistic

$$t_{n-1} = \frac{\sqrt{n}(\bar{x}-\mu_0)}{s_1}, \quad s_1^2 = \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1}. \tag{13.9}$$

Since the criterion is constructed for $\bar{x} < \mu_0$, the hypothesis $H_0$ is rejected for small values of $t_{n-1}$ or

$$\Pr\{t_{n-1} \leq -t_{n-1,\alpha}\} = \alpha. \tag{13.10}$$

Thus the criterion can be stated as follows: reject $H_0 : \mu \geq \mu_0$ when $t_{n-1} \leq -t_{n-1,\alpha}$ or when the observed $\bar{x} \leq \mu_0 - t_{n-1,\alpha}\frac{s_1}{\sqrt{n}}$, where $s_1^2$ the unbiased estimator of $\sigma^2$ is given in (13.9). Thus when $\sigma^2$ is unknown the standardized normal test statistic changes into a Student-$t$ statistic.

**Note 13.4.** Note from the examples and discussions so far that the rejection region is always in the direction of the alternate hypothesis. When the alternate is $\theta > \theta_0$, we reject at the right tail with probability $\alpha$; when $H_1$ is $\theta < \theta_0$ we reject at the left tail with probability $\alpha$ and when $H_1$ is $\theta \neq \theta_0$ we reject at both the tails (right and left with probabilities $\alpha_1$ and $\alpha_2$, such that $\alpha_1 + \alpha_2 = \alpha$, but for convenience we take $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ each).

**Example 13.5.** The following is the data on the time taken by a typist to type a page of mathematics in TEX: 20, 30, 25, 41. Assuming that the time taken is normally distributed $N(\mu,\sigma^2)$ with unknown $\sigma^2$, test the hypothesis $H_0 : \mu \geq \mu_0 = 30$, $H_1 : \mu < \mu_0$, at the level of rejection 5%.

**Solution 13.5.** The sample mean $\bar{x} = \frac{1}{4}(20 + 30 + 25 + 41) = 29$. Observed value of $s_1^2 = \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1} = \frac{242}{3}$. $t_{n-1,\alpha} = t_{3,0.05} = 2.353$. $\mu_0 = 30$. $\mu_0 - t_{n-1,\alpha}\frac{s_1}{\sqrt{n}} = 30 - 2.353(\frac{\sqrt{242}}{2\sqrt{3}}) = 30 - 10.57 = 19.43$. But the observed $\bar{x} = 29$, which is not less than 19.43, and hence we cannot reject $H_0$, at a 5% level of rejection.

In the above example, the observed value of the Student-$t$ variable, $t_{n-1} = \frac{\sqrt{n}(\bar{x}-\mu_0)}{s_1} = \frac{2(29-30)}{8.98} = -\frac{2}{8.98} = -0.22$. The observed value of the test statistic is $-0.22$ and the rejection region is on the left tail. Hence we can compute the probability that a $t_{n-1} \le -0.22$. This is called the $p$-value for this example.

**Definition 13.6** (The $p$-values). Compute the observed value of the test statistic used for testing a hypothesis $H_0$. Let the test statistic be denoted by $u$ and the observed value be denoted by $u_0$. If the rejection region is on the right, then the $p$-value is $\Pr\{u \ge u_0\}$; if the rejection region is on the left (in this case usually $u_0$ will be negative if the statistic can take negative values) then the $p$-value is $\Pr\{u \le u_0\}$ and if the rejection region is at both ends then the $p$-value is $\Pr\{u \ge |u_0|\} + \Pr\{u \le -|u_0|\}$ if $u$ has a symmetric distribution, symmetric about $u = 0$.

The advantage of $p$-values is that, instead of a pre-fixed $\alpha$, by looking at the $p$-values we can make decisions at various levels of $\alpha$, and conclude at which level $H_0$ is rejected and at which level $H_0$ is not rejected. We can summarize the inference on $\mu$, when $\sigma^2$ is unknown as follows:

**Case (4).** $H_0 : \mu \le \mu_0, H_1 : \mu > \mu_0$, population $N(\mu, \sigma^2)$, $\sigma^2$ unknown.

Test statistic: $\frac{\sqrt{n}(\bar{x}-\mu_0)}{s_1} \sim t_{n-1}, s_1^2 = \frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n-1}$;

Criterion: reject $H_0$ if the observed $t_{n-1} \ge t_{n-1,\alpha}$ or the observed value of $\bar{x} \ge \mu_0 + t_{n-1,\alpha}\frac{s_1}{\sqrt{n}}$.

**Case (5).** $H_0 : \mu \ge \mu_0, H_1 : \mu < \mu_0$, population $N(\mu, \sigma^2)$, $\sigma^2$ unknown.

Test statistic: same as in Case (4);

Criterion: reject $H_0$ if the observed value of $t_{n-1} \le -t_{n-1,\alpha}$ or the observed value of $\bar{x} \le \mu_0 - t_{n-1,\alpha}\frac{s_1}{\sqrt{n}}$.

**Case (6).** $H_0 : \mu = \mu_0, H_1 : \mu \ne \mu_0$, population $N(\mu, \sigma^2)$, $\sigma^2$ unknown.

Test statistic: same as in cases (4) and (5);

Criterion: reject $H_0$ if the observed value of $|\frac{\sqrt{n}(\bar{x}-\mu_0)}{s_1}| \ge t_{n-1,\alpha/2}$ or the observed value of $\bar{x} \ge \mu_0 + t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}}$ or $\bar{x} \le \mu_0 - t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}}$ as illustrated in Figure 13.7.

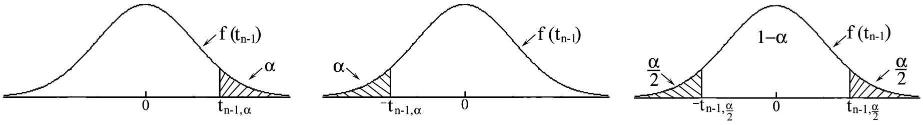


**Figure 13.7:** Illustration of $H_0$ on $\mu$ in $N(\mu, \sigma^2)$ when $\sigma^2$ is unknown.

### 13.3.3 Testing hypotheses on $\sigma^2$ in a $N(\mu, \sigma^2)$

Here, there are two cases to be considered. (a) when $\mu$ is known and (b) when $\mu$ is not known. The maximum of the likelihood function in the whole of the parameter space will be the following:

$$\max_{(\mu,\sigma^2)\in\Omega} L = \frac{1}{(2\pi s^2)^{\frac{n}{2}}}\mathrm{e}^{-n/2}, \quad s^2 = \frac{\sum_{j=1}^{n}(x_j-\mu)^2}{n}.$$

Replace $\mu$ by $\bar{x}$ if $\mu$ is unknown. Let us take the case $H_0 : \sigma^2 \le \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$ where $\mu$ is known. Then what is the maximum of the likelihood function under this $H_0$? Let $\theta = \sigma^2$. The likelihood equation is the following:

$$\frac{\partial}{\partial\theta}\ln L = 0 \quad \Rightarrow \quad \theta - s^2 = 0, \quad s^2 = \frac{\sum_{j=1}^{n}(x_j-\mu)^2}{n}.$$

Then if $s^2$ is an admissible value for $\theta$ then $\hat{\theta} = s^2$, otherwise assign the closest possible value to $s^2$ for $\theta$. The closest possible value to $s^2$ for $\theta$ is $\sigma_0^2$ for $H_0 : \sigma^2 \le \sigma_0^2$ (given), as well as for $H_0 : \sigma^2 \ge \sigma_0^2$. But

$$u = \frac{\sum_{j=1}^{n}(x_j-\mu)^2}{\sigma_0^2} \sim \chi_n^2$$

a chi-square with $n$ degrees of freedom. If $\mu$ is unknown, then $\mu$ will be replaced by $\bar{x}$ when doing the maximization and in this case $u \sim \chi_{n-1}^2$. This is the only difference. Hence the $\lambda$-criterion becomes

$$\lambda = \frac{u^{n/2}}{n^{n/2}}\mathrm{e}^{-\frac{1}{2}u}, \quad u \sim \chi_n^2. \tag{13.11}$$

The shape of this $\lambda$, as a function of $u$, is given in Figure 13.8.

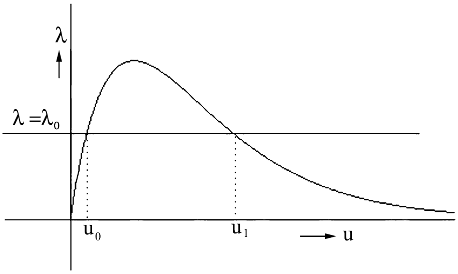


**Figure 13.8:** Shape of $\lambda$-criterion as a function of $u = \chi_n^2$.

We always reject for small values of $\lambda$, that is, $\Pr\{\lambda \le \lambda_0 | H_0\} = \alpha$. As shown in Figure 13.8, this statement is equivalent to the general probability statement for $u \le u_0$

and $u \geq u_1$ as shown in Figure 13.8. But in the above case we reject only for large values of $s^2$ compared to $\sigma_0^2$. Hence for the above case we reject only at the right tail. That is, we reject $H_0$ if $\chi_n^2 \geq \chi_{n,\alpha}^2$ . Similar procedures will yield the test criteria for the other cases. We can summarize the results as follows and illustration is given in Figure 13.9:

**Case (7).** $H_0 : \sigma^2 \leq \sigma_0^2$ (given), $H_1 : \sigma^2 > \sigma_0^2$, population $N(\mu, \sigma^2)$, $\mu$ known.

Test statistic: $\frac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \sim \chi_n^2$. (Replace the degrees of freedom $n$ by $n-1$ and $\mu$ by $\bar{x}$ when $\mu$ is unknown. Hence this situation is not listed separately. Make the changes for each case accordingly. When $\mu$ is known, we have the choice of making use of $\mu$ or ignoring this information.)

Criterion: reject $H_0$ if the observed value of $\frac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \geq \chi_{n,\alpha}^2$.

**Case (8).** $H_0 : \sigma^2 \geq \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$, population $N(\mu, \sigma^2), \mu$ known.

Test statistic: same as above.

Criterion: reject $H_0$ if the observed $\frac{\sum_{j=1}^n (x_j-\mu)^2}{\sigma_0^2} \leq \chi_{n,1-\alpha}^2$.

**Case (9).** $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 \neq \sigma_0^2$, population $N(\mu, \sigma^2)$, $\mu$ known.

Test statistic: same as above.

Criterion: reject $H_0$ if the observed $\chi_n^2 \leq \chi_{n,1-\frac{\alpha}{2}}^2$ or $\chi_n^2 \geq \chi_{n,\frac{\alpha}{2}}^2$. (Note from Figure 13.8 that the cut off areas are not equal to $\frac{\alpha}{2}$ at both ends but for convenience we will take them as equal.)

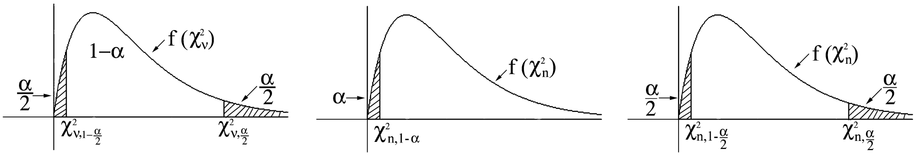


**Figure 13.9:** Testing hypotheses on $\sigma^2$ in $N(\mu, \sigma^2)$, $\mu$ known.

When $\mu$ is unknown, all steps and criteria are parallel. Replace $\mu$ by $\bar{x}$ and the degrees of freedom $n$ by $n-1$ for the chi-square. If $\mu$ is known but if we choose to ignore this information, then also replace $\mu$ by $\bar{x}$.

## Exercises 13.3

**13.3.1.** By using the likelihood ratio principle, derive the test criteria for testing the following hypotheses on $\mu$ in a $N(\mu, \sigma^2)$ where $\sigma^2$ is known, and assuming that a simple random sample of size $n$ is available from this population and here $\mu_0$ refers to a given number.

(1) $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ (It is known beforehand that $\mu$ can never be less than $\mu_0$.)

(2) $H_0 : \mu \geq \mu_0$, $H_1 : \mu < \mu_0$

(3) $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ (It is known beforehand that $\mu$ can never be greater than $\mu_0$.)

(4) $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$

**13.3.2.** Repeat Exercise 13.2.1 for the following situations when $\sigma^2$ is unknown:

(1) $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$ (It is known beforehand that $\mu$ can never be greater than $\mu_0$.)

**13.3.3.** By using the likelihood ratio principle, derive the test criteria for testing the following hypotheses on $\sigma^2$ in a $N(\mu, \sigma^2)$, assuming that a simple random sample of size $n$ is available. Construct the criteria for the cases (a) $\mu$ is known, (b) $\mu$ is unknown for the following situations, where $\sigma_0^2$ denotes a given quantity:

(1) $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$ (It is known beforehand that $\sigma^2$ can never be less than $\sigma_0^2$.)

(2) $H_0 : \sigma^2 \geq \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2$

(3) $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$ (It is known beforehand that $\sigma^2$ can never be greater than $\sigma_0^2$.)

(4) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$

**13.3.4.** Illustrate all cases in Exercise 13.3.1 if $\mu = 1\,\text{kg}$ and the observed data are the yields of tapioca on experimental test plots given by $5, 10, 3, 7$ kg.

**13.3.5.** Illustrate all cases in Exercises 13.3.2 for the same data in Exercise 13.3.4.

**13.3.6.** Illustrate all cases in Exercise 13.3.3 by using the data in Exercise 13.3.4 for the cases (a) $\mu = 1$, (b) $\mu$ is unknown. Cut off equal areas at both tails, for convenience.

## 13.4 Testing hypotheses in bivariate normal population

In a bivariate normal distribution, there are five parameters. If $(x_1, x_2)$ represents a bivariate random variable, then the parameters are $\mu_1 = E(x_1)$, $\mu_2 = E(x_2)$, $\text{Var}(x_1) = \sigma_1^2$, $\text{Var}(x_2) = \sigma_2^2$, $\rho$ = correlation between $x_1$ and $x_2$. If we have $n$ data points, then the sample will be of the form $X_i = (x_{1i}, x_{2i})$, $i = 1, \dots, n$, where capital $X_i$ represents a vector. If we have a simple random sample from $(x_1, x_2)$, then $X_i$, $i = 1, \dots, n$ are iid variables. In a practical situation, it may be possible to take observations but we may not have information about the five parameters. The situation may be that $x_1$ is the weight of an experimental animal before giving a special animal feed and $x_2$ may be the weight of the same animal after administering the special feed. Evidently, $x_1$ and $x_2$ are not independently distributed. Another example is the situation of administering a drug. $x_1$ may be the blood pressure before giving the drug and $x_2$ the same blood pressure after giving the drug. Our aim may be to test hypotheses of the type

$\mu_2 \le \mu_1$ or $\mu_2 - \mu_1 \le 0$. Such hypotheses can be tested without knowing the five parameters.

Let us consider a more general situation of testing a hypothesis on $a\mu_1 + b\mu_2 + c$ where $a, b, c$ are known constants such as $2\mu_1 - 3\mu_2 + 2 \le 0$. This can be tested because we know that when $(x_1, x_2)$ has a bivariate normal distribution, then all linear functions of $x_1$ and $x_2$ are also normally distributed. Hence if $u = ax_1 + bx_2 + c$, then $u \sim N(\mu, \sigma^2)$, where $\mu = E(u) = a\mu_1 + b\mu_2 + c$ and $\sigma^2 = \mathrm{Var}(u)$. Here, $\sigma^2$ is usually unknown. Hence the procedure is the following. Convert the observations $(x_{1i}, x_{2i})$ to observations on $u$, namely

$$u_i = ax_{1i} + bx_{2i} + c, \quad i = 1, \dots, n, \quad u \sim N(\mu, \sigma^2)$$

where $a, b, c$ are known. Let

$$\bar{u} = \frac{\sum_{j=1}^{n} u_j}{n}, \quad s_1^2 = \frac{\sum_{j=1}^{n}(u_j - \bar{u})^2}{n-1}.$$

Then from Section 13.3.2, the test statistic will be a Student-$t$ with $n - 1$ degrees of freedom, given by

$$\frac{\sqrt{n}(\bar{u} - \mu_0)}{s_1} \sim t_{n-1} \tag{13.12}$$

where $\mu_0$ is the hypothesized value of $E(u)$. The test criteria, at the level of rejection $\alpha$, are the following, where $\mu = a\mu_1 + b\mu_2 + c$, $a, b, c$ known:

**Case (10).** $H_0 : \mu \le \mu_0$, $H_1 : \mu > \mu_0$; reject $H_0$ if the observed value of $\frac{\sqrt{n}(\bar{u}-\mu_0)}{s_1} \ge t_{n-1,\alpha}$ or $\bar{u} \ge \mu_0 + t_{n-1,\alpha}\frac{s_1}{\sqrt{n}}$.

**Case (11).** $H_0 : \mu \ge \mu_0$, $H_1 : \mu < \mu_0$; reject $H_0$ if the observed value of $\frac{\sqrt{n}(\bar{u}-\mu_0)}{s_1} \le -t_{n-1,\alpha}$ or $\bar{u} \le \mu_0 - t_{n-1,\alpha}\frac{s_1}{\sqrt{n}}$.

**Case (12).** $H_0 : \mu = \mu_0$, $H_1 : \mu \ne \mu_0$; reject $H_0$ if the observed value of $|\frac{\sqrt{n}(\bar{u}-\mu_0)}{s_1}| \ge t_{n-1,\frac{\alpha}{2}}$ or $\bar{u} \ge \mu_0 + t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}}$ or $\le \mu_0 - t_{n-1,\frac{\alpha}{2}}\frac{s_1}{\sqrt{n}}$.

The illustration is the same as the one in Section 13.3.2, Figure 13.7. Let $\sigma^2 = \mathrm{Var}(u)$. Then we can test hypotheses on $\sigma^2$ also by using the same data.

**Case (13).** $H_0 : \sigma^2 \le \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$; reject $H_0$ if the observed value of $\chi_{n-1}^2 = \frac{\sum_{j=1}^{n}(u_j-\bar{u})^2}{\sigma_0^2} \ge \chi_{n-1,\alpha}^2$.

**Case (14).** $H_0 : \sigma^2 \ge \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$; reject $H_0$ if the observed value of the same $\chi_{n-1}^2 \le \chi_{n-1,1-\alpha}^2$.

**Case (15).** $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 \ne \sigma_0^2$; reject $H_0$ if the observed value of the same $\chi_{n-1}^2 \le \chi_{n-1,1-\frac{\alpha}{2}}^2$ or $\ge \chi_{n-1,\frac{\alpha}{2}}^2$.

**Note 13.5.** If $(x_1,\dots,x_p) \sim N_p(\tilde{\mu},\Sigma)$, a $p$-variate multinormal then the same procedure can be used for testing hypotheses on the expected value and variance of any given linear function $a_1x_1 + \dots + a_px_p + b$, where $a_1,\dots,a_p,b$ are known, without knowing the individual mean values or the covariance matrix $\Sigma$.

**Example 13.6.** The claim of a particular exercise routine is that the weight will be reduced at least by 5 kilograms (kg). A set of 4 people are selected at random from the set of individuals who went through the exercise routine. The weight before starting is $x_1$ and the weight at the finish is $x_2$. The following are the observations on $(x_1,x_2)$: $(50,50)$, $(60,55)$, $(70,60)$, $(70,75)$. Assuming a bivariate normal distribution for $(x_1,x_2)$ test the claim at a 2.5% level of rejection.

**Solution 13.6.** Let $\mu_1 = E(x_1)$, $\mu_2 = E(x_2)$. Then the claim is $\mu_1 - \mu_2 \ge 5$. Let $u = x_1 - x_2$. Then the observations on $u$ are $50 - 50 = 0$, $60 - 55 = 5$, $70 - 60 = 10$, $70 - 75 = -5$ and the observed $\bar{u} = \frac{1}{4}(0 + 5 + 10 - 5) = 2.5$. An observed value of $s_1^2 = \frac{1}{n-1}\sum_{j=1}^{n}(u_j - \bar{u})^2 = \frac{1}{3}[(0-2.5)^2 + (5-2.5)^2 + (10-2.5)^2 + (-5-2.5)^2] = \frac{125}{3}$, $s_1 \approx 6.45$, $\mu_0 = 5$. Hence $\mu_0 - t_{n-1,\alpha}\frac{s_1}{\sqrt{n}} = 5 - t_{3,0.025}\frac{6.45}{2} = 5 - 3.182(3.225) \approx -5.26$. The observed value of $\bar{u} = 2.25$ which is not less than $-5.26$, and hence the null hypothesis is not rejected.

**Remark 13.3.** Our conclusion should not be interpreted as the claim being correct and it cannot be interpreted that we can "accept" the claim. The above result only indicates that the data at hand does not enable us to reject the claim. Perhaps other data points might have rejected the hypothesis or perhaps the normality assumption may not be correct thereby the procedure becomes invalid or perhaps the claim itself may be correct.

## Exercises 13.4

**13.4.1.** Let $x_1$ be the grade of a student in a topic before subjecting to a special method of coaching and let $x_2$ be the corresponding grade after the coaching. Let $E(x_1) = \mu_1$, $E(x_2) = \mu_2$. The following are the paired observations on the grades of five independently selected students from the same set. $(80,85)$, $(90,92)$, $(85,80)$, $(60,70)$, $(65,68)$. Assuming $(x_1,x_2)$ to have a bivariate normal distribution test the following claims at 5% level of rejection: (1) $\mu_2 \ge \mu_1$; (2) $2\mu_2 - 3\mu_1 < 2$; (3) $5\mu_2 - 2\mu_1 \le 3$.

**13.4.2.** Let $t_1$ be the body temperature of a patient having some sort of fever before giving particular medicine and let $t_2$ be the temperature after giving the medicine. Let $\mu_1 = E(t_1)$, $\mu_2 = E(t_2)$, $\sigma_1^2 = \mathrm{Var}(t_1 - t_2)$, $\sigma_2^2 = \mathrm{Var}(2t_1 - 3t_2)$. Assume that $(t_1,t_2)$ has a bivariate normal distribution. The following are the observations on $(t_1,t_2)$ from 4 randomly selected patients having the same sort of fever. $(101,98)$, $(100,97)$, $(100,100)$, $(98,97)$. Test the following hypotheses at 5% level of rejection. (1) $H_0 : \mu_2 < \mu_1$; (2) $H_0 : \mu_2 \le \mu_1 + 1$; (3) $H_0 : \mu_1 + 1 \ge \mu_2$; (4) $H_0 : \sigma_1^2 \le 0.04$; (5) $H_0 : \sigma_1^2 = 0.2$; (6) $H_0 : \sigma_2^2 \ge 0.3$.

## 13.5 Testing hypotheses on the parameters of independent normal populations

Let $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ and let $x_1$ and $x_2$ be independently distributed. This situation is a special case of the situation in Section 13.4. We can test hypotheses on the mean values and variances of general linear functions of $x_1$ and $x_2$ by using similar procedure as adopted in Section 13.4, that is, hypotheses on $E(u)$ and $\mathrm{Var}(u)$, where $u = a_1x_1 + a_2x_2 + b$, $a_1, a_2, b$ are known constants. We will list here some standard situations such as $H_0 : \mu_1 - \mu_2 = \delta$, for given $\delta$. Let $x_{11}, \dots, x_{1n_1}$ and $x_{21}, \dots, x_{2n_2}$ be simple random samples of sizes $n_1$ and $n_2$ from $x_1$ and $x_2$, respectively. Let

$$\bar{x}_1 = \sum_{j=1}^{n_1} \frac{x_{1j}}{n_1}, \quad \bar{x}_2 = \sum_{j=1}^{n_2} \frac{x_{2j}}{n_2}, \quad s_1^2 = \sum_{j=1}^{n_1} \frac{(x_{1j} - \bar{x}_1)^2}{n_1},$$
$$s_2^2 = \sum_{j=1}^{n_2} \frac{(x_{2j} - \bar{x}_2)^2}{n_2}, \quad s_{1(1)}^2 = \sum_{j=1}^{n_1} \frac{(x_{1j} - \bar{x}_1)^2}{n_1 - 1}, \quad s_{2(1)}^2 = \sum_{j=1}^{n_2} \frac{(x_{2j} - \bar{x}_2)^2}{n_2 - 1},$$
$$s^2 = \frac{[\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2]}{n_1 + n_2 - 2}. \tag{13.13}$$

The likelihood ratio principle will lead to the following test criteria at $\alpha$ level of rejection:

**Case (16).** $\sigma_1^2$, $\sigma_2^2$ known, $\delta$ given. $H_0 : \mu_1 - \mu_2 \leq \delta$, $H_1 : \mu_1 - \mu_2 > \delta$; Test statistic: $z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$; Criterion: reject $H_0$ if the observed value of $z \geq z_\alpha$.

**Case (17).** $\sigma_1^2$, $\sigma_2^2$, $\delta$ known. $H_0 : \mu_1 - \mu_2 \geq \delta$, $H_1 : \mu_1 - \mu_2 < \delta$. Test statistic is the same $z$ as above. Test criterion: reject $H_0$ if the observed value of $z \leq -z_\alpha$.

**Case (18).** $\sigma_1^2$, $\sigma_2^2$, $\delta$ known. $H_0 : \mu_1 - \mu_2 = \delta$, $H_1 : \mu_1 - \mu_2 \neq \delta$. Test statistic is the same as above. Test criterion: reject $H_0$ if the observed value of $|z| \geq z_{\frac{\alpha}{2}}$.

Illustration is the same as the ones in Section 13.3.1, Figure 13.5.

**Case (19).** $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), $\delta$ given. $H_0 : \mu_1 - \mu_2 \leq \delta$, $H_1 : \mu_1 - \mu_2 > \delta$. Test statistic

$$\frac{[\bar{x}_1 - \bar{x}_2 - \delta]}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

or a Student-$t$ with $n_1 + n_2 - 2$ degrees of freedom, where $s^2$ is given in (13.13). Test criterion: reject $H_0$ if the observed value of $t_{n_1+n_2-2} \geq t_{n_1+n_2-2,\alpha}$.

**Case (20).** $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), $\delta$ given. $H_0 : \mu_1 - \mu_2 \geq \delta$, $H_1 : \mu_1 - \mu_2 < \delta$. Test statistic is the same Student-$t$ as above. Test criterion: reject $H_0$ if the observed value of $t_{n_1+n_2-2} \leq -t_{n_1+n_2-2,\alpha}$.

**Case (21).** $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), $\delta$ given. $H_0 : \mu_1 - \mu_2 = \delta$, $H_1 : \mu_1 - \mu_2 \neq \delta$. Test statistic is the same as above. Test criterion: reject $H_0$ if the observed value of $|t_{n_1+n_2-2}| \geq t_{n_1+n_2-2,\frac{\alpha}{2}}$.

The illustration is the same as in Section 13.3.2, Figure 13.7. We can also test a hypothesis on a constant multiple of the ratio of the variances in this case. Consider a typical hypothesis of the type $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq \eta > 0$ for a given $\eta$. Without loss of generality, we can take $\eta = 1$ and write the hypothesis as $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$. Let us examine the likelihood ratio principle here. Let $\theta$ represent all the parameters $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. The parameter space is

$$\theta \in \Omega \quad \Rightarrow \quad \Omega = \{\theta \mid -\infty < \mu_i < \infty, 0 < \sigma_i^2 < \infty, i = 1, 2\}.$$

The joint density function is given by

$$L = \left[\frac{1}{(2\pi\sigma_1^2)^{\frac{n_1}{2}}} \mathrm{e}^{-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n_1}(x_{1j}-\mu_1)^2}\right] \times \left[\frac{1}{(2\pi\sigma_2^2)^{\frac{n_2}{2}}} \mathrm{e}^{-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}(x_{2j}-\mu_2)^2}\right]$$

In the parameter space $\Omega$, the MLEs are $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2$ which are given in (13.13). Hence the maximum of the likelihood function is the following:

$$\max_{\theta\in\Omega} L = \frac{\mathrm{e}^{-\frac{1}{2}(n_1+n_2)}}{[(2\pi s_1^2)^{\frac{n_1}{2}}(2\pi s_2^2)^{\frac{n_2}{2}}]}. \tag{13.14}$$

Now, let us impose the hypothesis $H_0 : \sigma_1^2 \leq \sigma_2^2$. The MLEs of $\mu_1$ and $\mu_2$ remain the same as $\hat{\mu}_1 = \bar{x}_1$ and $\hat{\mu}_2 = \bar{x}_2$. We can also write this $H_0$ as $\sigma_1^2 = \delta\sigma_2^2$, $0 < \delta \leq 1$. Then the joint density, at $\hat{\mu}_1 = \bar{x}_1$, $\hat{\mu}_2 = \bar{x}_2$, denoted by $L_1$, becomes

$$L_1 = [(2\pi)^{\frac{n_1+n_2}{2}}\delta^{\frac{n_1}{2}}(\sigma_2^2)^{\frac{n_1+n_2}{2}}]^{-1} \times \exp\left\{-\frac{1}{2\sigma_2^2}\left[\frac{n_1 s_1^2}{\delta} + n_2 s_2^2\right]\right\}$$

where $s_1^2$ and $s_2^2$ are as defined in (13.13). Maximizing $L_1$ by using calculus we have the estimator for $\sigma_2^2$ as

$$\hat{\sigma}_2^2 = \frac{[n_1 s_1^2 \delta] + n_2 s_2^2}{n_1 + n_2}.$$

Therefore, substituting all these values we have the $\lambda$-criterion as the following. [In this representation of $\lambda$, we have used the property that

$$\frac{n_1(n_2-1)s_1^2}{n_2(n_1-1)s_2^2} = \frac{[\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)^2/(n_1-1)]}{[\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2/(n_2-1)]} \sim F_{n_1-1,n_2-1} \tag{13.15}$$

when $\sigma_1^2 = \sigma_2^2$, where $F_{n_1-1,n_2-1}$ is an $F$-random variable with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Then $\lambda$ is written in terms of $F$-variable, for convenience.]

$$\begin{aligned}\lambda &= \frac{(s_1^2)^{\frac{n_1}{2}}(s_2^2)^{\frac{n_2}{2}}}{\delta^{\frac{n_1}{2}}(\hat{\sigma}_2^2)^{\frac{1}{2}(n_1+n_2)}}\\ &= c_1\frac{[F_{n_1-1,n_2-1}]^{\frac{n_1}{2}}}{[1+c_2F_{n_1-,n_2-1}]^{\frac{1}{2}(n_1+n_2)}}\end{aligned}$$

where $c_1$ and $c_2$ are positive constants. The nature of $\lambda$, as a function of $F_{n_1-1,n_2-1}$ is given in Figure 13.10.

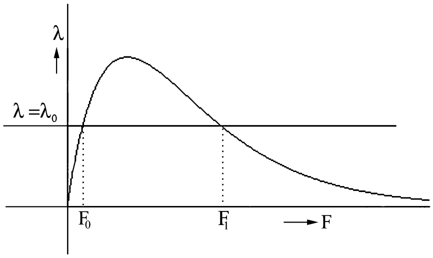


**Figure 13.10:** $\lambda$ as a function of $F_{n_1-1,n_2-1}$.

Note that in the above $H_0$ and alternate we will not be rejecting for small values of $\frac{\sigma_1^2}{\sigma_2^2}$ and we will reject only for large values or we will reject only for large values of $F_{n_1-1,n_2-1}$. Hence the criteria can be stated as the following and illustration is given in Figure 13.11.

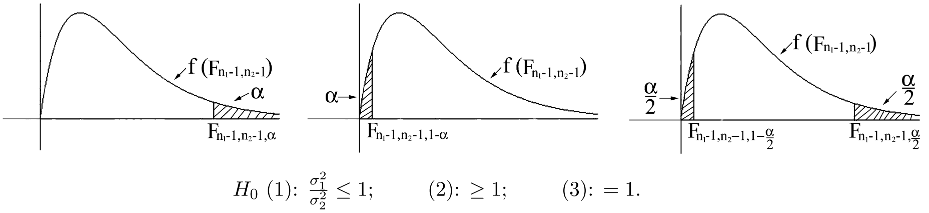


**Figure 13.11:** Illustration of the critical regions.

**Case (22).** Independent $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ populations. $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \le 1$, $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$. Test statistic: $F_{n_1-1,n_2-1}$ given in (13.15). Test criterion: reject $H_0$ if the observed value of $F_{n_1-1,n_2-1} \ge F_{n_1-1,n_2-1,\alpha}$.

**Case (23).** Same independent normal populations as above. $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \ge 1$, $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$. Test statistic is the same as above. Test criterion: reject $H_0$ if the observed value of $F_{n_1-1,n_2-1} \le F_{n_1-1,n_2-1,1-\alpha}$.

**Case (24).** Same independent normal populations as above. $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 \ne \sigma_2^2$. Test statistic is the same as above. Criterion: reject $H_0$ if the observed $F_{n_1-1,n_2-1} \le F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ or $F_{n_1-1,n_2-1} \ge F_{n_1-1,n_2-1,\frac{\alpha}{2}}$.

**Note 13.6.** The student will not find the lower percentage point for a $F$-density tabulated. This is because you can get lower percentage points from the upper percentage points of another $F$-density. The connection is the following:

$$F_{m,n} \le F_{m,n,1-\alpha} \quad \Rightarrow \quad F_{m,n} \ge \frac{1}{F_{n,m,\alpha}}$$

and hence the lower percentage points in $F_{m,n}$ are available from the reciprocal of the upper percentage points of $F_{n,m}$.

**Note 13.7.** If $\mu_1$ and $\mu_2$ are known, then we can use $\mu_1$ and $\mu_2$ instead of their estimators. In this case, replace $F_{n_1-1,n_2-1}$ by $F_{n_1,n_2}$. In this situation, if we replace $\mu_1$ and $\mu_2$ by their estimators and proceeded as above, then also the procedure is valid. Hence in this situation we can use both $F_{n_1,n_2}$ as well as $F_{n_1-1,n_2-1}$ as test statistics.

**Example 13.7.** Steel rods made through two different processes are tested for breaking strengths. Let the breaking strengths under the two processes be denoted by $x$ and $y$, respectively. The following are the observations on $x$: $5, 10, 12, 3$, and on $y$: $8, 15, 10$, respectively. Assuming that $x \sim N(\mu_1, \sigma_1^2)$ and $y \sim N(\mu_2, \sigma_2^2)$ and independently distributed, test the hypothesis at a 5% level of rejection that $\sigma_1^2 \le \sigma_2^2$.

**Solution 13.7.** Here, according to our notations, $n_1 = 4$, $n_2 = 3$, the observed values of $\bar{x} = \frac{1}{4}(5 + 10 + 12 + 3) = 7.5 = \bar{x}_1$, $\bar{y} = \frac{1}{3}(8 + 15 + 10) = 11 = \bar{x}_2$, $\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)^2 = (5 - 7.5)^2 + (10-7.5)^2 + (12-7.5)^2 + (3-7.5)^2 = 53$, $\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2 = (8-11)^2 + (15-11)^2 + (10-11)^2 = 26$. Therefore, the observed value of

$$F_{n_1-1,n_2-1} = F_{3,2} = \frac{(53/3)}{(26/2)} \approx 1.34.$$

From a $F$-table, we have $F_{3,2,0.05} = 19.2$. But 1.34 is not bigger than 19.2, and hence we cannot reject $H_0$.

## Exercises 13.5

**13.5.1.** The yields of ginger from test plots, under two different planting schemes, are the following: $20, 25, 22, 27$ (scheme 1), $23, 18, 28, 30, 32$ (scheme 2). If the yields under the two schemes are independently normally distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, test the following hypotheses at a 5% level of rejection, where the alternates are the natural alternates or negation of the null hypotheses: (1) $H_0 : \mu_1 - \mu_2 \le 1$, given that $\sigma_1^2 = 2$, $\sigma_2^2 = 3$; (2) $H_0 : 2\mu_1 - 3\mu_2 \ge 3$ given that $\sigma_1^2 = \sigma_2^2$; (3) $H_0 : \mu_1 = \mu_2$, given that $\sigma_1^2 = \sigma_2^2$; (4) $H_0 : \sigma_1^2 \le \sigma_2^2$ given that $\mu_1 = 0$, $\mu_2 = -2$; (5) $H_0 : \sigma_1^2 = \sigma_2^2$.

**13.5.2.** Derive the $\lambda$-criterion for the following cases and show that they agree with what is given in the text above, the alternates are the negation of the null hypotheses:

(1) $H_0 : \mu_1 \geq \mu_2$, given $\sigma_1^2 = 5$, $\sigma_2^2 = 2$; (2) $H_0 : 2\mu_1 - \mu_2 \leq 2$, given that $\sigma_1^2 = \sigma_2^2$; (3) $H_0 : \sigma_1^2 \geq \sigma_2^2$, given that $\mu_1 = 0$, $\mu_2 = -5$.

## 13.6 Approximations when the populations are normal

In Section 13.3.2, we dealt with the problem of testing hypotheses on the mean value $\mu$ of a normal population when the population variance $\sigma^2$ is unknown. We ended up with a Student-$t$ as the test statistic and the decisions were based on a $t_{n-1}$, Student-$t$ with $n-1$ degrees of freedom.

### 13.6.1 Student-$t$ approximation to normal

In many books, the student may find a statement that if the degrees of freedom are bigger than 30, then read off the percentage points from a standard normal table, instead of the Student-$t$ table, and Student-$t$ tables are not usually available also for large values of the degrees of freedom, beyond 30. But the student can use a computer and program it and compute to see the agreement of the tail areas with the tail areas of a standard normal density. You will see that even for degrees of freedom as large as 120 the agreement is not that good. The theoretical basis is given in Exercise 13.6.1 at the end of this section. Hence when replacing Student-$t$ values with standard normal values this point should be kept in mind. Hence the approximation used is the following:

$$u = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} \approx N(0,1), \quad \text{for large } n \tag{13.16}$$

where $s^2$ could be the unbiased estimator for $\sigma^2$ or we may use the MLE itself. When $n$ is large, dividing by $n$ or $n-1$ will not make much of a difference. Hence the test criteria can be stated as follows:

**Case (25).** Population $N(\mu, \sigma^2)$, $\sigma^2$ unknown, $n$ is very large. $H_0 : \mu \leq \mu_0$ (given), $H_1 : \mu > \mu_0$. Test statistic as given in (13.16). Criterion: reject $H_0$ if the observed $u \geq z_\alpha$ where $z_\alpha$ is the upper $100\alpha\%$ point from a standard normal.

**Case (26).** Population and situation as in Case (25). $H_0 : \mu \geq \mu_0$, $H_1 : \mu < \mu_0$. Test statistic: same $u$ as in (13.16). Criterion: reject $H_0$ if the observed $u \leq -z_\alpha$.

**Case (27).** Population details are as in Case (25). $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$. Test statistic: same $u$ in (13.16). Criterion: reject $H_0$ when the observed value of $|u| \geq z_{\frac{\alpha}{2}}$. Illustration is as in Section 13.3.1, Figure 13.5.

In the case of two independent normal populations $x_1 \sim N(\mu_1, \sigma_1^2)$, and $x_2 \sim N(\mu_2, \sigma_2^2)$, and samples of sizes $n_1$ and $n_2$ and the notations as used in (13.12), we may have a situation where $\sigma_1^2$ and $\sigma_2^2$ are unknown. Then replace the variances by the MLEs $s_1^2$ and $s_2^2$ and consider the following statistic:

$$v = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0,1) \tag{13.17}$$

when $n_1$ is large and $n_2$ is also large. Again, the approximation is poor and as a crude rule, we may use the approximation in (13.14) when $n_1 \geq 30$, $n_2 \geq 30$. In the following cases, we list only the types of hypotheses on $\mu_1 - \mu_2$ for convenience, but remember that similar hypotheses on linear functions of $\mu_1$ and $\mu_2$ can be formulated and tested.

**Case (28).** Independent normal populations $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, samples of sizes $n_1 \geq 30$ and $n_2 \geq 30$, $\sigma_1^2$, $\sigma_2^2$ unknown. $H_0 : \mu_1 - \mu_2 \leq \delta$ (given), $H_1 : \mu_1 - \mu_2 > \delta$. Test statistic: $v$ in (13.14). Criterion: reject $H_0$ if the observed value of $v \geq z_\alpha$.

**Case (29).** Same situation as in Case (28). $H_0 : \mu_1 - \mu_2 \geq \delta$ (given), $H_1 : \mu_1 - \mu_2 < \delta$. Test statistic: $v$ as in (13.17). Criterion: reject $H_0$ if the observed value of $v \leq -z_\alpha$.

**Case (30).** Same situation as in Case (28). $H_0 : \mu_1 - \mu_2 = \delta$ (given), $H_1 : \mu_1 - \mu_2 \neq \delta$. Test statistic: same $v$ as in (13.17). Criterion: reject $H_0$ if the observed value of $|v| \geq z_{\frac{\alpha}{2}}$. Illustrations as in Section 13.3.1, Figure 13.5.

In the case of independent normal populations, we may have a situation $\sigma_1^2 = \sigma_2^2 = \sigma^2$ where the common $\sigma^2$ may be unknown. Then we have seen that we can replace this $\sigma^2$ by a combined unbiased estimator as given in (13.13) and we will end up with a Student-$t$ with $n_1 + n_2 - 2$ degrees of freedom. If this $n_1 + n_2 - 2$ is really large, then we can have a standard normal approximation as follows:

$$w = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\hat{\sigma}\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1) \tag{13.18}$$

where for the approximation to hold $n_1 + n_2 - 2$ has to be really large, as explained above. Then we can have the following test criteria:

**Case (31).** Independent normal populations $N(\mu_1, \sigma_1^2)$,$N(\mu_2, \sigma_2^2)$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), $n_1 + n_2 - 2$ is very large. $H_0 : \mu_1 - \mu_2 \leq \delta$ (given), $H_1 : \mu_1 - \mu_2 > \delta$. Test statistic: $w$ in (13.18). Criterion: reject $H_0$ if the observed value of $w \geq z_\alpha$.

**Case (32).** Situation as in Case (31). $H_0 : \mu_1 - \mu_2 \geq \delta$ (given), $H_1 : \mu_1 - \mu_2 < \delta$. Test statistic: $w$ of (13.18). Criterion: reject $H_0$ if the observed value of $w \leq -z_\alpha$.

**Case (33).** Situation as in Case (31). $H_0 : \mu_1 - \mu_2 = \delta$ (given), $H_1 : \mu_1 - \mu_2 \neq \delta$. Test statistic: $w$ as in (13.18). Criterion: reject $H_0$ if the observed value of $|w| \geq z_{\frac{\alpha}{2}}$; see the illustrations as in Section 13.3.1, Figure 13.5.

### 13.6.2 Approximations based on the central limit theorem

The central limit theorem says that whatever be the population, designated by the random variable $x$, as long as the variance $\sigma^2$ of $x$, is finite, then for a simple random sample $x_1, \ldots, x_n$, the standardized sample mean goes to standard normal when $n \to \infty$.

$$u = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \to N(0,1), \quad \text{as } n \to \infty \tag{13.19}$$

where $\mu_0$ is a given value of the mean value of $x$. Hence when $\sigma^2$ is known, we can test hypotheses on $\mu$ if we have very large $n$. The approximate tests are the following:

**Case (34).** Any population with known $\sigma^2 < \infty$, $n$ very large. $H_0 : \mu \leq \mu_0$ (given), $H_1 : \mu > \mu_0$. Test statistic: $u$ in (13.19). Criterion: reject $H_0$ if the observed value of $u \geq z_\alpha$.

**Case (35).** Any population with known finite variance, $n$ very large. $H_0 : \mu \geq \mu_0$ (given), $H_1 : \mu < \mu_0$. Test statistic: $u$ in (13.19). Criterion: reject $H_0$ if the observed value of $u \leq -z_\alpha$.

**Case (36).** Any population with known finite variance, $n$ very large. Test statistic: $u$ in (13.19). Criterion: reject $H_0$ if the observed value of $|u| \geq z_{\frac{\alpha}{2}}$; see the illustrations as in Section 13.3.1, Figure 13.5.

In the approximation, $\sigma$ must be known. If $\sigma$ is not known and if $\sigma^2$ is replaced by an unbiased estimator for $\sigma^2$, then we do not have a Student-$t$ approximation even for large $n$. Hence the students are advised not to try to test hypotheses on $\mu$ by estimating $\sigma^2$ when the population is unknown.

## Exercises 13.6

**13.6.1.** Take a Student-$t$ density with $\nu$ degrees of freedom. Take the limit when $\nu \to \infty$ and show that the Student-$t$ density goes to a standard normal density when $\nu \to \infty$. [Hint: Use Stirling's formula on $\Gamma(\frac{\nu}{2})$ and $\Gamma(\frac{\nu+1}{2})$. Stirling's formula says that when $|z| \to \infty$ and $\alpha$ a bounded quantity, then

$$\Gamma(z + \alpha) \approx \sqrt{2\pi} z^{z+\alpha-\frac{1}{2}} \mathrm{e}^{-z}.$$

Take $z = \frac{\nu}{2}$ when using Stirling's formula on Student-$t$ density. The process $(1 + \frac{1}{n})^n \to \mathrm{e}$ is very slow, and hence no good approximation can come for small values of $\nu$. Even for $\nu = 120$, the approximation is not that close. Hence there should be caution when replacing a Student-$t$ variable with a standard normal variable.]

**13.6.2.** For a Student-$t$ with $\nu$ degrees of freedom, show that the $h$-th integer moment can exist only when $\nu > h$, and when $\nu > h$ then all odd moments for $h < \nu$ will be zeroes.

**13.6.3.** Two different varieties of sunflower plants are planted on 50 test plots under variety 1 and 60 test plots under variety 2. The summary data, in our notations in the text above, (see equation (13.13)), are the following: $\bar{x}_1 = 20$, $\bar{x}_2 = 25$, $s_1^2 = 4$, $s_2^2 = 9$. Test the following hypotheses, against the natural alternates, at a 5% level of rejection, assuming that $x$ and $y$, typical yields under variety 1 and variety 2, respectively, are independently normally distributed: (1) $H_0 : \mu_1 - \mu_2 \geq 3$, (2) $H_0 : \mu_2 - \mu_1 = 2$ but it is known that $\sigma_1^2 = \sigma_2^2$ (unknown).

## 13.7 Testing hypotheses in binomial, Poisson and exponential populations

In our binomial population, there are two parameters $(n,p)$, $0 < p < 1$ and in our Poisson population there is one parameter $\lambda$. We will consider the problems of testing hypotheses on $p$ when $n$ is known, which is also the same as testing hypotheses on $p$ of a point-Bernoulli population, and testing hypotheses on the Poisson parameter $\lambda$.

### 13.7.1 Hypotheses on $p$, the Bernoulli parameter

Consider $n$ iid variables from a Bernoulli population. Then the joint probability function, denoted by $L$ is the following:

$$L = p^{(\sum_{j=1}^{n} x_j)}(1-p)^{\sum_{j=1}^{n}(1-x_j)} = p^x(1-p)^{n-x},$$

where $x$ is a binomial random variable. The MLE of $p$ in the whole parameter space is $\hat{p} = \frac{x}{n}$. Under $H_0 : 0 < p \le p_0$ (given), $\lambda \equiv 1$ if $\frac{x}{n} \le p_0$. Hence we reject $H_0$ only when $\frac{x}{n} > p_0$ or when $x$ is large. The likelihood equation is

$$\frac{\partial}{\partial p} \ln L = 0 \quad \Rightarrow \quad p - \frac{x}{n} = 0.$$

If $\frac{x}{n}$ is an admissible value for $p$, then $\hat{p} = \frac{x}{n}$. Otherwise assign the closest value to $\frac{x}{n}$ for $p$, that is, the MLE under $H_0$ is $p_0$ for both $H_0 : p \le p_0$ (given) and $H_0 : p \ge p_0$. But $x$ has a binomial distribution. How large should $x$ be for rejecting $H_0 : p \le p_0$? Here, $x \ge x_0$ such that

$$\Pr\{x \ge x_0 | p = p_0\} \le \alpha \quad \text{or} \quad \sum_{x=x_0}^{n} \binom{n}{x} p_0^x (1-p_0)^{n-x} \le \alpha. \tag{13.20}$$

Look into binomial tables for $p = p_0$ and compute $x_0$ of (13.20). If the sum of the binomial probabilities does not hit $\alpha$ exactly, then take the nearest $x_0$ such that the probability is $\le\alpha$. For example, from the binomial tables, for $n = 10$, $p_0 = 0.40$, $\Pr\{x \ge 8|p = 0.4\} = 0.0123$, $\Pr\{x \ge 7|p = 0.4\} = 0.0548$. Hence if $\alpha = 0.05$ then we take $x_0 = 8$. Hence the criteria can be stated as follows:

**Case (37).** Bernoulli parameter $p$. $H_0 : p \le p_0$ (given), $H_1 : p > p_0$. Test statistic: binomial variable $x$. Criterion: reject $H_0$ if the observed number of successes is $\ge x_0$ where $\sum_{x=x_0}^{n} \binom{n}{x} p_0^x (1-p_0)^{n-x} \le \alpha$.

**Case (38).** Bernoulli parameter $p$. $H_0 : p \ge p_0$ (given), $H_1 : p < p_0$. Test statistic: binomial variable $x$. Criterion: reject $H_0$ if the observed number of successes $\le x_1$, such that $\sum_{x=0}^{x_1} \binom{n}{x} p_0^x (1-p_0)^{n-x} \le \alpha$.

**Case (39).** Bernoulli parameter $p$. $H_o : p = p_0$ (given), $H_1 : p \neq p_0$. Test statistic: binomial variable $x$. Criterion: reject $H_0$ if the observed value of $x \leq x_2$ or $x \geq x_3$ where

$$\sum_{x=0}^{x_2} \binom{n}{x} p_0^x (1-p_0)^{n-x} \leq \frac{\alpha}{2}, \quad \sum_{x=x_3}^{n} \binom{n}{x} p_0^x (1-p_0)^{n-x} \leq \frac{\alpha}{2},$$

as illustrated in Figure 13.12.

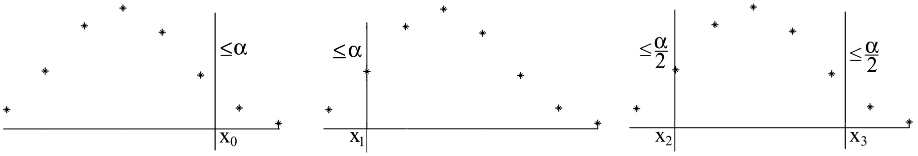


**Figure 13.12:** Hypotheses $H_0 : p \leq p_0$, $H_0 : p \geq p_0$, $H_0 : p = p_0$.

**Example 13.8.** Someone is shooting at a target. Let $p$ be the true probability of a hit. Assume that $p$ remains the same from trial to trial and that the trials are independent. Out of 10 trials, she has 4 hits. Test the hypothesis $H_0 : p \leq 0.45$, $H_1 : p > 0.45$ at the level of rejection $\alpha = 0.05$.

**Solution 13.8.** The number of hits $x$ is a binomial variable with parameters $(p, n = 10)$. We reject $H_0$ if the observed number of hits is $\geq x_0$, such that

$$\Pr\{x \geq x_0 | p = 0.45\} \leq 0.05 \quad \text{or} \quad \sum_{x=x_0}^{10} \binom{10}{x} (0.45)^x (0.65)^{10-x} \leq 0.05.$$

From a binomial table for $n = 10$ and $p = 0.45$, we have $\Pr\{x \geq 8 | p = 0.45\} = 0.0274$ and $\Pr\{x \geq 7 | p = 0.45\} = 0.102$. Hence $x_0 = 8$ at $\alpha = 0.05$. But our observed $x$ is 4, which is not $\geq 8$, and hence $H_0$ is not rejected.

**Note 13.8.** Note that the procedure in testing hypotheses on the binomial parameter $p$ is different from the procedure of establishing confidence intervals for $p$. In some of the tests on the parameters of the normal populations, it may be noted that the non-rejection intervals in hypotheses testing agree with some confidence intervals. Because of this, some authors mix up the process of constructing confidence intervals and tests of hypotheses and give credence to the statement "acceptance of $H_0$". It is already pointed out the logical fallacy of such misinterpretations of tests of hypotheses. Besides, tests are constructed by using some principles, such as the likelihood ratio principle, but confidence intervals can be constructed by picking any suitable pivotal quantity. Comments, similar to the ones here, also hold for the situations of tests of hypotheses and construction of confidence intervals in the Poisson case also, which will be discussed next.

### 13.7.2 Hypotheses on a Poisson parameter

The Poisson probability function is $f(x) = \frac{\theta^x}{x!}\mathrm{e}^{-\theta}$, $\theta > 0$ with the support $x = 0, 1, 2, \ldots$. Let $x_1, \ldots, x_n$ be iid variables from this Poisson population. [Usually the Poisson parameter is denoted by $\lambda$, but in order to avoid confusion with the $\lambda$ of the $\lambda$-criterion, we will use $\theta$ here for the Poisson parameter.] The joint probability function of $x_1, \ldots, x_n$ is

$$L = \frac{\theta^{x_1+\cdots+x_n}\mathrm{e}^{-n\theta}}{x_1!\cdots x_n!}$$

and the MLE of $\theta$ in the whole parameter space is $\hat{\theta} = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}$. Consider the hypothesis: $H_0 : \theta \le \theta_0$ (given), $H_1 : \theta > \theta_0$. If $\bar{x}$ falls in the interval $0 < \bar{x} \le \theta_0$, then the MLE in the parameter space $\Omega = \{\theta \mid 0 < \theta < \infty\}$ as well as under $H_0$ coincide and the likelihood ratio criterion $\lambda \equiv 1$, and we do not reject $H_0$. Hence we reject $H_0$ only when $\bar{x} > \theta_0$. The likelihood equation

$$\frac{\partial}{\partial\theta}\ln L = 0 \quad \Rightarrow \quad \theta - \bar{x} = 0.$$

Hence if $\bar{x}$ is an admissible value for $\theta$, then take $\hat{\theta} = \bar{x}$, otherwise assign the closest possible value to $\bar{x}$ as the estimate for $\theta$. Hence the MLE, under $H_0 : \theta \le \theta_0$ (given) as well as for $H_0 : \theta \ge \theta_0$, is $\theta_0$. Hence the test criterion will be the following: Reject $H_0$ if the observed value of $\bar{x}$ is large or the observed value of $u = x_1 + \cdots + x_n$ is large. But $u$ is Poisson distributed with parameter $n\theta$. Therefore, we reject $H_0 : \theta \le \theta_0$ if $u \ge u_0$ such that $\Pr\{u \ge u_0 | \theta = \theta_0\} \le \alpha$ or

$$\sum_{u=u_0}^{\infty} \frac{(n\theta_0)^u}{u!}\mathrm{e}^{-n\theta_0} \le \alpha \quad \text{or} \quad \sum_{u=0}^{u_0-1} \frac{(n\theta_0)^u}{u!}\mathrm{e}^{-n\theta_0} \ge 1 - \alpha$$

where

$$u = (x_1 + \cdots + x_n) \sim \text{Poisson}(n\theta_0). \tag{13.21}$$

We can summarize the results as follows:

**Case (40).** Parameter $\theta$ in a Poisson distribution. $H_0 : \theta \le \theta_0$ (given), $H_1 : \theta > \theta_0$. Test statistic: $u$ in (13.21). Criterion: reject $H_0$ if the observed value of $u \ge u_0$ such that $\Pr\{u \ge u_0 | \theta = \theta_0\} \le \alpha$.

**Case (41).** Poisson parameter $\theta$. $H_0 : \theta \ge \theta_0$ (given), $H_1 : \theta < \theta_0$. Test statistic: $u$ in (13.21). Criterion: reject $H_0$ if the observed value of $u \le u_1$ such that $\Pr\{u \le u_1 | \theta = \theta_0\} \le \alpha$.

**Case (42).** Poisson parameter $\theta$. $H_0 : \theta = \theta_0$ (given), $H_1 : \theta \ne \theta_0$. Test statistic: $u$ in (13.21). Criterion: reject $H_0$ if the observed value of $u \le u_2$ or $\ge u_3$, such that $\Pr\{u \le$

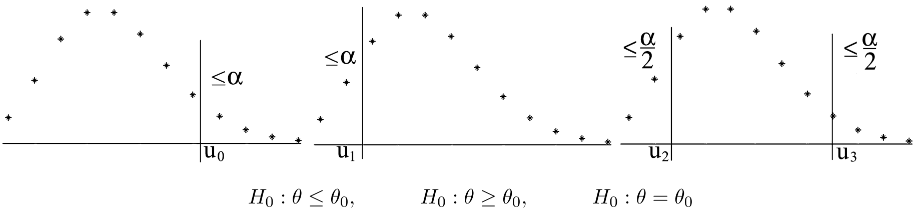


**Figure 13.13:** Hypotheses in a Poisson population.

$u_2|\theta = \theta_0\} \le \frac{\alpha}{2}$ and $\Pr\{u \ge u_3|\theta = \theta_0\} \le \frac{\alpha}{2}$. (Equal cut off areas at both ends are taken for convenience only.) For illustration see Figure 13.13.

**Example 13.9.** The number of snake bites per year in a certain village is found to be Poisson distributed with expected number of bites $\theta$. A randomly selected 3 years gave the number of bites as $0, 8, 4$. Test the hypothesis, at 5% level of rejection, that $\theta \le 2$.

**Solution 13.9.** According to our notation, the observed value of $u = (0 + 8 + 4) = 12$, $\theta_0 = 2$, $n = 3$, $n\theta_0 = 6$, $\alpha = 0.05$. We need to compute $u_0$ such that

$$\sum_{u=u_0}^{\infty} \frac{6^u}{u!} e^{-6} \le 0.05 \quad \Rightarrow \quad \sum_{u=0}^{u_0-1} \frac{6^u}{u!} e^{-6} \ge 0.95.$$

From the Poisson table, we have $u_0 - 1 = 10$ or $u_0 = 11$. Our observed $u$ is $12 > 11$, and hence we reject $H_0$.

### 13.7.3 Hypotheses in an exponential population

The exponential population is given by the density

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \quad x \ge 0,\ \theta > 0$$

and zero elsewhere. Consider $x_1, \dots, x_n$ iid variables from this population. Then the joint density function of $x_1, \dots, x_n$ is

$$L = \frac{1}{\theta^n} e^{-\frac{1}{\theta}(x_1+\dots+x_n)}.$$

The likelihood equation is given by

$$\frac{\partial}{\partial \theta} \ln L = 0 \quad \Rightarrow \quad \theta - \bar{x} = 0.$$

Hence if $\bar{x}$ is an admissible value for $\theta$ then $\hat{\theta} = \bar{x}$, otherwise assign for $\theta$ the admissible closest value to $\bar{x}$. Therefore, the MLE in the whole parameter space is $\hat{\theta} = \bar{x}$. But for

$H_0 : \theta \le \theta_0$ (given) as well as for $H_0 : \theta \ge \theta_0$ the MLE, under $H_0$, is $\theta_0$. In $H_0 : \theta \le \theta_0$ we reject only for $\bar{x} > \theta_0$ (otherwise $\lambda \equiv 1$) and in the case $H_0 : \theta \ge \theta_0$ we reject only when $\bar{x} < \theta_0$. Hence for $H_0 : \theta \le \theta_0$ we reject for large values of $\bar{x}$ or for large values of $u = (x_1 + \cdots + x_n)$. When $x_j$ has exponential distribution, $u$ has a gamma distribution with parameters $(\alpha = n, \beta = \theta_0)$, under $H_0$, or $v = \frac{u}{\theta_0}$ is gamma with parameters $(n, 1)$ or $w = \frac{2u}{\theta_0} \sim \chi^2_{2n}$, a chi-square with $2n$ degrees of freedom. Hence for testing the hypothesis, we can use either $u$ or $v$ or $w$. Hence the criteria are the following, given in terms of a chi-square variable.

**Case (43).** Exponential population with parameter $\theta$. $H_0 : \theta \le \theta_0$ (given), $H_1 : \theta > \theta_0$. Test statistic: $w = \frac{2(x_1+\cdots+x_n)}{\theta_0} \sim \chi^2_{2n}$. Criterion: reject $H_0$ if the observed value of $w \ge \chi^2_{2n,\alpha}$.

**Case (44).** Same population as in Case (43). $H_0 : \theta \ge \theta_0$ (given), $H_1 : \theta < \theta_0$. Test statistic is the same $w$ as in Case (43). Criterion: reject $H_0$ if the observed value of $w \le \chi^2_{2n,1-\alpha}$.

**Case (45).** Same population as in Cases (43). $H_0 : \theta = \theta_0$ (given), $H_1 : \theta \ne \theta_0$. Test statistic is the same $w$ as above. Criterion: reject $H_0$ if the observed value of $w \le \chi^2_{2n,1-\frac{\alpha}{2}}$ or $w \ge \chi^2_{2n,\frac{\alpha}{2}}$. [Equal tail areas are taken only for convenience.] The illustrations for Cases (43), (44), (45) are given in Figure 13.9, with degrees of freedom $2n$.

## Exercises 13.7

**13.7.1.** In a drug-testing experiment on mice, some mice die out before the experiment is completed. All mice are of identical genotype and age. Assuming that the probability of a mouse dying is $p$ and it is the same for all mice, test the hypotheses (1) $p \le 0.4$, $H_1 : p > 0.4$; (2) $H_0 : p = 0.4$, $H_1 : p \ne 0.4$, at a 5% level of rejection, and based on the data in one experiment of 10 mice where (a) 5 died, (b) 4 died.

**13.7.2.** The number of floods in Meenachil River in the month of September is seen to be Poisson distributed with expected number $\theta$. Test the hypotheses, at 5% level of rejection, that (1) $H_0 : \theta = 2$, $H_1 : \theta \ne 2$, (2): $H_0 : \theta \ge 2$, $H_1 : \theta < 2$. Two years are selected at random. The numbers of floods in September are 8, 1.

**13.7.3.** A typist makes mistakes on every page she types. Let $p$ be the probability of finding at least 5 mistakes in any page she types. Test the hypotheses (1) $H_0 : p = 0.8$, $H_1 : p \ne 0.8$, (2) $H_0 : p \ge 0.8$, $H_1 : p < 0.8$, at a 5% level of rejection, and based on the following observations: Out of 8 randomly selected pages that she typed, 6 had mistakes of more than 5 each.

**13.7.4.** Lightning strikes in the Pala area in the month of May is seen to be Poisson distributed with expected number $\theta$. Test the hypotheses (1) $H_0 : \theta \le 3$, $H_1 : \theta > 3$; (2) $H_0 : \theta = 3$, $H_1 : \theta \ne 3$, at a 5% level of rejection, and based on the following observations. Three years are selected at random and there were 2, 2, 3 lightning strikes in May.

**13.7.5.** The waiting time in a queue at a check-out counter in a grocery store is assumed to be exponential with expected waiting time $\theta$ minutes, time being measured in minutes. The following are observations on 3 randomly selected occasions at this check-out counter: $10, 8, 2$. Check the following hypotheses at a 5% level of rejection: (1) $H_0 : \theta \leq 2$, $H_1 : \theta > 2$; (2) $H_0 : \theta = 8$, $H_1 : \theta \neq 8$.

**Remark 13.4.** After having examined various hypotheses and testing procedures, the students can now evaluate the whole procedure of testing hypotheses. All the test criteria or test statistics were constructed by prefixing the probability of rejection of $H_0$ under $H_0$, namely $\alpha$, and the probability of rejection of $H_0$ under the alternate, namely $1 - \beta$. Thus the whole thing is backing up only the situation of rejecting $H_0$. If $H_0$ is not rejected, then the theory has nothing to say. In a practical situation, the person calling for a statistical test may have the opposite in mind, to accept $H_0$ but the theory does not justify such a process of "accepting $H_0$".

There are two standard technical terms, which are used in this area of testing of statistical hypotheses. These are null and non-null distributions of test statistics.

**Definition 13.7** (Null and non-null distributions). The distribution of a test statistic $\lambda$ under the null hypothesis $H_0$ is called the null distribution of $\lambda$ and the distribution of $\lambda$ under negation of $H_0$ is called the non-null distribution of $\lambda$.

Observe that we need the null distribution of a test statistic for carrying out the test and non-null distribution for studying the power of the test so that the test can be compared with others tests of the same size $\alpha$.

## 13.8 Some hypotheses on multivariate normal

In the following discussion, capital Latin letters will stand for vector or matrix variables and small Latin letter for scalar variables, and Greek letters (small and capital) will be used to denote parameters, a prime will denote the transpose. We only consider real scalar, vector, matrix variables here. Let $X$ be $p \times 1$, $X' = (x_1, \dots, x_p)$. The standard notation $N_p(\mu, \Sigma)$, $\Sigma > 0$ means that the $p \times 1$ vector $X$ has a $p$-variate non-singular normal distribution with mean value vector $\mu' = (\mu_1, \dots, \mu_p)$ and non-singular positive definite $p \times p$ covariance matrix $\Sigma$, and with the density function

$$f(X) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(X - \mu)' \Sigma^{-1} (X - \mu)\right\} \tag{13.22}$$

where $-\infty < \mu_i < \infty$, $-\infty < x_i < \infty$, $i = 1, \dots, p$, $\Sigma = (\sigma_{ij}) = \Sigma' > 0$. If we have a simple random sample of size $N$ from this $N_p(\mu, \Sigma)$, then the sample values and the sample matrix can be represented as follows:

$$X_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{pi} \end{bmatrix}, \quad i = 1, \dots, N; \quad \text{sample matrix } X = \begin{bmatrix} x_{11} & \dots & x_{1N} \\ x_{21} & \dots & x_{2N} \\ \vdots & \dots & \vdots \\ x_{p1} & \dots & x_{pN} \end{bmatrix}.$$

The sample average or sample mean vector, a matrix of sample means and the deviation matrix are the following:

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}, \quad \bar{x}_i = \frac{\sum_{k=1}^{N} x_{ik}}{N},$$

$$\tilde{X} = \begin{bmatrix} \bar{x}_1 & \dots & \bar{x}_1 \\ \bar{x}_2 & \dots & \bar{x}_2 \\ \vdots & \dots & \vdots \\ \bar{x}_p & \dots & \bar{x}_p \end{bmatrix}$$

$$X - \tilde{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & \dots & x_{1N} - \bar{x}_1 \\ x_{21} - \bar{x}_2 & \dots & x_{2N} - \bar{x}_2 \\ \vdots & \dots & \vdots \\ x_{p1} - \bar{x}_p & \dots & x_{pN} - \bar{x}_p \end{bmatrix}. \tag{13.23}$$

Then the sample sum of products matrix, denoted by $S$, is given by the following:

$$S = [X - \tilde{X}][X - \tilde{X}]' = (s_{ij})$$

where

$$s_{ij} = \sum_{k=1}^{N} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j), \quad s_{ii} = \sum_{k=1}^{N} (x_{ik} - \bar{x}_i)^2. \tag{13.24}$$

Then the joint density of the sample values, for simple random sample of size $N$ (iid variables), denoted by $L(X, \mu, \Sigma)$ is the following:

$$\begin{aligned} L(X, \mu, \Sigma) &= \prod_{j=1}^{N} f(X_j) \\ &= \frac{1}{(2\pi)^{\frac{Np}{2}} |\Sigma|^{\frac{N}{2}}} \\ &\quad \times \exp\left\{ -\frac{1}{2} \sum_{j=1}^{N} (X_j - \mu)' \Sigma^{-1} (X_j - \mu) \right\}. \end{aligned} \tag{13.25}$$

The exponent can be simplified and written in the following form by using the fact that the trace of a 1 × 1 matrix or trace of a scalar quantity is itself. Also we have a general result on trace, namely, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ as long as the products $AB$ and $BA$ are defined, $AB$ need not be equal to $BA$.

$$\begin{aligned}\sum_{j=1}^{N}(X_j-\mu)'\Sigma^{-1}(X_j-\mu) &= \operatorname{tr}\left\{\sum_{j=1}^{N}(X_j-\mu)'\Sigma^{-1}(X_j-\mu)\right\}\\ &= \sum_{j=1}^{N}\operatorname{tr}(X_j-\mu)'\Sigma^{-1}(X_j-\mu)\\ &= \sum_{j=1}^{N}\operatorname{tr}\{\Sigma^{-1}(X_j-\mu)(X_j-\mu)'\}\\ &= \operatorname{tr}\left\{\Sigma^{-1}\left[\sum_{j=1}^{N}(X_j-\mu)(X_j-\mu)'\right]\right\}\\ &= \operatorname{tr}\{\Sigma^{-1}S\}\\ &\quad + N(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu). \end{aligned} \tag{13.26}$$

Then the maximum likelihood estimators (MLE) of $\mu$ and $\Sigma$ are

$$\hat{\mu}=\bar{X},\quad \hat{\Sigma}=\frac{S}{N}. \tag{13.27}$$

Those who are familiar with vector and matrix derivatives may do the following. Take $\ln L(X,\mu,\Sigma)$, operate with $\frac{\partial}{\partial\mu}$ and $\frac{\partial}{\partial\Sigma}$, equate to null vector and null matrix respectively and solve. Use the form as in (13.26), which will provide the solutions easily. For the sake of illustration, we will derive one likelihood ratio criterion or $\lambda$-criterion.

### 13.8.1 Testing the hypothesis of independence

Let the hypothesis be that the individual components $x_1,\dots,x_p$ are independently distributed. In the normal case, this will imply that the covariance matrix $\Sigma$ is diagonal, $\Sigma=\operatorname{diag}(\sigma_{11},\dots,\sigma_{pp})$, where $\sigma_{ii}=\sigma_i^2=\operatorname{Var}(x_i)$, $i=1,\dots,p$. In the whole parameter space $\Omega$, the MLE are $\hat{\mu}=\bar{X}$ and $\hat{\Sigma}=\frac{S}{N}$ and the supremum of the likelihood function is

$$\sup_{\Omega} L=\frac{e^{-\frac{Np}{2}}}{(2\pi)^{\frac{Np}{2}}|\frac{S}{N}|^{\frac{N}{2}}}.$$

Under the null hypothesis of independence of components, the likelihood function splits into product. Let $L_0$ denote $L$ under $H_0$. Then

$$L_0=\prod_{j=1}^{p}\frac{e^{-\frac{1}{2\sigma_j^2}\sum_{k=1}^{N}(x_{jk}-\mu_j)^2}}{[\sigma_j^2(2\pi)]^{\frac{N}{2}}}$$

which leads to the MLE as $\hat{\mu}_j=\bar{x}_j$, $\hat{\sigma}_j^2=\frac{s_{ii}}{N}$. Hence

$$\sup_{H_0} L=\frac{e^{-\frac{Np}{2}}}{(2\pi)^{\frac{Np}{2}}\prod_{j=1}^{p}(\hat{\sigma}_j^2)^{\frac{N}{2}}}.$$

Hence

$$\lambda = \frac{|\frac{S}{N}|^{\frac{N}{2}}}{\prod_{j=1}^{p}(\frac{s_{ii}}{N})^{\frac{N}{2}}} \quad \Rightarrow \quad u = c\lambda^{\frac{2}{N}} = \frac{|S|}{s_{11}\cdots s_{pp}}$$

where $c$ is a constant. The null distribution of $u = c\lambda^{\frac{2}{N}}$ is the distribution of $u$ when the population is $N_p(\mu, \Sigma) = \prod_{j=1}^{p} N(\mu_j, \sigma_j^2)$ and the non-null distribution of $u$ is the distribution of $u$ when the population is $N_p(\mu, \Sigma), \Sigma > 0$. These can be evaluated by using the real matrix-variate gamma distribution of $S$. This material is beyond the scope of this book.

## Exercises 13.8

**13.8.1.** If the $p \times 1$ vector $X$ has a non-singular $p$-variate normal distribution, $N_p(\mu, \Sigma)$, $\Sigma > 0$, write down the densities when (1) $\Sigma$ = a diagonal matrix, (2) $\Sigma = \sigma^2 I$ where $\sigma^2$ is a scalar quantity and $I$ is the identity matrix, (3) $\Sigma$ is a block diagonal matrix $\Sigma = \mathrm{diag}(\Sigma_{11}, \dots, \Sigma_{qq})$, $q \le p$.

**13.8.2.** If a simple random sample of size $N$ is available from a $N_p(\mu, \Sigma)$, $\Sigma > 0$, then derive the MLE of $\mu$ and $\Sigma$.

**13.8.3.** For the problem in Exercise 13.8.1, derive the MLE of the parameters under the special cases (1), (2) and (3) there.

**13.8.4.** Fill in the steps in the derivation of the result in (13.26).

**13.8.5.** Construct the $\lambda$-criterion for testing the hypotheses that $\Sigma$ is of the form of (2) and (3) of Exercise 13.8.1.

## 13.9 Some non-parametric tests

In previous sections, we have been dealing with hypotheses on the parameters of a pre-selected distribution. In other words, we have already selected a model (density or probability function) and we are concerned about one or more parameters in this selected model. Now, we will consider situations, which are not basically of this nature. Some parameters or models may enter into the picture at a secondary stage. Suppose that we want to test whether there is any association between two qualitative characteristics or two quantitative characteristics or one qualitative and one quantitative characteristic, such as the habit of wearing a tall hat (qualitative) and longevity of life (quantitative), intelligence (qualitative/quantitative) and weight (quantitative), color of eyes (qualitative) and behavior (qualitative), preference of certain types of clothes (qualitative) and beauty (qualitative), etc., then this is not of a parametric type that we have considered so far. Suppose that we have some data at hand, such as data on heights of students in a class and we would like to see whether height is normally distributed. Such problems usually fall in the category of tests known as “lack-of-fit” or “goodness-of-fit” tests. Here, we are concerned about the selection of a model for the data at hand. In a production process where there is a serial production of a machine part, some of the parts may be defective in the sense of not satisfying quality specifications. We would like to see whether the defective item is occurring at random. In a cloth weaving process, sometime threads get tangled and uniformity of the cloth is lost and we would like to see whether such an event is occurring at random or not. This is the type of situation where we want to test for randomness of an event. We will start with lack-of-fit tests first.

### 13.9.1 Lack-of-fit or goodness-of-fit tests

Two popular test statistics, which are used for testing goodness-of-fit or lack-of-fit of the selected model to given data, are Pearson’s $X^2$ statistic and the Kolmogorov–Smirnov statistic. We will examine both of these procedures. The essential difference is that for applying Pearson’s $X^2$ statistic we need data in a categorized form. We do not need the actual data points. We need only the information about the numbers of observations falling into various categories or classes. If the actual data points are available and if we wish to use Pearson’s $X^2$, then the data are to be categorized. This brings in arbitrariness. If different people categorize the data they may use different class intervals and these may result in contradictory conclusions in the applications of Pearson’s $X^2$ statistic. The Kolmogorov–Smirnov statistic makes use of the actual data points, and not applicable to the data which are already categorized in the form of a frequency table. [In a discrete situation, actual data will be of the form of actual points the random variable can take, and the corresponding frequencies, which is not considered to be a categorized data.] This is the essential difference. Both of the procedures make use of some sort of distance between the observed points and expected points, expected under the hypothesis or under the assumed model. Pearson’s $X^2$ and the Kolmogorov–Smirnov statistics make use of distance measures. Pearson’s statistic is of the following form:

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}, \tag{13.28}$$

where $o_i$ = observed frequency in the $i$-th class, $e_i$ = the expected frequency in the $i$-th class under the assumed model, for $i = 1, \dots, k$. It can be shown that (13.28) is a generalized distance between the vectors $(o_1, \dots, o_k)$ and $(e_1, \dots, e_k)$. It can also be shown that $X^2$ is approximately distributed as a chi-square with $k - 1 - s$ degrees of freedom when $s$ is the number of parameters estimated while computing the $e_i$’s. If no parameter is estimated, then the degrees of freedom is $k - 1$. This approximation is good only when $k \geq 5$, $e_i \geq 5$ for each $i = 1, \dots, k$. Hence (13.28) should be used under these conditions, otherwise the chi-square approximation is not good. The Kolmogorov–Smirnov statistics are the following:

$$D_n = \sup_x |S_n(x) - F(x)| \tag{13.29}$$

and

$$W^2 = E|S_n(x) - F(x)|^2 \tag{13.30}$$

where $S_n(x)$ is the sample distribution function (discussed in Section 11.5.2) and $F(x)$ is the population distribution function or the distribution function under the assumed model or assumed distribution. In $W^2$, the expected value is taken under the hypothesis or under $F(x)$. Both $D_n$ and $(W^2)^{\frac{1}{2}}$ are mathematical distances between $S_n(x)$ and $F(x)$, and hence Kolmogorov–Smirnov tests are based on actual distance between the observed values represented by $S_n(x)$ and expected values represented by $F(x)$. For example, suppose that we have data on waiting times in a queue. Our hypothesis may be that the waiting time $t$ is exponentially distributed with expected waiting time 10 minutes, time being measured in minutes. Then the hypothesis $H_0$ says that the underlying distribution is

$$H_0: \quad f(x) = \frac{1}{10} \mathrm{e}^{-\frac{t}{10}}, \quad t \geq 0 \text{ and zero elsewhere.} \tag{13.31}$$

Then, under $H_0$, the population distribution function is

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \int_0^x \frac{\mathrm{e}^{-\frac{t}{10}}}{10}\mathrm{d}t = 1 - \mathrm{e}^{-\frac{x}{10}}.$$

Suppose that the number of observations in the class $0 \leq t \leq 2$ is 15 or it is observed that the waiting time of 15 people in the queue is between 0 and 2 minutes. Then we may say that the observed frequency in the first interval $o_1 = 15$. What is the corresponding expected value $e_1$? Suppose that the total number of observations is 100 or waiting times of 100 people are taken. Then the expected value $e_1 = 100p_1$, where $p_1$ is the probability of finding a person whose waiting time $t$ is such that $0 \leq t \leq 2$. Under our assumed model, we have

$$e_1 = 100 \times \int_0^2 \frac{\mathrm{e}^{-\frac{t}{(10)}}}{10}\mathrm{d}t = 100[1 - \mathrm{e}^{-\frac{2}{(10)}}] = 100[1 - \mathrm{e}^{-\frac{1}{5}}].$$

In general, we have the following situation:

| Classes | 1 | 2 | … | $k$ |
|---|---|---|---|---|
| Observed frequencies | $n_1$ | $n_2$ | … | $n_k$ |
| Expected frequencies | $e_1 = np_1$ | $e_2 = np_2$ | … | $e_k = np_k$ |

where $n = n_1 + \cdots + n_k$ = total frequency and $p_i$ is the probability of finding an observation in the $i$-th class, under the hypothesis, $i = 1, \dots, k$, $p_1 + \cdots + p_k = 1$. Then

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \approx \chi^2_{k-1}. \tag{13.32}$$

For $k \geq 5$, $np_i \geq 5$, $i = 1, \dots, k$ we have a good approximation in (13.26).

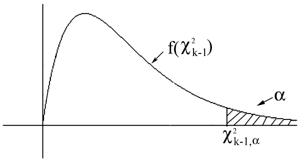


**Figure 13.14:** Critical region in Pearson's $X^2$ test.

When using a distance measure for testing the goodness-of-fit of a model, we will always reject if the distance is large or we reject when the observed $X^2$ is large. Hence, when using $X^2$ statistic the criterion is to reject $H_0$ if the observed $X^2 \geq \chi^2_{k-1,\alpha}$, for a test at the level of rejection $\alpha$ as shown in Figure 13.14, where

$$\Pr\{X^2 \geq \chi^2_{k-1,\alpha}\} = \alpha.$$

**Example 13.10.** A tourist resort is visited by tourists from many countries. The resort operator has the following data in the month of January:

| Country | USA | UK | Canada | Italy | Germany | France | Asia |
|---|---|---|---|---|---|---|---|
| Frequency | 22 | 12 | 18 | 10 | 20 | 18 | 30 |

The resort operator has a hypothesis that the proportions of tourists visiting in the month of January of any year is 2 : 1 : 2 : 1 : 2 : 2 : 3. Test this hypothesis, at a 5% level of rejection.

**Solution 13.10.** The total frequency $n = 22 + 12 + 18 + 10 + 20 + 18 + 30 = 130$. Under the hypothesis, the expected frequencies are the following: For the USA, it is 2 out of $13 = 2 + 1 + 2 + 1 + 2 + 2 + 3$ of 130 or $130 \times \frac{2}{13} = 20$. Calculating the expected frequencies like this, we have the following table:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Observed frequency | 22 | 12 | 18 | 10 | 20 | 18 | 30 |
| Expected frequency | 20 | 10 | 20 | 10 | 20 | 20 | 30 |

The degrees of freedom for Pearson's $X^2$ is $k - 1 = 7 - 1 = 6$. From a chi-square table, $\chi^2_{6,0.05} = 12.59$. Also our conditions $k \geq 5$, $e_i \geq 5$, $i = 1, \dots, 7$ are satisfied, and hence a good chi-square approximation can be expected. The observed value of $X^2$ is given by

$$
\begin{aligned}
X^2 &= \frac{(20-20)^2}{20} + \frac{(12-10)^2}{10} + \frac{(18-20)^2}{20} \\
&\quad + \frac{(10-10)^2}{10} + \frac{(20-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(30-30)^2}{30} \\
&= \frac{4}{20} + \frac{4}{10} + \frac{4}{20} + \frac{4}{20} + 0 + 0 + 0 = \frac{20}{20} = 1.
\end{aligned}
$$

The observed value of $X^2$ is not greater than the tabulated value 12.59, and hence we cannot reject $H_0$. Does it mean that our model is a good fit or our hypothesis can be "accepted"? Remember that within that distance of less than 12.59 units there could be several distributions, and hence "accepting the claim" is not a proper procedure. For example, try the proportions 3 : 1 : 1 : 1 : 2 : 2 : 3 and see that the hypothesis is not rejected. At the most what we can say is only that the data seem to be consistent with the hypothesis of resort owner's claim of the proportions.

**Note 13.9.** In statistical terminology, our hypothesis in Example 13.10 was on the multinomial probabilities, saying that the multinomial probabilities are $p_1 = \frac{2}{13}$, $p_2 = \frac{1}{13}, \dots, p_7 = \frac{3}{13}$, in a 6-variate multinomial probability law.

**Example 13.11.** Test the goodness-of-fit of a normal model, $x \sim N(\mu = 80, \sigma^2 = 100)$, $x$ = grade obtained by students in a particular course, to the following data, at 5% level of rejection.

| | | | |
|---|---|---|---|
| Classes | $x < 50$ | $50 \le x < 60$ | $60 \le x < 70$ |
| Frequency | 225 | 220 | 235 |
| Classes | $70 \le x < 80$ | $80 \le x < 90$ | $90 \le x \le 100$ |
| Frequency | 240 | 230 | 220 |

**Solution 13.11.** Total frequency $n = 1370$. Let

$$
f(x) = \frac{1}{10(\sqrt{2\pi})} \mathrm{e}^{-\frac{1}{2\times 100}(x-80)^2}, \quad y = \frac{x-80}{10}, \quad g(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}.
$$

The probability $p_1$ of finding an observation in the interval $-\infty < x < 50$ is given by

$$
p_1 = \int_{-\infty}^{50} f(x)\mathrm{d}x = \int_{-\infty}^{-3} g(y)\mathrm{d}y = 0.0014
$$

from $N(0,1)$ tables, where $y = \frac{50-80}{10} = -3$. We have the following results from the computations:

$$
\begin{aligned}
p_1 &= 0.0014 \\
e_1 &= np_1 = 1370 \times 0.0014 = 3.84 \\
p_2 &= \int_{50}^{60} f(x)\mathrm{d}x = \int_{-3}^{-2} g(y)\mathrm{d}y = 0.4998 - 0.4773 = 0.0213 \\
e_2 &= np_2 = 1370 \times 0.0213 = 29.18
\end{aligned}
$$

$$
\begin{aligned}
p_3 &= \int_{60}^{70} f(x)\mathrm{d}x = \int_{-2}^{-1} g(y)\mathrm{d}y = 0.4773 - 0.3414 = 0.1359 \\
e_3 &= np_3 = 1370 \times 0.1359 = 186.18 \\
p_4 &= \int_{70}^{80} f(x)\mathrm{d}x = \int_{-1}^{0} g(y)\mathrm{d}y = 0.3414 - 0 = 0.3414 \\
e_4 &= np_4 = 1370 \times 0.3414 = 467.72 \\
p_5 &= \int_{80}^{90} f(x)\mathrm{d}x = \int_{0}^{1} g(y)\mathrm{d}y = 0.3414 = 0.3414 \\
e_5 &= np_5 = 1370 \times 0.3414 = 467.72 \\
p_6 &= \int_{90}^{100} f(x)\mathrm{d}x = \int_{1}^{2} g(y)\mathrm{d}y = 0.1359 \\
e_6 &= np_6 = 1370 \times 0.1359 = 186.18.
\end{aligned}
$$

Since $e_1 < 5$, we may combine the first and second classes to make the expected frequency greater than 5 to have a good approximation. Then add up the observed frequencies in the first two classes $o_1 + o_2 = 225 + 220 = 445$, call it $o_2'$ and expected frequencies $e_1 + e_2 = 3.84 + 29.18 = 33.02 = e_2'$. Thus, the effective number of classes is now $6 - 1 = 5$ still the condition $k \geq 5$ is satisfied. The degrees of freedom is reduced by one and the new degrees of freedom is $k' - 1 = 6 - 1 = 5$, $k' = k - 1$. The tabulated value of $\chi^2_{k'-1,\alpha} = \chi^2_{4,0.05} = 9.49$. The observed value of $X^2$ is given by

$$
\begin{aligned}
X^2 = {} & \frac{(446.33 - 33.02)^2}{33.02} + \frac{(235 - 186.18)^2}{186.18} + \frac{(240 - 467.72)^2}{467.72} \\
& + \frac{(230 - 467.72)^2}{467.72} + \frac{(220 - 186.18)^2}{186.18}.
\end{aligned}
$$

We have to see only whether the observed $X^2$ is greater than 9.49 or not. For doing this, we do not have to compute every term in $X^2$. Start with the term involving the largest deviation first, then the second largest deviation, etc. The term with the largest deviation is the first term. Hence let us compute this first:

$$
\frac{(445 - 33.02)^2}{33.02} > 9.49
$$

and hence the hypothesis is rejected. We do not have to compute any other term.

**Note 13.10.** When this particular normal distribution with $\mu = 80$ and $\sigma^2 = 100$ is rejected that does not mean that other normal distributions are also rejected. We can try other normal populations with specified $\mu$ and $\sigma^2$. We can also try the goodness-of-fit of any normal population. In this case, we do not know $\mu$ and $\sigma^2$. Then we estimate $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$ and try to fit $N(\hat{\mu}, \hat{\sigma}^2)$. MLE or moment estimates can be used. In this case, two degrees of freedom for the chi-square are lost. In our example, the effective degrees of freedom will be then $k' - 1 - 2 = 5 - 2 = 3$, $k' = k - 1$ due to combining two classes, and then our approximation is not good also.

**Note 13.11.** The testing procedure based on Pearson's $X^2$ statistic and other so called "goodness-of-fit" tests are only good for rejecting the hypothesis, which defeats the purpose of the test. Usually, one goes for such tests to claim that the selected model is a good model. Testing procedure defeats this purpose. If the test did not reject the hypothesis, then the maximum claim one can make is only that the data seem to be consistent with the hypothesis or simply claim that the discrepancy, measured by $X^2$, is small, and hence we may take the model as a good model, remembering that several other models would have given the same or smaller values of $X^2$.

**Note 13.12.** In our problem, the variable $x$ is a Gaussian random variable, which ranges from $-\infty$ to $\infty$. But our data are only for $x \le 100$. Logically, we should have taken the last cell or last interval as $90 \le x < \infty$ or $x \ge 90$, instead of taking $90 \le x \le 100$.

### 13.9.2 Test for no association in a contingency table

A two-way contingency table is a table of frequencies where individuals or items are classified according to two qualitative or two quantitative or one qualitative and one quantitative characteristics. For example, if a random sample of 1, 120 people are classified according to their ability to learn statistics and their weights and suppose that we have the following two-way contingency table:

**Example 13.12.** Test for no association between the characteristics of classification in the following two-way contingency table, where 1, 120 people are classified according to their ability to learn statistics and weights. Test at a 5% level of rejection, where $I$ = excellent, $II$ = very good, $III$ = good, $IV$ = poor, $W_1$ = <50 kg, $W_2$ = 50 – 60 kg, $W_3$ = >60 kg.

| Weight → | $W_1$ | $W_2$ | $W_3$ |
|---|---|---|---|
| Ability ↓ | | | |
| $I$ | $50 = n_{11}(p_{11})$ | $100 = n_{12}(p_{12})$ | $120 = n_{13}(p_{13})$ |
| $II$ | $100 = n_{21}(p_{21})$ | $120 = n_{22}(p_{22})$ | $80 = n_{23}(p_{23})$ |
| $III$ | $80 = n_{31}(p_{31})$ | $90 = n_{32}(p_{32})$ | $100 = n_{33}(p_{33})$ |
| $IV$ | $90 = n_{41}(p_{41})$ | $100 = n_{42}(p_{42})$ | $90 = n_{43}(p_{43})$ |
| Sum | $n_{.1} = 320(p_{.1})$ | $n_{.2} = 410(p_{.2})$ | $n_{.3} = 390(p_{.3})$ |

Due to overflow in the page, the last column is given below:

Row sum

$$\begin{aligned}
n_{1.} &= 270(p_{1.})\\
n_{2.} &= 270(p_{2.})\\
n_{3.} &= 270(p_{3.})\\
n_{4.} &= 280(p_{4.})\\
n_{..} &= 1120(p_{..} = 1)
\end{aligned}$$

Here, the number of people or frequency in the $i$-th ability group and $j$-th weight group, or in the $(i,j)$-th cell, is denoted by $n_{ij}$. The probability of finding an observation in the $(i,j)$-th cell is denoted by $p_{ij}$. These are given in the brackets in each cell. The following standard notations are used. These notations will be used in model building also. A summation with respect to a subscript is denoted by a dot. Suppose that $i$ goes from 1 to $m$ and $j$ goes from 1 to $n$. Then

$$n_{i.} = \sum_{j=1}^{n} n_{ij}; \quad n_{.j} = \sum_{i=1}^{m} n_{ij}; \quad \sum_{i=1}^{m}\sum_{j=1}^{n} n_{ij} = n_{..} = \text{grand total}. \tag{13.33}$$

Similar notations are used on $p_{ij}$'s so that the total probability $p_{..} = 1$. In the above table of observations, the row sums of frequencies are denoted by $n_{1.}, n_{2.}, n_{3.}, n_{4.}$ and the column sums of frequencies are denoted by $n_{.1}, n_{.2}, n_{.3}$.

**Solution 13.12.** If the two characteristics of classification are independent or in the sense that there is no association between the two characteristics of classification, then the probability in the $(i,j)$-th cell is the product of the probabilities of finding an observation in the $i$-th row and in the $j$-th column or $p_{ij} = p_{i.}p_{.j}$. Note that in a multinomial probability law the probabilities are estimated by the corresponding relative frequencies. For example, $p_{i.}$ and $p_{.j}$ are estimated by

$$\hat{p}_{i.} = \frac{n_{i.}}{n_{..}}, \quad \hat{p}_{.j} = \frac{n_{.j}}{n_{..}} \tag{13.34}$$

and, under the hypothesis of independence of the characteristics of classification, the expected frequency in the $(i,j)$-th cell is estimated by

$$\hat{e}_{ij} = n_{..}\hat{p}_{i.}\hat{p}_{.j} = n_{..} \times \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}} = \frac{n_{i.}n_{.j}}{n_{..}}. \tag{13.35}$$

Hence, multiply by the marginal totals and then divide by the grand total to obtain the expected frequency in each cell, under the hypothesis of independence of the characteristics of classification. For example, in the first row, first column or $(1,1)$-th cell the expected frequency is $\frac{270\times 320}{1120} \approx 77.14$. The following table gives the observed frequencies and the expected frequencies. The expected frequencies are given in the brackets.

| | | |
|---|---|---|
| 50(77.14) | 100(98.84) | 120(94.02) |
| 100(85.71) | 120(109.82) | 80(104.46) |
| 80(77.14) | 90(98.84) | 100(94.02) |
| 90(80) | 100(102.5) | 90(97.5) |

Then an observed value of Pearson's $X^2$ statistic is the following:

$$\begin{aligned} X^2 = {} & \frac{(50-77.14)^2}{77.14} + \frac{(100-98.84)^2}{98.84} + \frac{(120-94.02)^2}{94.02} \\ & + \frac{(100-85.71)^2}{85.71} + \frac{(120-109.02)^2}{109.02} + \frac{(80-104.46)^2}{104.46} \\ & + \frac{80-77.14)^2}{77.14} + \frac{(90-98.84)^2}{98.84} + \frac{(100-94.02)^2}{94.02} \\ & + \frac{(90-80)^2}{80} + \frac{(100-102.5)^2}{102.5} + \frac{(90-97.5)^2}{97.5} \end{aligned}$$

In general, what is the degrees of freedom of Pearson's $X^2$ statistic in testing hypothesis of independence in a two-way contingency table? If there are $m$ rows and $n$ columns, then the total number of cells is $mn$ but we have estimated $p_{i.}$, $i = 1,\dots,m$ which gives $m-1$ parameters estimated. Similarly, $p_{.j}$, $j = 1,\dots,n$ gives $n-1$ parameters estimated because in each case the total probability $p_{..} = 1$. Thus the degrees of freedom is $mn-(m-1)-(n-1)-1 = (m-1)(n-1)$. In our case above, $m = 4$ and $n = 3$, and hence the degrees of freedom is $(m-1)(n-1) = (3)(2) = 6$. The chi-square approximation is good when the expected frequency in each cell $e_{ij} \geq 5$ for all $i$ and $j$ and $mn \geq 5$. In our example above, the conditions are satisfied and we can expect a good chi-square approximation for Pearson's $X^2$ statistic or

$$X^2 \approx \chi^2_{(m-1)(n-1)}, \quad \text{for } mn \geq 5, \quad e_{ij} \geq 5 \quad \text{for all } i \text{ and } j. \tag{13.36}$$

In our example, if we wish to test at a 5% level of rejection, then the tabulated value of $\chi^2_{6,0.05} = 12.59$. Hence it is not necessary to compute each and every term in $X^2$. Compute the terms with the largest deviations first. The (1, 1)-th term gives $\frac{(50-77.14)^2}{77.14} = 9.55$. The (1, 3)-th term gives $\frac{(120-94.02)^2}{94.02} = 7.18$. Hence the sum of these two terms alone exceeded the critical point 12.59 and hence we reject the hypothesis of independence of the characteristics of classification here, at a 5% level of rejection.

**Note 13.13.** Rejection of our hypothesis of independence of the characteristics of classification does not mean that there is association between the characteristics. In the beginning stages of the development of statistics as a discipline, people were making all sorts of contingency tables and claiming that there were association between characteristics of classification such as the habit of wearing tall hats and longevity of life, etc. Misuses went to the extent that people were claiming that "anything and everything could be proved by statistical techniques". Remember that nothing is established or proved by statistical techniques, and as remarked earlier, that non-rejection of $H_0$ cannot be given any meaningful interpretations because the statistical procedures do not support or justify to make any claim if $H_0$ is not rejected.

### 13.9.3 Kolmogorov–Smirnov statistic $D_n$

$D_n$ is already stated in (13.29), which is,

$$D_n = \sup_x |S_n(x) - F(x)|$$

where $S_n(x)$ is the sample distribution function and $F(x)$ is the population distribution, under the population assumed by the hypothesis. Let the hypothesis $H_0$ be that the underlying distribution is continuous, has density $f(x)$ and distribution function $F(x)$, such as $f(x)$ is an exponential density. Then $F(x)$ will produce a continuous curve and $S_n(x)$ will produce a step function as shown in Figure 13.15.

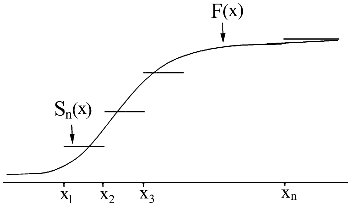


**Figure 13.15:** Sample and population distribution functions $S_n(x)$ and $F(x)$.

We will illustrate the computations with a specific example.

**Example 13.13.** Check to see whether the following data could be considered to have come from a normal population $N(\mu = 13, \sigma^2 = 1)$. Data: 16, 15, 15, 9, 10, 10, 12, 12, 11, 13, 13, 13, 14, 14.

**Solution 13.13.** Here, $H_0$: is that the population density is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-13)^2}$$

and, therefore,

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

At $x = 9$,

$$F(9) = \int_{-\infty}^{9} f(t)\mathrm{d}t = \int_{-\infty}^{-4} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\mathrm{d}y = 0.0000$$

from $N(0,1)$ tables. For $x = 10$, $F(10) = 0.0013$ from $N(0,1)$ tables. But at $x = 9$ the sample distribution function is $S_n(x) = s_{14}(9) = \frac{1}{14} = 0.0714$ and this remains the same from $9 \le x < 10$. Hence, theoretically we have $|S_n(x) - F(x)| = 0.0714 - 0.0000 = 0.0714$ at $x = 9$. But $\lim_{x\to 10_-} S_n(x) = \frac{1}{14}$ , and hence we should take the difference at the point $x = 10$ also, which is $|S_{14}(9) - F(10)| = |0.0714 - 0.0013| = 0.0701$. Hence, for each interval we should record these two values and take the largest of all such values to obtain an observed value of $D_n$. The computed values are given in the following table:

| $x$ | Frequency | $F(x)$ | $S_{14}(x)$ |
|---|---|---|---|
| 9 | 1 | 0.0000 | $1/14 = 0.0714,\ 9 \le x < 10$ |
| 10 | 2 | 0.0013 | $3/14 = 0.2142,\ 10 \le x < 11$ |
| 11 | 1 | 0.0228 | $4/14 = 0.2856,\ 11 \le x < 12$ |
| 12 | 2 | 0.1587 | $6/14 = 0.4284,\ 112 \le x < 13$ |
| 13 | 3 | 0.5000 | $9/14 = 0.6426,\ 13 \le x < 14$ |
| 14 | 2 | 0.8413 | $11/14 = 0.7854,\ 14 \le x < 15$ |
| 15 | 2 | 0.9772 | $13/14 = 0.9282,\ 15 \le x < 16$ |
| 16 | 1 | 0.9987 | $14/14 = 1.0000,\ 16 \le x < \infty$ |

In the following table, we have two points for each interval. These are given against the $x$-values

| $x$ | $\lvert S_{14}(x) - F(x)\rvert$ |
|---|---|
| 9 | 0.0714, 0.0701 |
| 10 | 0.2129, 0.1914 |
| 11 | 0.2628, 0.1269 |
| 12 | **0.2697**, 0.0716 |
| 13 | 0.1426, 0.1987 |
| 14 | 0.1987, 0.1918 |
| 15 | 0.1918, 0.0705 |
| 16 | 0.0705, 0.0000 |

The largest of the entries in the last two columns is 0.2697, and hence the observed value of $D_{14} = 0.2697$. Tables of $D_n$ are available. The tabled value of $D_{14} = 0.35$. We reject the hypothesis only when the distance is large or when the observed $D_n$ is bigger than the tabulated $D_n$. In the above example, the hypothesis is not rejected since the observed value 0.2697 is not greater than the tabulated value 0.35.

**Note 13.14.** When considering goodness-of-fit tests, we have assumed the underlying populations to be continuous. How do we test the goodness-of-fit of a discrete distribution to the given data? Suppose that the data is the following: $3, 3, 5, 5, 5, 5, 6, 6$ or $x = 3$ with frequency $n_1 = 2$, $x = 5$ with frequency $n_2 = 4$, $x_3 = 6$ with frequency $n_3 = 2$. Whatever is seen here is the best fitting discrete distribution, namely,

$$f(x) = \begin{cases} 2/8, & x = 3 \\ 4/8, & x = 5 \\ 2/8, & x = 6 \\ 0, & \text{elsewhere.} \end{cases}$$

There is no better fitting discrete distribution to this data. Hence, testing goodness-of-fit of a discrete distribution to the data at hand does not have much meaning.

There are a few other non-parametric tests called sign test, rank test, run test, etc. We will give a brief introduction to these. For more details, the students must consult books on non-parametric statistics.

### 13.9.4 The sign test

The sign test is applicable when the population is continuous and when we know that the underlying population is symmetric about a parameter $\theta$ or the population density $f(x)$ is symmetric about $x = \theta$, such as a normal population, which is symmetric about $x = \mu$ where $\mu = E(x)$. We wish to test a hypothesis on $\theta$. Let $H_0 : \theta = \theta_0$ (given). Then under $H_0$ the underlying population is symmetric about $x = \theta_0$. In this case, the probability of getting an observation from this population less than $\theta_0$ is $\frac{1}{2}$ = the probability of getting an observation greater than $\theta_0$, and due to continuity, the probability of getting an observation equal to $\theta_0$ is zero. Hence finding an observation above $\theta_0$ can be taken as a Bernoulli success and finding an observation below $\theta_0$ as a failure, or vice versa. Therefore, the procedure is the following: Delete all observations equal to $\theta_0$. Let the resulting number of observations be $n$. Put a plus sign for an observation above $\theta_0$ and put a minus sign for observations below $\theta_0$. Count the number of + signs. This number of + signs can be taken as the number of successes in $n$ Bernoulli trials. Hence we can translate the hypothesis $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$ into $H_0 : p = \frac{1}{2}, H_1 : p \neq \frac{1}{2}$ where $p$ is the probability of success in a Bernoulli trial. Then the test criterion for a test at the level of rejection $\alpha$ can be stated as follows: Reject $H_0$ if the observed number of plus signs is small or large. For convenience, we may cut off equal tail probabilities $\frac{\alpha}{2}$ at both ends. Let $y$ be the number of plus signs then compute $y_0$ and $y_1$ such that

$$\sum_{y=0}^{y_0} \binom{n}{y} (0.5)^y (1-0.5)^{n-y} \leq \frac{\alpha}{2} \tag{a}$$

and

$$\sum_{y=y_1}^{n} \binom{n}{y} (0.5)^y (1-0.5)^{n-y} \leq \frac{\alpha}{2}. \tag{b}$$

If the observed number of plus signs is less than $y_0$ or greater than $y_1$, then reject $H_0$. Since the test is based on signs, it is called a *sign test*.

**Example 13.14.** The following is the data on the yield of wheat from 12 test plots: $5, 1, 8, 9, 11, 4, 7,\ 12.5, 6, 8, 9$. Assume that the population is symmetric about $\mu = E(x)$ where $x$ is the yield in a test plot. Test the hypothesis that $\mu = 9$ at the level of rejection of 5%.

**Solution 13.14.** In our notation, $\theta_0 = 9$. There are two observations equal to 9, and hence delete these and there are 10 remaining observations. Mark the observations bigger than 9 by a plus sign: 11(+), 12(+). There are two observations bigger than 9 and then the observed number of successes in 10 Bernoulli trials is 2. Here, $\alpha = 0.05$ or $\frac{\alpha}{2} = 0.025$. From a binomial table for $p = \frac{1}{2}$ , we have $y_0 = 1$ and $y_1 = 9$, which are the closest points where the probability inequalities in (a) and (b) above are satisfied. Our observed value is 2 which is not in the critical region, and hence the hypothesis is not rejected at the 5% level of rejection.

**Note 13.15.** Does it mean that there is a line of symmetry at $x = 9$ for the distribution? For the same data if $H_0$ is $\theta_0 = 8$, then again $n = 10$ and the observed number of successes is 4 and $H_0$ is not rejected. We can have many such values for $\theta_0$ and still the hypotheses will not be rejected. That does not mean that the underlying distribution has symmetry at all these points. Besides, $p = \frac{1}{2}$ is not uniquely determining a line of symmetry. Hence trying to give an interpretation for non-rejection is meaningless. Due to this obvious fallacy some people modify the hypothesis saying that at $x = \theta_0$ there is the median of the underlying population. Again, $p = \frac{1}{2}$ does not uniquely determine $x = \theta_0$ as the median point. There could be several points qualifying to be the median. Hence non-rejection of $H_0$ cannot be given a justifiable interpretation.

### 13.9.5 The rank test

This is mainly used for testing the hypothesis that two independent populations are identical, against the alternative that they are not identical. If $x$ is the typical yield of ginger from a test plot under organic fertilizer and $y$ is the yield under a chemical fertilizer, then we may want to claim that $x$ and $y$ have identical distributions, whatever be the distributions. Our observations may be of the following forms: $n_1$ observations on $x$ and $n_2$ observations on $y$ are available. Then for applying a rank test the procedure is the following: Pool the observations on $x$ and $y$ and order them according to their magnitudes. Give to the smallest observation rank 1, the second smallest rank 2 and the last one rank $n_1 + n_2$. If two or more observations have the same magnitude, then give the average rank to each. For example, if there are two smallest numbers then each of these numbers gets the ranks $\frac{(1+2)}{2} = 1.5$ and the next number gets the rank 3, and so on. Keep track of the ranks occupied by each sample. A popular test statistic based on ranks is the Mann–Whitney $u$-test where

$$u = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{13.37}$$

where $n_1$ and $n_2$ are the sample sizes and $R_1$ is the sum of ranks occupied by the sample with size $n_1$. Under the hypothesis of identical distributions for $x$ and $y$, it can be shown that the mean value and the variance are $E(u) = \frac{n_1 n_2}{2}$ and $\mathrm{Var}(u) = \sigma_u^2 = n_1 n_2 \frac{(n_1+n_2+1)}{12}$ and that

$$v = \frac{u - E(u)}{\sigma_u} \approx N(0,1) \tag{13.38}$$

or $v$ is approximately a standard normal, and a good approximation is available for $n_1 \geq 8$, $n_2 \geq 8$. Hence we reject the hypothesis if the observed value of $|v| \geq z_{\frac{\alpha}{2}}$ [see Figure 13.5].

**Example 13.15.** The following are the observations on waiting times, $x$, at one checkout counter in a departmental store on 10 randomly selected occasions: $10, 5, 18, 12, 3, 8, 5, 8, 9, 12$. The following are the waiting times, $y$, on randomly selected 8 occasions at another checkout counter of the same store: $2, 5, 3, 4, 6, 5, 6, 9$. Test the hypotheses, at a 5% level of rejection, that $x$ and $y$ are identically distributed.

**Solution 13.15.** Since the size of the second sample is smaller, we will take that as the one for computing the sum of the ranks. Then in our notation, $n_1 = 8$, $n_2 = 10$. Let us pool and order the numbers. A subscript $a$ is put for numbers coming from the sample with size $n_1$ for identification. The following table gives the numbers and the corresponding ranks:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Numbers | $2_a$, | $3_a$, | 3, | $4_a$, | $5_a$, | $5_a$, | 5, | 5, | $6_a$ |
| Ranks | $1_a$ | $2.5_a$ | $2.5_a$ | $4_a$ | $6.5_a$ | $6.5_a$ | 6.5 | 6.5 | $9.5_a$ |
| Numbers | $6_a$, | 8, | 8, | $9_a$, | 9, | 10, | 12, | 12, | 18. |
| Ranks | $9.5_a$ | 11.5 | 11.5 | $13.5_a$ | 13.5 | 15 | 16.5 | 16.5 | 18 |

Total number of ranks occupied by the sample of size $n_1$, and the observed values of other quantities are the following:

$$R_1 = 1.0 + 2.5 + +2.5 + 4.0 + 6.5 + 6.5 + 9.5 + 9.5 + 13.5 = 55.5;$$

$$E(u) = \frac{(8)(10)}{2} = 40;$$

$$\sigma_u^2 = \frac{(8)(10)(8+10+1)}{12} \approx 126.67, \quad \sigma_u \approx 11.25;$$

$$u = (8)(10) + \frac{(8)(9)}{2} - 55.5 = 61.0; \quad \frac{u - E(u)}{\sigma_u} = \frac{61 - 40}{11.25} \approx 1.87.$$

At the 5% level of rejection in a standard normal case, the critical point is 1.96, and hence we do not reject the hypothesis here.

**Note 13.16.** Again, non-rejection of the hypothesis of identical distribution cannot be given meaningful interpretation. The same observations could have been obtained if the populations were not identically distributed. Observed values of $u$ and $R_1$ or the formula in (13.38) do not characterize the property of identical distributions for the underlying distributions. Hence, in this test as well as in the other tests to follow, non-rejection of the hypothesis should not be given all sorts of interpretations.

Another test based on the rank sums, which can be used in testing the hypotheses that $k$ given populations are identical, is the *Kruskal–Wallis H-test*. Suppose that the samples are of sizes $n_i$, $i = 1, \dots, k$. Again, pool the samples and order the observations from the smallest to the largest. Assign ranks from 1 to $n = n_1 + \cdots + n_k$, distributing the averages of the ranks when some observations are repeated. Let $R_i$ be the sum of the ranks occupied by the $i$-th sample. Then

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \approx \chi^2_{k-1} \tag{13.39}$$

where $n = n_1 + \cdots + n_k$, and $H$ is approximately a chi-square with $k - 1$ degrees of freedom under the null hypothesis that all the $k$ populations are identical. The approximation is good for $n_i \geq 5$, $i = 1, \dots, k$. Here, we reject the hypothesis for large values of $H$ only.

### 13.9.6 The run test

This test is usually used for testing randomness of an event. Suppose that in a production queue, an item is produced in succession. If the item satisfies the quality specifications, then we call it a good item, denoted by $G$ and if the item does not satisfy the quality specifications then we call it defective, denoted by $D$. Hence the production queue can be described by a chain of the letters $G$ and $D$, such as $GGGDGGGDDGG$, etc. A succession of identical symbol is called a run. In our example, there is one run of size 3 of $G$, then one run of size 1 of $D$ then one run of size 3 of $G$, then one run of size 2 of $D$, then one run of size 2 of $G$. Thus there are 5 runs here. The number of times the symbol $G$ appears is $n_1 = 8$ and the number of times the symbol $D$ appears is $n_2 = 3$ here. Consider a sequence of two symbols, such as $G$ and $D$, where the first symbol appears $n_1$ times and the second symbol appears $n_2$ times. [Any one of the two symbols can be called the first symbol and the other the second symbol.] Let the total number of runs in the sequence be $R$. Then under the hypothesis that the first symbol (or the second symbol) is appearing in the sequence at random or it is random occurrence and does not follow any particular pattern, we can compute the expected value and variance of $R$. Then it can be shown that the standardized $R$ is approximately a standard normal for large $n_1$ and $n_2$. The various quantities, under the hypothesis of randomness, are the following:

$$E(R) = \frac{2n_1n_2}{n_1+n_2} + 1; \quad \sigma_R^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)},$$
$$T = \frac{R - E(R)}{\sigma_r} \approx N(0,1) \quad \text{for } n_1 \geq 10,\ n_2 \geq 10. \tag{13.40}$$

The ideas will be clear from the following example. Note that we reject the null hypothesis of randomness if $|T| \geq z_{\frac{\alpha}{2}}$ for a test at the level of rejection $\alpha$.

**Example 13.16.** Test the hypothesis that in a production process, where an item can be $D$ = defective and $G$ = good, the defective items or $D$'s are appearing at random, based on the following observed sequence of $D$'s and $G$'s, at the 5% level of rejection: Observed sequence

$$GGGDDGGDGGGGGDDDGGGGDDGDDD$$

**Solution 13.16.** Let $n_1$ be the number of $D$'s and $n_2$ the number of $G$'s in the given sequence. Here, $n_1 = 11$, $n_2 = 15$. The number of runs $R = 10$. Under the hypothesis of randomness of $D$, the observed values are the following:

$$E(R) = \frac{2(11)(15)}{11+15} + 1 \approx 13.69; \quad \sigma_R^2 = \frac{2(11)(15)(2(11)(15) - 11 - 15)}{(11+15)^2(11+15-1)}$$
$$\approx 5.94; \quad |T| = \left|\frac{R - E(R)}{\sigma_r}\right| = \left|\frac{10 - 13.69}{\sqrt{5.94}}\right| \approx 1.51.$$

Here, $\alpha = 0.05$ or $\frac{\alpha}{2} = 0.025$ and then $z_{0.025} = 1.96$ from standard normal tables. The observed value of $|T|$ is 1.51, which is not greater than 1.96, and hence the hypothesis is not rejected.

## Exercises 13.9

**13.9.1.** A bird watcher reported that she has spotted birds belonging to 6 categories in a particular bird sanctuary and her claim is that these categories of birds are frequenting this sanctuary in the proportions 1 : 1 : 2 : 3 : 1 : 2 in these 6 categories of birds. Test this claim, at 5% level of rejection, if the following data are available:

| Category | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency | 6 | 7 | 13 | 17 | 6 | 10 |

**13.9.2.** The telephone calls received by a switchboard in successive 5-minute intervals is given in the following table:

| Number of calls = $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Number of intervals (frequency) | 25 | 30 | 20 | 15 | 10 | 8 | 10 |

Test the hypothesis that the data is compatible with the assumption that the telephone calls are Poisson distributed. Test at a 5% level of rejection.

**13.9.3.** The following table gives the increase in sales in a particular shop after placing advertisements. Test the "goodness-of-fit" of an exponential distribution to this data, at a 5% level of rejection.

| Increase | 5 – 10 | 11 – 15 | 16 – 22 | 23 – 27 | 28 – 32 | > 33 |
|---|---|---|---|---|---|---|
| Frequency | 200 | 100 | 170 | 140 | 100 | 25 |

[Note: Make the intervals continuous and take the mid-points as representative values, when estimating the mean value.]

**13.9.4.** The following contingency table gives the frequencies of people classified according to their mood and intelligence. Test, at a 1% level of rejection, to see whether there is any association between these two characteristics of classification, where $I$ = intelligent, $II$ = average, $III$ = below average.

| Intelligence → | $I$ | $II$ | $III$ |
|---|---|---|---|
| Mood ↓ | | | |
| Good | 15 | 10 | 10 |
| Tolerable | 8 | 10 | 10 |
| Intolerable | 8 | 10 | 15 |

**13.9.5.** The following table gives the telephone calls received by a switchboard on 265 days. Test whether or not a Poisson distribution with parameter $\lambda = 2$ is a good fit, by using the Kolmogorov–Smirnov test $D_n$. [The tabulated value of $D_{256} = 0.085$ at 5% level of rejection.]

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 52 | 68 | 60 | 40 | 22 | 10 | 3 | 1 | 0 |

**13.9.6.** The following table gives the waist measurements of 35 girls. Test the goodness-of-fit of a $N(\mu = 15, \sigma^2 = 4)$ to the data by using $D_n$ statistic. [The observed value of $D_{35} = 0.27$ at 0.01 level of rejection.]

| $x$ | 10 | 12 | 13 | 14 | 15 | 17 | 18 |
|---|---|---|---|---|---|---|---|
| Frequency | 2 | 4 | 6 | 10 | 7 | 4 | 2 |

**13.9.7.** The following are the observations on the grades obtained by 14 students in a particular course. Assume that the typical grade $x$ has a symmetric distribution around $E(x) = \mu$. Test the hypothesis at a 5% level of rejection that $\mu = 80$, against $\mu \neq 80$. Data: $90, 100, 60, 40, 80, 60, 50, 40, 55, 62, 80, 80, 30, 80$.

**13.9.8.** The following are the grades of a randomly selected 10 students under method 1 of teaching: $90, 95, 80, 80, 85, 70, 73, 82, 83, 80$ and the following are the grades of 8 randomly selected students under method 2 of teaching: $40, 50, 55, 60, 65, 45, 70, 100$. If the students have the same background test, the hypothesis that both the methods are equally effective, by using a $u$-test at a 5% level of rejection.

**13.9.9.** Use a Kruskal–Wallis' $H$-test to test at a 5% level of rejection that the following three populations, designated by the random variables $x_1$, $x_2$ and $x_3$, are identical, based on the following data: Data on

$$
\begin{aligned}
&x_1 : 5, 2, 5, 6, 8, 10, 12, 11, 10;\\
&x_2 : 15, 16, 2, 8, 10, 14, 15, 15, 18\\
&x_3 : 20, 18, 30, 15, 10, 11, 8, 15, 18, 12.
\end{aligned}
$$

**13.9.10.** Test the hypothesis of randomness of the occurrence of the symbol $D$ in a production process, based on the following observed sequence [test at a 1% level of rejection]:

$$GGDDDGDGGGDDGGGGDGGDDGGGGGDGGGGD.$$

# 14 Model building and regression

## 14.1 Introduction

There are various types of models that one can construct for a given set of data. The types of model that is chosen depends upon the type of data for which the model is to be constructed. If the data are coming from a deterministic situation, then there may be already an underlying mathematical formula such as a physical law. Perhaps the law may not be known yet. When the physical law is known, then there is no need to fit a model, but for verification purposes, one may substitute data points into the speculated physical law. For example, a simple physical law for gases says that the pressure $P$ multiplied by volume $V$ is a constant under a constant temperature. Then the physical law that is available is

$$PV = c$$

where $c$ is a constant. When it is a mathematical relationship, then all pairs of observations on $P$ and $V$ must satisfy the equation $PV = c$. If $V_1$ is one observation on $V$ and if the corresponding observation is $P_1$ for $P$, then $P_1V_1 = c$ for the same constant $c$. Similarly, other pairs of observations $(P_2, V_2), \ldots$ will satisfy the equation $PV = c$. If there are observational errors in observing $P$ and $V$, then the equation may not be satisfied exactly by a given observational pair. If the model proposed $PV = c$ is not true exactly, then of course the observational pairs $(P_i, V_i)$ for some specific $i$ may not satisfy the equation $PV = c$. There are many methods of handling deterministic situations. The usual tools are differential equations, difference equations, functional equations, integral equations, etc. For more details on deterministic situations and the corresponding model, see the CMS publication of 2010 SERC Notes.

## 14.2 Non-deterministic models

Deterministic situation is governed by definite mathematical rules. There is no chance variation involved. But most of the practical situations, mostly in social sciences, economics, commerce, management, etc. as well as many physical phenomena, are non-deterministic in nature. An earthquake at a place cannot be predetermined but with sophisticated prediction tools we may be able to predict the occurrence to some extent. We know that Meenachil River will be flooded during the monsoon season but we cannot tell in advance what the flood level will be on July 1, 2020, in front of Pastoral Institute. Even though many factors about weight gain are known, we cannot state for sure how much the weight gain will be on a cow under a certain feed. A student who is going to write an examination cannot tell beforehand what exact grade she is going to get. She may be able to predict that it will be above 80% from her knowledge

https://doi.org/10.1515/9783110562545-014

about the subject matter. But after writing the examination she might be able to give a better prediction that she would get at least 95% if not 100%. She could improve her prediction by knowing additional information of having written the examination.

Situations described above and many other situations of the same type are not deterministic in nature. When chance variation is involved, prediction can be made by using properties of random variables or chance variables or measurement of chance or probabilities.

Since a lot of practical situations are random or non-deterministic in nature, when we talk about model building, people naturally think that we are trying to describe a random or non-deterministic situation by mathematical modeling. Attempts to describe random situations have given birth to different branches of science. Stochastic process is one such area where we study a collection of random variables. Time series analysis is an area where we study a collection of random variables over time. Regression is an area where we try to describe random situations by analyzing conditional expectations. As can be seen, even to give a basic description of all the areas and disciplines where we build models, it will take hundreds of pages. Hence what we will do here is to pick a few selected topics and give a basic introduction to these topics.

### 14.2.1 Random walk model

Consider a simple practical situation of a drunkard coming out of the liquor bar, denoted by $S$ in Figure 14.1.

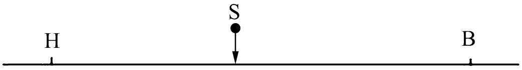


**Figure 14.1:** Random walk on a line.

He tries to walk home. Since he is fully drunk, assume that he walks in the following manner. At every minute, he will either take a step forward or backward. Let us assume a straight line path. Suppose that he covers 1 foot (about a third of a meter) at each step. He came out from the bar as indicated by the arrow. Then if his first step is to the right, then he is one foot closer to home, whereas if his first step is to the left, then he is farther way from home by one foot. At the next minute, he takes the next step either forward or backward. If he had taken the second step also to backward, then now he is farther away from home by two feet. One can associate a chance or probability for taking a step to the left (backward) or right (forward) at each stage. If the probabilities are $\frac{1}{2}$ each, then at each step there is 50% chance that the step will be forward and 50% chance that the step will be backward. If the probabilities of going forward and backward are 0.6 and 0.4, respectively, then there is a 60% chance that his first step will be forward.

Some interesting questions to ask in this case are the following: What is the chance that eventually he will reach home? What is the chance that eventually he will get lost and walk away from home to infinity? Where is his position after $n$ steps? There can be several types of modifications to the simple random walk on a line. In Figure 14.1, a point is marked as $B$. This $B$ may be a barrier. This barrier may be such that once he hits the barrier he falls down dead or the walk is finished, or the barrier may be such that if he hits the barrier there is a certain positive chance of bouncing back to the previous position so that the random walk can continue and there may be a certain chance that the walk is finished, and so on.

An example of a 2-dimensional random walk is the case of Mexican jumping beans. There is a certain variety of Mexican beans (lentils). If you place a dried bean on the tabletop, then after a few seconds it jumps by itself in a random direction to another point on the table. This is due to an insect making the dry bean as its home and the insect moves around by jumping. This is an example of a two-dimensional random walk. We can also consider random walk in space such as a dust particle moving around in the atmosphere and random walk in higher dimensions.

### 14.2.2 Branching process model

In nature, there are many species which behave in the following manner. There is a mother and the mother gives birth to $k$ offsprings once in her lifetime. After giving birth, the mother dies. The number of offsprings could be $0, 1, 2, \dots, k$ where $k$ is a finite number, not infinitely large. A typical example is the banana plant. One banana plant gives only one bunch of bananas. You cut the bunch and the mother plant is destroyed. The next generation offsprings are the new shoots coming from the bottom. The number of shoots could be $0, 1, 2, 3, 4, 5$, usually a maximum of 5 shoots. These shoots are the next generation plants. Each shoot, when mature, can produce one bunch of bananas each. Usually, after cutting the mother banana plant, the farmer will separate the shoots and plant all shoots, except one, elsewhere so that all have good chances of growing up into healthy banana plants.

Another example is the pineapple. Again one pineapple plant gives only one pineapple. The pineapple itself will have one shoot of plant at the top of the fruit and other shoots will be coming from the bottom. Again the mother plant is destroyed when the pineapple is plucked. Another example is certain species of spiders. The mother carries the sack of eggs around and dies after the new offsprings are born. Another example is salmon fish. From the wide ocean, the mother salmon enters into a fresh water river, goes to the birthplace of the river, overcoming all types of obstacles on the way, and lays one bunch of eggs and then promptly dies. Young salmon come out of these eggs and they find their way down river to the ocean. The life cycle is continued.

Assume that the last name of a person, for example, “Rumfeld” is carried only by the sons and not by the daughters. It is assumed that the daughters take the last names of their husbands. Then there is a chance that the father’s last name will be extinct after some generations. What is the chance that the name Rumfeld will disappear from the Earth?

These examples are examples for branching processes. Interesting questions to ask are the following: What is the chance that the species will be extinct eventually? This can happen if there is a positive probability of having no offspring in a given birth. What is the expected population size after $n$ generations? The branching process is a subject matter by itself and it is a special case of a general process known as birth and death processes.

### 14.2.3 Birth and death process model

This can be explained with a simple example. Suppose that there is a good pool area in a river, a good fishing spot for fishermen. Fish move in and move out of the pool area. If $N(t)$ is the number of fish in the pool area at time $t$ and if one fish moved out at the next unit of time, then $N(t+1) = N(t) - 1$. On the other hand, if one fish moved into the pool area at the next time unit then $N(t+1) = N(t) + 1$. When one addition is there, then one can say that there is one birth and when one deletion is there we can say that there is one death. Thus if we are modeling such a process where there is possibility of birth and death, then we call it a birth and death model.

### 14.2.4 Time series models

Suppose that we are monitoring the flood level at the Meenachil River at the Pastoral Institute. If $x(t)$ denotes the flood level on the $t$-th day, time $= t$ being measured in days, then at the zeroth day or starting of the observation period the flood level is $x(0)$, the next day the flood level is $x(1)$ and so on. We are observing a phenomenon namely flood level over time. In this example, time is measured in discrete time units. Details of various types of processes and details of the various techniques available for time series modeling of data are available in SERC School Notes of 2005–2010. Since these are available to the students, the material will not be elaborated here. But some more details will be given in the coming chapters.

## 14.3 Regression type models

The first prediction problem that we are going to handle is associated with a concept called regression and our models will be regression-type models. When Sreelakshmy was born, her doctor predicted by looking at the heights of parents and grandparents that she would be $5'5''$ at the age of 11. When she grew up, when she hit 11 years of age and her height was only $5'2''$. Thus the true or observed height was $5'2''$ against the predicted height of $5'5''$. Thus the prediction was not correct and the error in the prediction = observed minus predicted = $5'2'' - 5'5'' = -3''$. Thus the prediction was off by $3''$ in magnitude. [We could have also described error as predicted minus observed.] Of course, the guiding principle is that smaller the distance between the observed and predicted, better the prediction. Thus we will need to consider some measures of distance between the observed and predicted. When random variables are involved, some measures of distances between the real scalar random variable $x$ and a fixed point $a$ are the following:

$$E|x-a| = \text{mean deviation or expected difference between } x \text{ and } a \tag{i}$$

$$[E|x-a|^2]^{\frac{1}{2}} = \text{mean square deviation between } x \text{ and } a \tag{ii}$$

$$[E|x-a|^r]^{\frac{1}{r}}, \quad r=1,2,3,\ldots \tag{iii}$$

where $E$ denotes the expected value. Many such measures can be proposed. Since we are going to deal with only mean deviation and mean square deviations mainly, we will not look into other measures of distance between $x$ and $a$ here. For the sake of curiosity, let us see what should be $a$, if $a$ is an arbitrary constant, such that $E|x-a|$ is a minimum or what should be an arbitrary constant $b$ such that $[E|x-b|^2]^{\frac{1}{2}}$ is a minimum?

### 14.3.1 Minimization of distance measures

**Definition 14.1** (A measure of scatter). A measure of scatter in real scalar random variable $x$ from an arbitrary point $\alpha$ is given by $\sqrt{E(x-\alpha)^2}$.

What should be $\alpha$ such that this dispersion is the least? Note that minimization of $\sqrt{E(x-\alpha)^2}$ is equivalent to minimizing $E(x-\alpha)^2$. But

$$\begin{aligned} E(x-\alpha)^2 &= E(x-E(x)+E(x)-\alpha)^2 \quad \text{by adding and subtracting } E(x) \\ &= E(x-E(x))^2 + E(E(x)-\alpha)^2 + 2E(x-E(x))(E(x)-\alpha) \\ &= E(x-E(x))^2 + (E(x)-\alpha)^2 \end{aligned} \tag{14.1}$$

because the cross product term is zero due to the fact that $(E(x)-\alpha)$ is a constant and, therefore, the expected value applies on $(x-E(x))$, that is,

$$E(x-E(x)) = E(x) - E(E(x)) = E(x) - E(x) = 0$$

since $E(x)$ is a constant. In the above computations, we assumed that $E(x)$ is finite or it exists. In (14.1), the only quantity containing $\alpha$ is $[E(x) - \alpha]^2$ where both the quantities $E(x)$ and $\alpha$ are constants, and hence (14.1) attains its minimum when the non-negative quantity $[E(x) - \alpha]^2$ attains its minimum which is zero. Therefore,

$$[E(x) - \alpha]^2 = 0 \quad \Rightarrow \quad E(x) - \alpha = 0 \quad \Rightarrow \quad E(x) = \alpha.$$

Hence the minimum is attained when $\alpha = E(x)$. [The maximum value that $E(x - \alpha)^2$ can attain is $+\infty$ because $\alpha$ is arbitrary.] We will state this as a result.

**Result 14.1.** *For real scalar random variable $x$, for which $E(x)$ exists or a fixed finite quantity, and for an arbitrary real number $\alpha$*

$$\min_{\alpha} [E(x - \alpha)^2]^{\frac{1}{2}} \quad \Rightarrow \quad \min_{\alpha} E(x - \alpha)^2 \quad \Rightarrow \quad \alpha = E(x). \tag{14.2}$$

In a similar fashion, we can show that the mean deviation is least when the deviation is taken from the median. This can be stated as a result.

**Result 14.2.** *For a real scalar random variable $x$, having a finite median, and for an arbitrary real number $b$,*

$$\min_{b} E|x - b| \quad \Rightarrow \quad b = M$$

*where $M$ is the median of $x$.*

The median $M$ is the middle value for $x$ in the sense that

$$\Pr\{x \leq M\} \geq \frac{1}{2} \quad \text{and} \quad \Pr\{x \geq M\} \geq \frac{1}{2}$$

where $\Pr\{\cdot\}$ denotes the probability of the event $\{\cdot\}$. The median $M$ can be unique or there may be many points qualifying to be the median for a given $x$ depending upon the distribution of $x$.

### 14.3.2 Minimum mean square prediction

First, we will consider the problem of predicting one real scalar random variable by using one real scalar random variable. Such situations are plenty in nature. Let $y$ be the variable to be predicted and let $x$ be the variable by using which the variable $y$ is predicted. Some times we call the variable $y$ to be predicted as *dependent* variable and the variable $x$, which is independently preassigned to predict $y$, is called the independent variable. This terminology should not be confused with statistical independence of random variables. Let $y$ be the marks obtained by a student in a class test and $x$ be the amount of study time spent on that subject. The type of question that we would like to ask is the following: Is $y$ a function of $x$? If so, what is the functional relationship so that we can use it to evaluate $y$ at a given value of $x$. If there is no obvious functional relationship, can we use a preassigned value of $x$ to predict $y$? Can we answer a question such as if 20 hours of study time is used what will be the grade in the class test? In this case, irrespective of whether there exists a relationship between $x$ and $y$ or not, we would like to use $x$ to predict $y$.

As another example, let $y$ be the growth of a plant seedling, (measured in terms of height), in one week, against the amount of water $x$ supplied. As another example, let $y$ be the amount of evaporation of certain liquid in one hour and $x$ be the total exposed surface area.

If $x$ is a variable that we can preassign or observe, what is a prediction function of $x$ in order to predict $y$? Let $\phi(x)$ be an arbitrary function of $x$ that we want to use as a predictor of $y$. We may want to answer questions like: what is the predicted value of $y$ if the function $\phi(x)$ is used to predict $y$ at $x = x_0$ where $x_0$ is a given point. Note that infinitely many functions can be used as a predictor for $y$. Naturally, the predicted value of $y$ at $x = x_0$ will be far off from the true value if $\phi(x)$ is not a good predictor for $y$. What is the “best” predictor, “best” in some sense? If $y$ is predicted by using $\phi(x)$ then the ideal situation is that $\phi(x)$, at every given value of $x$, coincides with the corresponding observed value of $y$. This is the situation of a mathematical relationship, which may not be available in a problem in social sciences, physical sciences and natural sciences. For the example of the student studying for the examination if a specific function $\phi(x)$ is there, where $x$ is the number of hours of study, then when $x = 3$ hours $\phi(x)|_{x=3} = \phi(3)$ should produce the actual grade obtained by the student by spending 3 hours of study. Then $\phi(x)$ should give the correct observation for every given value of $x$. Naturally this does not happen. Then the error in using $\phi(x)$ to predict the value of $y$ at a given $x$ is

$$y - \phi(x) \quad \text{or we may take as} \quad \phi(x) - y.$$

The aim is to minimize a “distance” between $y$ and $\phi(x)$ and thus construct $\phi$. Then this $\phi$ will be a “good” $\phi$. This is the answer from common sense. We have many mathematical measures of “distance” between $y$ and $\phi(x)$ or measures of scatter in $e = y - \phi(x)$. One such measure is $\sqrt{E[y - \phi(x)]^2}$ and another measure is $E|y - \phi(x)|$, where $y$ is a real scalar random variable and $x$ is a preassigned value of another random variable or more precisely, for the first measure of scatter the distance is $\sqrt{E[y - \phi(x = x_0)]^2}$, that is, at $x = x_0$. Now, if we take this measure then the problem is to minimize over all possible functions $\phi$.

$$\min_{\phi} \sqrt{E[y - \phi(x = x_0)]^2} \quad \Rightarrow \quad \min_{\phi} E[y - \phi(x = x_0)]^2.$$

But from (14.2) it is clear that the “best” function $\phi$, best in the minimum mean square sense, that is, minimizing the expected value or mean value of the squared error, is

$\phi = E(y|x = x_0)$ or simply $\phi = E(y|x)$ = conditional expectation of $y$ given $x$. Hence this "best predictor", best in the minimum mean square sense, is defined as the regression of $y$ on $x$. Note that if we had taken any other measure of scatter in the error $e$ we would have ended up with some other function for the best $\phi$. Hereafter, when we say "regression", we will mean the best predictor, best in the minimum mean square sense or the conditional expectation of $y$ given $x$. Naturally, for computing $E(y|x)$ we should either have the joint distribution of $x$ and $y$ or at least the conditional distribution of $y$, given $x$.

**Definition 14.2** (Regression of $y$ on $x$)**.** It is defined as $E(y|x)$ = conditional expectation of $y$ given $x$ whenever it exists.

**Note 14.1.** There is quite a lot of misuse in this area of "regression". In some applied statistics books, the concept of *regression* is mixed up with the *least square estimates*. Hence the students must pay special attention to the basic concepts and logics of derivations here so that the topic is not mixed up with least square estimation problem. Regression has practically very little to do with least square estimation.

**Example 14.1.** If $x$ and $y$ have a joint distribution given by the following density function, where both $x$ and $y$ are normalized to the interval $[0,1]$,

$$f(x,y) = \begin{cases} x+y, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & \text{elsewhere,} \end{cases}$$

what is the "best" predictor of $y$ based on $x$, best in the minimum mean square sense? Also predict $y$ at (i) $x = \frac{1}{3}$, (ii) $x = 1.5$.

**Solution 14.1.** As per the criterion of "best", we are asked to compute the regression of $y$ on $x$ or $E(y|x)$. From elementary calculus, the marginal density of $x$, denoted by $f_1(x)$, is given by

$$f_1(x) = \int_y f(x,y)\mathrm{d}y = \int_0^1 (x+y)\mathrm{d}y = \begin{cases} x + \frac{1}{2}, & 0 \le x \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Then the conditional density of $y$ given $x$ is

$$g(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{x+y}{x+\frac{1}{2}}, \quad 0 \le y \le 1$$

and zero elsewhere. Hence the conditional expectation of $y$ given $x$ is then

$$E(y|x) = \int_{y|x} yg(y|x)\mathrm{d}y = \int_{y=0}^1 y\left[\frac{x+y}{x+\frac{1}{2}}\right]\mathrm{d}y$$

$$= \frac{1}{x+\frac{1}{2}} \int_{y=0}^{1} (xy + y^2) \mathrm{d}y = \frac{\frac{x}{2}+\frac{1}{3}}{x+\frac{1}{2}}.$$

Hence the best predictor of $y$ based on $x$ in Example 14.1 is

$$E(y|x) = \frac{\frac{x}{2}+\frac{1}{3}}{x+\frac{1}{2}}$$

for all given admissible values of $x$. This answers the first question. Now, to predict $y$ at a given $x$ we need to only substitute the value of $x$. Hence the predicted value at $x = \frac{1}{3}$ is

$$E\left(y|x = \frac{1}{3}\right) = \frac{\frac{1}{6}+\frac{1}{3}}{\frac{1}{3}+\frac{1}{2}} = \frac{3}{5}.$$

The predicted value of $y$ at $x = 1.5$ is not available from the above formula because 1.5 is not an admissible value of $x$ or it is outside the support $0 \le x \le 1$ of the density $g(y|x)$. The question contradicts with what is given as the density in Example 14.1.

**Example 14.2.** Suppose that it is found that $x$ and $y$ have a joint distribution given by the following: [Again we will use the same notation to avoid introducing too many symbols, even though $f$ in Example 14.1 is different from $f$ in Example 14.2.]

$$f(x,y) = \begin{cases} \mathrm{e}^{-x-y}, & 0 \le x < \infty,\ 0 \le y < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

Evaluate the "best" predictor of $y$ based on $x$, best in the minimum mean square sense, and predict $y$ at $x = x_0$.

**Solution 14.2.** We need the conditional expectation of $y$ given $x$, for which we need the conditional density of $y$ given $x$. From the above joint density, it is clear that the marginal density of $x$ is

$$f_1(x) = \begin{cases} \mathrm{e}^{-x}, & 0 \le x < \infty \\ 0, & \text{elsewhere} \end{cases}$$

and hence the conditional density of $y$ given $x$ is given by

$$g(y|x) = \frac{f(x,y)}{f_1(x)}$$

$$g(y|x) = \begin{cases} \mathrm{e}^{-y}, & 0 \le y < \infty \\ 0, & \text{elsewhere} \end{cases}$$

and the conditional expectation of $y$ given $x$ is

$$E(y|x) = \int_0^\infty y\mathrm{e}^{-y}\mathrm{d}y = 1$$

[evaluated by using a gamma function as $\Gamma(2) = 1!$ or by integration by parts]. Here, $E(y|x)$ is not a function of $x$, which means that whatever be the preassigned value of $x$ the predicted value of $y$ is simply 1. In other words, there is no meaning in using $x$ to predict the value of $y$ because the conditional distribution is free of the conditioned variable $x$. This happens because in this example, $x$ and $y$ are statistically independently distributed. Hence $x$ cannot be used to predict $y$ and vice versa.

But note that, in Examples 14.1 and 14.2 we have more information about the variables $x$ and $y$ than what is necessary to construct the "best" predictor. The best predictor is $E(y|x)$, and hence we need only the conditional distribution of $y$ given $x$ to predict $y$ and we do not need the joint distribution. Knowing the joint distribution means knowing the whole surface in a 3-dimensional space. Knowing the conditional distribution means knowing only the shape of the curve when this surface $f(x,y)$ is cut by the plane $x = x_0$ for some preassigned $x_0$. [The reader is asked to look at the geometry of the whole problem in order to understand the meaning of the above statement.] Thus we can restrict our attention to conditional distributions only for constructing a regression function.

**Example 14.3.** The strength $y$ of an iron rod is deteriorating depending upon the amount of rust $x$ on the rod. The more rust means less strength and finally rust will destroy the iron rod. The conditional distribution of $y$ given $x$ is seen to be of exponential decay model with the density

$$g(y|x) = \begin{cases} \frac{1}{1+x}\mathrm{e}^{-\frac{y}{1+x}}, & 0 \le y < \infty,\ 1+x > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

Construct the best predictor function for predicting the strength $y$ at the preassigned amount $x$ of rust and predict the strength when $x = 2$ units.

**Solution 14.3.** From the density given above, it is clear that at every $x$ the conditional density of $y$ given $x$ is exponential with expected value $1+x$ (comparing with a negative exponential density). Hence

$$E(y|x) = 1 + x$$

is the best predictor. The predicted value of $y$ at $x = 2$ is $1 + 2 = 3$ units.

**Example 14.4.** The marks $y$ obtained by the students in an examination is found to be normally distributed with polynomially increasing expected value with respect to the amount $x$ of time spent. The conditional distribution of $y$, given $x$, is found to be normal with the density

$$g(y|x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-70-2x-x^2)^2}, \quad -\infty < y < \infty, \ 0 \le x \le 4.$$

Construct the best predictor of $y$ based on $x$ and predict $y$ at $x = 3$ hours.

**Solution 14.4.** From the density itself, it is clear that the conditional density of $y$ given $x$ is $N(\mu, \sigma^2)$ with $\sigma^2 = 1$ and $\mu = 70 + 2x + x^2$ and, therefore,

$$E(y|x) = 70 + 2x + x^2$$

is the best predictor of $y$, and the predicted marks at $x = 3$ is

$$70 + 2(3) + 3^2 = 85.$$

In Examples 14.1 and 14.4, the regression function, that is, $E(y|x)$, is a non-linear function of $x$.

$$E(y|x) = \frac{\frac{x}{2} + \frac{1}{3}}{x + \frac{1}{2}} \quad \text{in Example 14.1}$$

$$E(y|x) = 70 + 2x + x^2 \quad \text{in Example 14.4}$$

whereas

$$E(y|x) = 1 + x \quad \text{in Example 14.3}$$

which is a linear function in $x$. Thus the regression of $y$ on $x$ may or may not be linear function in $x$. In Example 14.3, if $x$ and $y$ had a joint bivariate normal distribution, then we know for sure that $E(y|x)$ is a linear function in $x$. Thus one should not conclude that regression being a linear function in $x$ means that the variables are jointly normally distributed because it is already seen that in Example 14.3 the regression is linear in $x$ but it is not a case of joint normal distribution.

**Note 14.2.** The word “regression” means to go back, to regress means to go back. But in a regression-type prediction problem we are not going back to something. We are only computing the conditional expectation. The word “regression” is used for historical reasons. The original problem, when regression was introduced, was to infer something about ancestors by observing offsprings, and thus going back.

### 14.3.3 Regression on several variables

Again, let us examine the problem of predicting a real scalar random variable $y$ at preassigned values of many real scalar variables $x_1, x_2, \dots, x_k$. As examples, we can cite many situations.

**Example 14.5.**

$$\begin{aligned}
y &= \text{the marks obtained by a student in an examination}\\
x_1 &= \text{the amount of time spent on it}\\
x_2 &= \text{instructor's ability measured in the scale } 0 \le x_2 \le 10.\\
x_3 &= \text{instructor's knowledge in the subject matter}\\
x_4 &= \text{student's own background preparation in the area}
\end{aligned}$$

**Example 14.6.**

$$\begin{aligned}
y &= \text{cost of living}\\
x_1 &= \text{unit price for staple food}\\
x_2 &= \text{unit price for vegetables}\\
x_3 &= \text{cost of transportation}
\end{aligned}$$

and so on. There can be many factors contributing to $y$ in Example 14.5 as well as in Example 14.6. We are not sure in which form these variables $x_1, \dots, x_k$ will enter into the picture. If $\psi(x_1, \dots, x_k)$ is the prediction function for $y$, then predicting exactly as in the case of one variable situation the best prediction function, best in the sense of minimum mean square error, is again the conditional expectation of $y$ given $x_1, \dots, x_k$. That is, $E(y|x_1, \dots, x_k)$. Hence the regression of $y$ on $x_1, \dots, x_k$ is again defined as the conditional expectation.

**Definition 14.3.** The regression of $y$ on $x_1, \dots, x_k$ is defined as the conditional expectation of $y$ given $x_1, \dots, x_k$, that is, $E(y|x_1, \dots, x_k)$.

As before, depending upon the conditional distribution of $y$ given $x_1, \dots, x_k$ the regression function may be a linear function of $x_1, \dots, x_k$ or may not be a linear function.

**Example 14.7.** If $y, x_1, x_2$ have a joint density,

$$f(y, x_1, x_2, x_3) = \begin{cases} \frac{2}{3}(y + x_1 + x_2), & 0 \le y, x_1, x_2 \le 1\\ 0, & \text{elsewhere}\end{cases}$$

evaluate the regression of $y$ on $x_1$ and $x_2$.

**Solution 14.7.** The joint marginal density of $x_1$ and $x_2$ is given by integrating out $y$. Denoting it by $f_1(x_1, x_2)$, we have

$$\begin{aligned}
f_1(x_1, x_2) &= \int_{y=0}^{1} \frac{2}{3}(y + x_1 + x_2)\mathrm{d}y\\
&= \begin{cases} \frac{2}{3}(\frac{1}{2} + x_1 + x_2), & 0 \le x_1 \le 1,\ 0 \le x_2 \le 1\\ 0, & \text{elsewhere.}\end{cases}
\end{aligned}$$

Hence the conditional density of $y$ given $x_1$ and $x_2$ is

$$g(y|x_1,x_2) = \begin{cases} \frac{\frac{2}{3}(y+x_1+x_2)}{\frac{2}{3}(\frac{1}{2}+x_2+x_3)} = \frac{y+x_1+x_2}{\frac{1}{2}+x_1+x_2}, & 0 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Therefore, the regression of $y$ on $x_1$, $x_2$ is given by

$$E(y|x_1,x_2) = \int_{y=0}^{1} y\left[\frac{y+x_1+x_2}{\frac{1}{2}+x_1+x_2}\right]\mathrm{d}y = \frac{\frac{1}{3}+\frac{x_1}{2}+\frac{x_2}{2}}{\frac{1}{2}+x_1+x_2}.$$

This is the best predictor of $y$. For example, the predicted value of $y$ at $x_1 = 0$, $x_2 = \frac{1}{2}$ is given by

$$\frac{\frac{1}{3}+0+\frac{1}{4}}{\frac{1}{2}+0+\frac{1}{2}} = \frac{7}{12}.$$

In this example, how can we predict $y$ based on $x_2$ alone, that is, $E(y|x_2)$? First, integrate out $x_1$ and obtain the joint marginal density of $y$ and $x_2$. Then proceed as in the one variable case. Note that $E(y|x_2)$ is not the same as $E(y|x_2, x_1 = 0)$. These two are two different statements and two different items.

Again, note that for constructing the regression function of $y$ on $x_1,\dots,x_k$, that is, $E(y|x_1,\dots,x_k)$, we need only the conditional distribution of $y$ given $x_1,\dots,x_k$ and we do not require the joint distribution of $y, x_1,\dots,x_k$.

**Example 14.8.** If the conditional density of $y$ given $x_1,\dots,x_k$ is given by

$$g(y|x_1,\dots,x_k) = \begin{cases} \frac{1}{5+x_1+\dots+x_k}\mathrm{e}^{-\frac{y}{5+x_1+\dots+x_k}}, & 0 \le y < \infty,\ 5+x_1+\dots+x_k > 0 \\ 0, & \text{elsewhere} \end{cases}$$

evaluate the regression function.

**Solution 14.8.** The conditional density is the exponential density with the mean value

$$E(y|x_1,\dots,x_k) = 5 + x_1 + \dots + x_k$$

and this is the regression of $y$ on $x_1,\dots,x_k$, which here is a linear function in $x_1,\dots,x_k$ also.

When the variables are all jointly normally distributed also, one can obtain the regression function to be linear in the regressed variables. Example 14.8 illustrates that joint normality is not needed for the regression function to be linear. When the regression function is linear, we have some interesting properties.

## Exercises 14.2–14.3

**14.3.1.** Prove that $\min_a E|y-a| \Rightarrow a =$ median of $y$, where $a$ is an arbitrary constant. Prove the result for continuous, discrete and mixed cases for $y$.

**14.3.2.** Let

$$f(x,y) = \frac{c}{(y+x)^3}, \quad 1 \le y < \infty,\ 0 \le x \le 1,$$

and zero elsewhere. If $f(x,y)$ is a joint density function of $x$ and $y$, then evaluate (i) the normalizing constant $c$; (ii) the marginal density of $y$; (iii) the marginal density of $x$; (iv) the conditional density of $y$ given $x$.

**14.3.3.** By using the joint density in Exercise 14.3.2, evaluate the regression of $y$ on $x$ and then predict $y$ at $x = \frac{1}{2}$.

**14.3.4.** Consider the function

$$f(x,y) = cy^{x^2-1}, \quad 0 \le y \le 1,\ 1 \le x \le 2,$$

and zero elsewhere. If this is a joint density function, then evaluate (i) the normalizing constant $c$; (ii) the marginal density of $x$; (iii) the conditional density of $y$ given $x$; (iv) the regression of $y$ given $x$; (v) the predicted value of $y$ at $x = \frac{1}{3}$.

**14.3.5.** Let

$$f(x_1,x_2,x_3) = c(1+x_1+x_2+x_3), \quad 0 \le x_i \le 1,\ i = 1,2,3$$

and zero elsewhere be a density function. Then evaluate the following: (i) the normalizing constant $c$; (ii) the regression of $x_1$ on $x_2$, $x_3$; (iii) the predicted value of $x_1$ at $x_2 = \frac{1}{2}$, $x_3 = \frac{1}{4}$; (iv) the predicted value of $x_1$ at $x_2 = \frac{1}{3}$.

## 14.4 Linear regression

Let $y, x_1, \dots, x_k$ be real scalar random variables and let the regression of $y$ on $x_1, \dots, x_k$ be a linear function in $x_1, \dots, x_k$, that is,

$$E(y|x_1,\dots,x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{14.3}$$

where $\beta_0, \beta_1, \dots, \beta_k$ are constants. If joint moments up to the second order exist, then we can evaluate the constants $\beta_0, \beta_1, \dots, \beta_k$ in terms of product moments. In order to achieve this, we will need two results from elementary statistics. These will be listed here as lemmas. These were given earlier but for ready reference these will be listed here again.

**Lemma 14.1.**

$$E(u) = E_v[E(u|v)]$$

*whenever the expected values exist. Here, $E(u|v)$ is in the conditional space of $u$ given $v$ or computed from the conditional distribution of $u$ given $v$, as a function of $v$. Then the resulting quantity is treated as a function of the random variable $v$ in the next step of taking the expected value $E_v(\cdot)$.*

**Lemma 14.2.**

$$\mathrm{Var}(u) = \mathrm{Var}[E(u|v)] + E[\mathrm{Var}(u|v)]$$

*That is, the sum of the variance of the conditional expectation and the expected value of the conditional variance is the unconditional variance of any random variable $u$, as long as the variances exist.*

All the expected values and variances defined there must exist for the results to hold. The proofs follow from the definitions themselves and are left to the students. Let us look into the implications of these two lemmas with the help of some examples.

**Example 14.9.** Consider the joint density of $x$ and $y$, given by

$$f(x,y) = \begin{cases} \frac{1}{x^2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-2-x)^2}, & -\infty < y < \infty,\ 1 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

Evaluate the regression of $y$ on $x$, and also verify Lemma 14.1.

**Solution 14.9.** Integrating out $y$, one has the marginal density of $x$. Integration with respect to $y$ can be effected by looking at a normal density in $y$ with expected value $2+x$. Then the marginal density of $x$ is given by

$$f(x) = \begin{cases} \frac{1}{x^2}, & 1 \le x < \infty \\ 0, & \text{elsewhere,} \end{cases}$$

because the joint density is the product of conditional and marginal densities. Therefore, the conditional density of $y$ given $x$ is normal with expected value $2+x$ and hence the regression of $y$ on $x$ is given by

$$E(y|x) = 2 + x \tag{14.4}$$

which is a linear function in $x$ and well behaved smooth function of $x$. Expected value of the right side of (14.2) is then

$$E(2+x) = 2 + E(x)$$

$$
\begin{aligned}
&= 2 + \int_1^\infty \frac{x}{x^2}\mathrm{d}x \\
&= 2 + [\ln x]_1^\infty = \infty.
\end{aligned}
$$

Thus the expected value does not exist and Lemma 14.1 is not applicable here.

Note that in Lemma 14.1 the variable $v$ could be a single real scalar variable or a collection of real scalar variables. But since Lemma 14.2 is specific about variance of a single variable, the formula does not work if $v$ contains many variables.

**Example 14.10.** Verify Lemma 14.1 for the following joint density:

$$
f(x,y) = \begin{cases} 2, & 0 \le x \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}
$$

**Solution 14.10.** Here, the surface $z = f(x,y)$ is a prism sitting on the $(x,y)$-plane the non-zero part of the density is in the triangle $0 \le x \le y \le 1$. Thus the region can be defined as either $0 \le x \le y$ and $0 \le y \le 1$ or $x \le y \le 1$ and $0 \le x \le 1$. Marginally, $0 \le x \le 1$ as well as $0 \le y \le 1$. The marginal densities of $x$ and $y$ are respectively

$$
\begin{aligned}
f_1(x) &= \int_{y=x}^1 2\mathrm{d}y = \begin{cases} 2(1-x), & 0 \le x \le 1 \\ 0, & \text{elsewhere;} \end{cases} \\
f_2(y) &= \int_{x=0}^y 2\mathrm{d}x = \begin{cases} 2y, & 0 \le y \le 1 \\ 0, & \text{elsewhere.} \end{cases}
\end{aligned}
$$

Hence

$$
E(y) = \int_0^1 y(2y)\mathrm{d}y = \frac{2}{3} \quad \text{and} \quad E(x) = \int_0^1 x[2(1-x)]\mathrm{d}x = \frac{1}{3}.
$$

The conditional density of $y$ given $x$ is given by

$$
g(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad x \le y \le 1
$$

and zero elsewhere. Note that when $x$ is fixed at some point then $y$ can only vary from that point to 1. Therefore, the conditional expectation of $y$, given $x$,

$$
E(y|x) = \int_{y=x}^1 \frac{y}{1-x}\mathrm{d}y = \frac{1-x^2}{2(1-x)} = \frac{1+x}{2}.
$$

From this, by taking expected value we have

$$
E_x[E(y|x)] = \frac{1}{2}[E(1+x)] = \frac{1}{2}[1+E(x)] = \frac{1}{2}\left[1+\frac{1}{3}\right] = \frac{2}{3} = E(y).
$$

Thus the result is verified. [Note that when we take $E_x(\cdot)$ we replace the preassigned $x$ by the random variable $x$ or we consider all values taken by $x$ and the corresponding density or we switch back to the density of $x$ and we are no longer in the conditional density.]

Coming back to the linear regression in (14.3), we have

$$E(y|x_1,\dots,x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{14.5}$$

Taking expected value on both sides, which means expected value in the joint marginal density of $x_1,\dots,x_k$ and by Lemma 14.1,

$$\begin{aligned} E[E(y|x_1,\dots,x_k)] &= E(y) \\ E[\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k] &= \beta_0 + \beta_1 E(x_1) + \cdots + \beta_k E(x_k). \end{aligned} \tag{14.6}$$

Note that taking expected value of $x_j$ in the joint distribution of $x_1,\dots,x_k$ is equivalent to taking the expected value of $x_j$ in the marginal distribution of $x_j$ alone because the other variables can be integrated out (or summed up, if discrete) first to obtain the marginal density of $x_j$ alone. [The student may work out an example to grasp this point.] From (14.5) and (14.6), one has

$$E(y|x_1,\dots,x_k) - E(y) = \beta_1[x_1 - E(x_1)] + \cdots + \beta_k[x_k - E(x_k)]. \tag{14.7}$$

Multiply both sides of (14.7) by $x_j - E(x_j)$ for a specific $j$ and then take expected value with respect to $x_1,\dots,x_k$. The right side gives the following:

$$\begin{aligned} &\beta_1 \operatorname{Cov}(x_1,x_j) + \beta_2 \operatorname{Cov}(x_2,x_j) + \cdots \\ &\qquad + \beta_j \operatorname{Var}(x_j) + \cdots + \beta_k \operatorname{Cov}(x_k,x_j) \end{aligned} \tag{14.8}$$

because

$$E\{[x_j - E(x_j)][x_r - E(x_r)]\} = \begin{cases} \operatorname{Cov}(x_j,x_r), & \text{if } j \neq r \\ \operatorname{Var}(x_j), & \text{if } j = r. \end{cases}$$

The left side of (14.7) leads to the following:

$$E\{[x_j - E(x_j)]E(y)\} = E(y)\{E[x_j - E(x_j)]\} = 0$$

since for any variable $x_j$, $E[x_j - E(x_j)] = 0$ as long as the expected value exists.

$$\begin{aligned} E\{[x_j - E(x_j)]E(y|x_1,\dots,x_k)\} &= E\{E(y(x_j - E(x_j))|x_1,\dots,x_k\} \\ &= E[y(x_j - E(x_j))] \end{aligned}$$

since in the conditional expectation $x_1,\dots,x_k$ are fixed and, therefore, one can take $x_j - E(x_j)$, being constant, inside the conditional expected value and write $y(x_j - E(x_j))$ given $x_1,\dots,x_k$. But

$$E[y(x_j - E(x_j))] = \mathrm{Cov}(y,x_j)$$

because for any two real scalar random variables $x$ and $y$,

$$\begin{aligned}\mathrm{Cov}(x,y) &= E\{(x-E(x))(y-E(y))\}\\ &= E\{x[y-E(y)]\} = E\{y[x-E(x)]\}\end{aligned}$$

because

$$E\{E(x)[y-E(y)]\} = E(x)E\{y-E(y)\} = 0$$

since $E[y-E(y)] = E(y) - E(y) = 0$ and similarly $E\{E(y)[x-E(x)]\} = 0$. Therefore, we have

$$\sigma_{jy} = \beta_1\sigma_{1j} + \beta_2\sigma_{2j} + \cdots + \beta_k\sigma_{kj} \tag{14.9}$$

where $\sigma_{ij} = \mathrm{Cov}(x_i,x_j)$ and $\sigma_{jy} = \sigma_{yj} = \mathrm{Cov}(x_j,y)$. Writing (14.9) explicitly, one has

$$\begin{aligned}\sigma_{1y} &= \beta_1\sigma_{11} + \beta_2\sigma_{12} + \cdots + \beta_k\sigma_{1k}\\ \sigma_{2y} &= \beta_1\sigma_{21} + \beta_2\sigma_{22} + \cdots + \beta_k\sigma_{2k}\\ \vdots\ &\ \vdots\\ \sigma_{ky} &= \beta_1\sigma_{k1} + \beta_2\sigma_{k2} + \cdots + \beta_k\sigma_{kk}.\end{aligned}$$

Writing in matrix notation, we have

$$\begin{aligned}&\Sigma_y = \Sigma\beta, \quad \Sigma = (\sigma_{ij})\\ &\beta = \begin{bmatrix}\beta_1\\ \vdots\\ \beta_k\end{bmatrix}, \quad \Sigma_y = \begin{bmatrix}\sigma_{1y}\\ \vdots\\ \sigma_{ky}\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\sigma_{11} & \sigma_{12} & \dots & \sigma_{1k}\\ \vdots & \vdots & \dots & \vdots\\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk}\end{bmatrix}.\end{aligned}$$

Note that the covariance matrix $\Sigma$ is symmetric since $\sigma_{ij} = \sigma_{ji}$ for all $i$ and $j$. If $\Sigma$ is nonsingular, then

$$\beta = \Sigma^{-1}\Sigma_y \tag{14.10}$$

where $\Sigma^{-1}$ denotes the regular inverse of the covariance matrix or the variance-covariance matrix $\Sigma$. This notation $\Sigma$ is another awkward symbol in statistics. This can be easily confused with the summation symbol $\sum$. But since it is very widely used, we will also use it here. Is $\Sigma$ likely to be non-singular? If $\Sigma$ is singular, then it means that at least one of the rows (columns) of $\Sigma$ is a linear function of other rows (columns). This can happen if at least one of the variables $x_1, \dots, x_k$ is a linear function of the other variables. Since these variables are preassigned, and hence nobody will preassign one vector $(x_1, \dots, x_k)$ and another point as a constant multiple $\alpha(x_1, \dots, x_k)$ because the second point does not give any more information. Thus when the points are preassigned as in a regression problem, one can assume, without loss of generality, that $\Sigma$ is non-singular. But when $\Sigma$ is estimated, since observations are taken on the variables, near singularity may occur. We will look into this aspect later. From (14.10) and (14.6), one has

$$\begin{aligned} \beta_0 &= E(y) - \beta_1 E(x_1) - \cdots - \beta_k E(x_k) \\ &= E(y) - \beta' E(X) \end{aligned} \tag{14.11}$$

where a prime denotes the transpose

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad E(X) = \begin{bmatrix} E(x_1) \\ \vdots \\ E(x_k) \end{bmatrix}.$$

From (14.10), we have for example, $\beta_1 = \Sigma^{(1)}\Sigma_y$ where $\Sigma^{(1)}$ is the first row of $\Sigma^{-1}$, $\beta_j = \Sigma^{(j)}\Sigma_y$ where $\Sigma^{(j)}$ is the $j$-th row of $\Sigma^{-1}$ for $j = 1, \dots, k$.

Instead of denoting the variables as $y$ and $x_1, \dots, x_k$ we may denote the variables simply as $x_1, \dots, x_k$ and the problem is to predict $x_1$ by preassigning $x_2, \dots, x_k$. In this notation, we can write the various quantities in terms of the sub-matrices of $\Sigma$. For this purpose, let us write

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

$$\Sigma_{21} = \Sigma_{12}' = \begin{bmatrix} \sigma_{21} \\ \vdots \\ \sigma_{k1} \end{bmatrix}, \quad \Sigma_{22} = \begin{bmatrix} \sigma_{22} & \dots & \sigma_{2k} \\ \sigma_{32} & \dots & \sigma_{3k} \\ \vdots & \dots & \vdots \\ \sigma_{k2} & \dots & \sigma_{kk} \end{bmatrix}. \tag{14.12}$$

Then the best predictor, best in the minimum mean square sense, for predicting $x_1$ at preassigned values of $x_2, \dots, x_k$ is given by $E(x_1|x_2, \dots, x_k)$ and if this regression of $x_1$ on $x_2, \dots, x_k$ is linear in $x_2, \dots, x_k$ then it is of the form

$$E(x_1|x_2, \dots, x_k) = \alpha_0 + \alpha_2 x_2 + \cdots + \alpha_k x_k \tag{14.13}$$

where $\alpha_0, \alpha_2, \dots, \alpha_k$ are constants. Then from (14.10),

$$\alpha = \begin{bmatrix} \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} = \Sigma_{22}^{-1}\Sigma_{21} \tag{14.14}$$

and the regression of $x_1$ on $x_2, \dots, x_k$ or the best predictor of $x_1$ at preassigned values of $x_2, \dots, x_k$, when the regression is linear, is given by

$$E(x_1|x_2, \dots, x_k) = \alpha_0 + \alpha' X_2, \quad \alpha = \begin{bmatrix} \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_2 \\ \vdots \\ x_k \end{bmatrix}. \tag{14.15}$$

For example, when $k = 2$

$$\begin{aligned} E(x_1|x_2) &= E(x_1) + \frac{\operatorname{Cov}(x_1, x_2)}{\operatorname{Var}(x_2)}(x_2 - E(x_2)) \\ &= \mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2) \\ &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) \end{aligned} \tag{14.16}$$

where $E(x_1) = \mu_1$, $E(x_2) = \mu_2$, $\operatorname{Var}(x_1) = \sigma_1^2$, $\operatorname{Var}(x_2) = \sigma_2^2$ and $\rho$ is the correlation between $x_1$ and $x_2$. This is a very useful result when we consider a linear regression of one real scalar variable on another real scalar variable.

Let us compute the correlation between $x_1$ and its best linear predictor, that is, between $x_1$ and the predicting function in (14.15).

**Example 14.11.** If $X' = (x_1, x_2, x_3)$ has the mean value,

$$E(X') = (E(x_1), E(x_2), E(x_3)) = (2, 1, -1)$$

and the covariance matrix

$$\operatorname{Cov}(X) = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

construct the regression function for predicting $x_1$ at given values of $x_2$ and $x_3$, if it is known that the regression is linear in $x_2, x_3$.

**Solution 14.11.** As per our notation,

$$\Sigma_{11} = \sigma_{11} = 2, \quad \Sigma_{12} = (-1, 0), \quad \Sigma_{21} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \Sigma_{22} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

and hence

$$\Sigma_{22}^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}, \quad \Sigma_{12}\Sigma_{22}^{-1} = (-1, 0)\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} = (-1, 1).$$

Hence the best predictor is

$$E(x_1|x_2, x_3) = E(x_1) + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - E(X_2))$$

$$
\begin{aligned}
&= 2 + (-1, 1) \begin{bmatrix} x_2 - 1 \\ x_3 + 1 \end{bmatrix} \\
&= 2 - x_2 + 1 + x_3 + 1 = 4 - x_2 + x_3
\end{aligned}
$$

is the best prediction function for predicting $x_1$ at preassigned values of $x_2$ and $x_3$.

### 14.4.1 Correlation between $x_1$ and its best linear predictor

In order to compute the correlation between $x_1$ and $E(x_1|x_2, \dots, x_k)$, we need the variances of these two quantities and the covariance between them. As per our notation in (14.12), we have

$$\operatorname{Var}(x_1) = \sigma_{11}. \tag{14.17}$$

The variance of $\alpha_0 + \alpha' X$ can be computed by using the result on variance of a linear function of scalar variables. These will be stated as lemmas. These follow directly from the definition itself.

**Lemma 14.3.** *Consider a linear function of real scalar random variables $y_1, \dots, y_n$ with covariances, $\operatorname{Cov}(y_i, y_j) = v_{ij}$, $i, j = 1, \dots, n$ thereby $v_{ii} = \operatorname{Var}(y_i)$ and let the variance-covariance matrix in $(y_1, \dots, y_n)$ be denoted by $V = (v_{ij})$. Let*

$$
\begin{aligned}
u &= a_0 + a_1 y_1 + a_2 y_2 + \dots + a_n y_n = a_0 + a' Y \\
v &= b_0 + b_1 y_1 + b_2 y_2 + \dots + b_n y_n = b_0 + b' Y
\end{aligned}
$$

*where*

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

*and a prime denotes the transpose. Then*

$$\operatorname{Var}(u) = a' V a, \quad \operatorname{Var}(v) = b' V b, \quad \operatorname{Cov}(u, v) = a' V b = b' V a.$$

Note that $V$ is a symmetric matrix. Then with the help of Lemma 14.3, we have

$$
\begin{aligned}
\operatorname{Var}[E(x_1|x_2, \dots, x_k)] &= \operatorname{Var}[\alpha_0 + \alpha' X] = \operatorname{Var}[\alpha' X] \\
&= \alpha' \operatorname{Cov}(X) \alpha = \alpha' \Sigma_{22} \alpha
\end{aligned} \tag{14.18}
$$

where $\operatorname{Cov}(X)$ means the covariance matrix in $X$. But from (14.14) and (14.18), the variance of the best linear predictor

$$\alpha' \Sigma_{22} \alpha = [\Sigma_{22}^{-1} \Sigma_{21}]' \Sigma_{22} [\Sigma_{22}^{-1} \Sigma_{21}] = \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \tag{14.19}$$

because $\Sigma_{22}' = \Sigma_{22}$, $\Sigma_{21}' = \Sigma_{12}$. The covariance between $x_1$ and its best linear predictor is then

$$\begin{aligned}\operatorname{Cov}[x_1, E(x_1|x_2,\dots,x_k)] &= \operatorname{Cov}[x_1, \alpha_0 + \alpha' X] \\ &= \operatorname{Cov}[x_1, \alpha' X] = \alpha' \operatorname{Cov}(x_1, X) = \alpha' \Sigma_{21} \\ &= [\Sigma_{22}^{-1}\Sigma_{21}]'\Sigma_{21} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.\end{aligned}$$

Strangely enough, the covariance between $x_1$ and its best linear predictor is the same as the variance of the best linear predictor. Variance being non-negative, it is clear that the covariance in this case is also non-negative, and hence the correlation is also non-negative. Denoting the correlation by $\rho_{1.(2\dots k),}$ we have

$$\rho_{1.(2\dots k)}^2 = \frac{(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^2}{\sigma_{11}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}. \tag{14.20}$$

But note that the best predictor $E(x_1|x_2,\dots,x_k)$ for predicting $x_1$ at preassigned values of $x_2,\dots,x_k$ need not be linear. We have cited several examples where the regression function is non-linear. But if the regression is linear then the correlation between $x_1$ and its best linear predictor has the nice from given in (14.20).

## Exercises 14.4

**14.4.1.** Write the following linear functions by using vector, matrix notations. For example, $4 + x_1 - x_2 + 5x_3 = 4 + a'X = b'Y$ where the prime denotes the transpose and

$$a = \begin{bmatrix} 1 \\ -1 \\ 5 \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad b = \begin{bmatrix} 4 \\ 1 \\ -1 \\ 5 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

(i) $y = 2 + x_1 + x_2 - x_3$; (ii) $y = 1 + 2x_1 - x_2$; (iii) $y = 5 + x_1 + x_2 - 2x_3 + x_4$.

**14.4.2.** Write down the following quadratic forms in the form $X'AX$ where $A = A'$:
(i) $x_1^2 + 2x_2^2 - 3x_1x_2$;
(ii) $2x_1^2 + x_2^2 - x_3^2 + 2x_1x_2 - x_2x_3$;
(iii) $x_1^2 + x_2^2 + \dots + x_k^2$.

**14.4.3.** Write the same quantities in Exercise 14.4.2 as $X'AX$ where $A \neq A'$.

**14.4.4.** Can the following matrices represent covariance matrices, if so prove and if not explain why?

$$A_1 = \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & -3 \\ -3 & 2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 4 \end{bmatrix}, \quad A_5 = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 2 \\ 0 & 2 & 2 \end{bmatrix}, \quad A_6 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & -3 & 1 \\ 0 & 1 & 4 \end{bmatrix}.$$

**14.4.5.** Let $X' = (x_1, x_2, x_3)$ have a joint distribution with the following mean value and covariance matrix:

$$E(X) = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, \quad \mathrm{Cov}(X) = V = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 2 \end{bmatrix}.$$

Let the regression of $x_1$ on $x_2$ and $x_3$ be a linear function of $x_2$ and $x_3$. (i) Construct the best linear predictor $E(x_1|x_2,x_3)$ for predicting $x_1$ at preassigned values of $x_2$ and $x_3$; (ii) predict $x_1$ at $x_2 = 1$, $x_3 = 0$; (iii) compute the variance of the best predictor $E(x_1|x_2,x_3)$; (iv) compute the covariance between $x_1$ and the best linear predictor; (v) compute the correlation between $x_1$ and its best linear predictor.

## 14.5 Multiple correlation coefficient $\rho_{1.(2\ldots k)}$

The multiple correlation coefficient is simply defined as

$$\rho_{1.(2\ldots k)} = \sqrt{\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}} \tag{14.21}$$

with the notations as given in (14.12). It does not mean that we are assuming that there is a linear regression. If the regression is linear, then the multiple correlation coefficient is also the correlation between $x_1$ and its best linear predictor. The expression in (14.20) itself has many interesting properties and the multiple correlation coefficient, as defined in (14.21), has many statistical properties.

**Example 14.12.** For the same covariance matrix in Example 14.11, compute $\rho^2_{1.(2.3)}$.

**Solution 14.12.** We need to compute $\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}$, out of which $\Sigma_{12}\Sigma_{22}^{-1}$ is already computed in Example 14.11 as $\Sigma_{12}\Sigma_{22}^{-1} = (-1, 1)$, and $\Sigma_{21} = \binom{-1}{0}$. Therefore,

$$\rho^2_{1.(2.3)} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}} = \frac{1}{2}(-1, 1)\begin{pmatrix} -1 \\ 0 \end{pmatrix} = \frac{1}{2}.$$

### 14.5.1 Some properties of the multiple correlation coefficient

Let $b_2x_2 + \cdots + b_kx_k = b'X_2$ where $b' = (b_2, \ldots, b_k)$ and $X_2' = (x_2, \ldots, x_k)$ be an arbitrary linear predictor of $x_1$. That is, $x_1$ is predicted by using $b'X_2$. Let us compute the correlation between $x_1$ and this arbitrary predictor $b'X_2$. Note that, from Lemma 14.3 we have

$$\mathrm{Var}(b'X_2) = b'\Sigma_{22}b \quad \text{and} \quad \mathrm{Cov}(x_1, b'X_2) = b'\Sigma_{21}.$$

Then the square of the correlation between $x_1$ and an arbitrary linear predictor, denoted by $\eta^2$, is given by the following:

$$\eta^2 = \frac{(b'\Sigma_{21})^2}{(b'\Sigma_{22}b)\sigma_{11}}.$$

But from Cauchy–Schwarz inequality we have

$$(b'\Sigma_{21})^2 = [(b'\Sigma_{22}^{\frac{1}{2}})(\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21})]^2 \le [b'\Sigma_{22}b][\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]$$

where $\Sigma_{22}^{\frac{1}{2}}$ is the symmetric positive definite square root of $\Sigma_{22}$. Hence

$$\eta^2 = \frac{(b'\Sigma_{21})^2}{(b'\Sigma_{22}b)\sigma_{11}} \le \frac{[b'\Sigma_{22}b][\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]}{[b'\Sigma_{22}b]\sigma_{11}} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}. \tag{14.22}$$

In other words, the maximum value of $\eta^2$ is $\rho^2_{1.(2\dots k)}$ the square of the multiple correlation coefficient given in (14.20).

**Result 14.3.** *Multiple correlation coefficient of $x_1$ on $(x_2, \dots, x_k)$, where $x_1, x_2, \dots, x_k$ are all real scalar random variables, is also the maximum correlation between $x_1$ and an arbitrary linear predictor of $x_1$ based on $x_2, \dots, x_k$.*

**Note 14.3** (Cauchy–Schwarz inequality). Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be two sequences of real numbers. Then

$$\sum_{i=1}^{n} a_i b_i \le \left[\sum_{i=1}^{n} a_i^2\right]^{\frac{1}{2}} \left[\sum_{i=1}^{n} b_i^2\right]^{\frac{1}{2}} \tag{N14.1}$$

and the equality holds when $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ are linearly related. In terms of real scalar random variables $x$ and $y$, this inequality is the following:

$$|\mathrm{Cov}(x, y)| \le [\mathrm{Var}(x)]^{\frac{1}{2}} [\mathrm{Var}(y)]^{\frac{1}{2}} \tag{N14.2}$$

and the equality holds when $x$ and $y$ are linearly related.

Proof is quite simple. (N14.1) is nothing but the statement $|\cos\theta| \le 1$ where $\theta$ is the angle between the vectors $\vec{a} = (a_1, \dots, a_k)$ and $\vec{b} = (b_1, \dots, b_k)$ and (N14.2) is the statement that $|\rho| \le 1$ where $\rho$ is the correlation coefficient between the real scalar variables $x$ and $y$. Thus the covariance as well as correlation between the real scalar random variables $x$ and $y$ can be described as measuring $\cos\theta$ where $\theta$ is measuring angular dispersion between $x$ and $y$ or scatter in the point $(x, y)$. There are various variations and extensions of Cauchy–Schwarz inequality. But what we need to use in our discussions are available from (N14.1) and (N14.2).

**Note 14.4** (Determinants and inverses of partitioned matrices). Consider a matrix $A$ and its regular inverse $A^{-1}$, when $A$ is non-singular and let $A$ be partitioned as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix} \tag{N14.3}$$

where $A_{11}, A_{12}, A_{21}, A_{22}$ are submatrices in $A$ and $A^{11}, A^{12}, A^{21}, A^{22}$ are submatrices in $A^{-1}$. For example, let

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 1 & -1 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where let

$$A_{11} = [2], \quad A_{12} = [0,1], \quad A_{21} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad A_{22} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Then from elementary theory of matrices and determinants we have the following, denoting the determinant of $A$ by $|A|$:

$$|A| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}| \quad \text{if } |A_{11}| \neq 0 \tag{N14.4}$$

$$= |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}| \quad \text{if } |A_{22}| \neq 0. \tag{N14.5}$$

For our illustrative example, the determinant of $A_{11}$ is $|A_{11}| = 2$ and

$$\begin{aligned} |A_{22} - A_{21}A_{11}^{-1}A_{12}| &= \left| \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} - \begin{bmatrix} 0 \\ 2 \end{bmatrix} \frac{1}{2}[0,1] \right| \\ &= \left| \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right| \\ &= \begin{vmatrix} 1 & 1 \\ 1 & -2 \end{vmatrix} = -3 \end{aligned}$$

and, therefore,

$$|A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}| = (2)(-3) = -6.$$

The student may evaluate the determinant of $A$ directly and verify the result and as well as use (N14.5) and verify that result also. The proof for establishing (N14.4) and (N14.5) are quite simple. From the axioms defining a determinant, it follows that if linear functions of one or more rows are added to one or more rows the value of the determinant remains the same. This operation can be done one at a time or several steps together. Consider (N14.3). What is a suitable linear function of the rows containing $A_{11}$ to be added to the remaining rows containing $A_{21}$ so that a null matrix appears at the position of $A_{21}$. The appropriate linear combination is obtained by a pre-multiplication by

$$-A_{21}A_{11}^{-1}\begin{bmatrix}A_{11} & A_{12}\end{bmatrix} = \begin{bmatrix}-A_{21} & -A_{21}A_{11}^{-1}A_{12}\end{bmatrix}.$$

Hence the resulting matrix and the corresponding determinant are the following:

$$|A| = \begin{vmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{vmatrix} = \begin{vmatrix}A_{11} & A_{12}\\ O & A_{22}-A_{21}A_{11}^{-1}A_{12}\end{vmatrix}.$$

This is a triangular block matrix and hence the determinant is a product of the determinants of the diagonal blocks. That is,

$$|A| = |A_{11}||A_{22}-A_{21}A_{11}^{-1}A_{12}|.$$

If $A^{-1}$ exists then $AA^{-1}=I=A^{-1}A$. In the partitioned format in (N14.3),

$$AA^{-1}=I \quad \Rightarrow \quad \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}A^{11} & A^{12}\\ A^{21} & A^{22}\end{bmatrix} = \begin{bmatrix}I_r & O\\ O & I_s\end{bmatrix}.$$

Thus, by straight multiplication the following equations are determined where $r$ and $s$, denote the orders of the identity matrices where we assumed that $A_{11}$ is $r\times r$ and $A_{22}$ is $s\times s$.

$$\begin{aligned}A_{11}A^{11}+A_{12}A^{21}&=I_r\\ A_{11}A^{12}+A_{12}A^{22}&=O\\ A_{21}A^{11}+A_{22}A^{21}&=O\\ A_{21}A^{12}+A_{22}A^{22}&=I_s.\end{aligned} \tag{N14.6}$$

Solving the system in (N14.6), we have the following representations, among other results:

$$\begin{aligned}A^{11}&=[A_{11}-A_{12}A_{22}^{-1}A_{21}]^{-1}, \quad A_{11}^{-1}=A^{11}-A^{12}(A^{22})^{-1}A^{21}\\ A^{22}&=[A_{22}-A_{21}A_{11}^{-1}A_{12}]^{-1}, \quad A_{22}^{-1}=A^{22}-A^{21}(A^{11})^{-1}A^{12}.\end{aligned} \tag{N14.7}$$

Note that the submatrices in the inverse are not the inverses of the corresponding submatrices in the original matrix. That is, $A^{11}\neq A_{11}^{-1}$, $A^{22}\neq A_{22}^{-1}$. From (N14.6), one can also derive formulae for $A^{21}$ and $A^{12}$ in terms of the submatrices in $A$ and vice versa, which are not listed above. [This is left as an exercise to the student.]

For our illustrative example, it is easily verified that

$$A^{-1}=\frac{1}{3}\begin{bmatrix}1 & -\frac{1}{2} & \frac{1}{2}\\ -1 & 2 & 1\\ 1 & 1 & -1\end{bmatrix} = \begin{bmatrix}A^{11} & A^{12}\\ A^{21} & A^{22}\end{bmatrix}$$

where

$$A^{11} = \left[\frac{1}{3}\right], \quad A^{12} = \left[-\frac{1}{6}, \frac{1}{6}\right],$$

$$A^{21} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}, \quad A^{22} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} \end{bmatrix}.$$

From the computations earlier, we have, for example,

$$A_{22} - A_{21}A_{11}^{-1}A_{12} = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}$$

and hence

$$[A_{22} - A_{21}A_{11}^{-1}A_{12}]^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}^{-1} = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} = A^{22}.$$

Thus one result is verified. The remaining verifications are left to the students.

**Note 14.5** (Correlation coefficient)**.** This is very often a misused concept in applied statistics. The phrase “correlation” indicates “relationship”, and hence people misinterpret it as a measure of relationship between two real scalar random variables and the corresponding sample value as measuring the relationship between the pairs of numbers. There is extensive literature trying to evaluate the strength of the relationship, “negative relationship”, “positive relationship”, “increasing and decreasing nature of the relationship” etc by studying the correlation. But it is very easy to show that correlation does not measure relationship at all. Let $\rho$ denote the correlation between two real scalar random variables $x$ and $y$. Then it is easy to show that $-1 \le \rho \le 1$. This follows from Cauchy–Schwarz inequality or by using the property $|\cos\theta| \le 1$ or by considering two random variables:

$$u = \frac{x}{\sigma_1} + \frac{y}{\sigma_2} \quad \text{and} \quad v = \frac{x}{\sigma_1} - \frac{v}{\sigma_2}$$

where $x$ and $y$ are non-degenerate random variables with $\mathrm{Var}(x) = \sigma_1^2 > 0$ and $\mathrm{Var}(y) = \sigma_2^2 > 0$. Take $\mathrm{Var}(u)$ and $\mathrm{Var}(v)$ and use the fact that they are non-negative to show that $-1 \le \rho \le 1$. In the Cauchy–Schwarz inequality, equality is attained or the boundary values $\rho = +1$ and $\rho = -1$ are attained if and only if $x$ and $y$ are linearly related, that is, $y = a + bx$ where $a$ and $b \neq 0$ are constants. This relationship must hold almost surely meaning that there could be non-linear relationships but the total probability measure on the non-linear part must be zero or there must exist an equivalent linear function with probability 1. This aspect will be illustrated with an example later. Coming back to $\rho$, let us consider a perfect mathematical relationship between $x$ and $y$ in the form:

$$y = a + bx + cx^2, \quad c \neq 0. \tag{N14.8}$$

Since we are computing correlations, without loss of generality, we can omit $a$.

Further, for convenience let us assume that $x$ has a symmetrical distribution so that all odd moments disappear. Then $E(x) = 0$, $E(x^3) = 0$ and $E(y) = a + cE(x^2)$. Also we rule out degenerate variables when computing correlation, and hence it is assumed that $\operatorname{Var}(x) \neq 0$.

$$\begin{aligned}
\operatorname{Cov}(x,y) &= E\{x[y - E(y)]\} = E\{x[bx + c(x^2 - E(x^2))]\} \\
&= 0 + b\operatorname{Var}(x) + 0 = b\operatorname{Var}(x). \\
\operatorname{Var}(y) &= E[bx + c(x^2 - E(x^2))]^2 \\
&= b^2\operatorname{Var}(x) + c^2\{E(x^4) - [E(x^2)]^2\}.
\end{aligned}$$

Then

$$\begin{aligned}
\rho &= \frac{b\operatorname{Var}(x)}{\sqrt{\operatorname{Var}(x)}\sqrt{b^2\operatorname{Var}(x) + c^2\{E(x^4) - (E(x^2))^2\}}} \\
&= \frac{b}{|b|\sqrt{1 + \frac{c^2}{b^2}\{\frac{E(x^4)-(E(x^2))^2}{\operatorname{Var}(x)}\}}}, \quad \text{for } b \neq 0 \\
&= 0 \quad \text{if } b = 0 \\
&= \frac{1}{\sqrt{1 + \frac{c^2}{b^2}\{\frac{E(x^4)-(E(x^2))^2}{\operatorname{Var}(x)}\}}}, \quad \text{if } b > 0 \\
&= -\frac{1}{\sqrt{1 + \frac{c^2}{b^2}\{\frac{E(x^4)-(E(x^2))^2}{\operatorname{Var}(x)}\}}} \quad \text{if } b < 0.
\end{aligned} \tag{N14.9}$$

Let us take $x$ to be a standard normal variable, that is, $x \sim N(0,1)$, then we know that $E(x) = 0$, $E(x^2) = 1$, $E(x^4) = 3$. In this case,

$$\rho = \pm\frac{1}{\sqrt{1 + 2\frac{c^2}{b^2}}}, \tag{N14.10}$$

positive if $b > 0$, negative if $b < 0$ and zero if $b = 0$. Suppose that we would like to have $\rho = 0.01$ and at the same time a perfect mathematical relationship between $x$ and $y$, such as the one in (N14.8) existing. Then let $b > 0$ and let

$$\begin{aligned}
&\frac{1}{\sqrt{1 + 2\frac{c^2}{b^2}}} = 0.01 \quad \Rightarrow \quad 1 + 2\frac{c^2}{b^2} = \frac{1}{(0.01)^2} = 10\,000 \quad \Rightarrow \\
&2\frac{c^2}{b^2} = 9\,999 \quad \Rightarrow \quad c^2 = \frac{9\,999b^2}{2}.
\end{aligned}$$

Take any $b > 0$ such that $c^2 = \frac{9\,999b^2}{2}$. There are infinitely many choices. For example, $b = 1$ gives $c^2 = \frac{9\,999}{2}$. Similarly, if we want $\rho$ to be zero, then take $b = 0$. If we want $\rho$ to be a very high positive number such as 0.999 or a number close to $-1$ such as $-0.99$, then also there are infinitely many choices of $b$ and $c$ such that a perfect mathematical relationship existing between $x$ and $y$ and at the same time $\rho$ can be any small or large quantity between $-1$ and $+1$. Thus $\rho$ is not an indicator of relationship between $x$ and $y$. Even when there is a relationship between $x$ and $y$, other than linear relationship, $\rho$ can be anything between $-1$ and $+1$, and $\rho = \pm 1$ when and only when there is a linear relationship almost surely. From the quadratic function that we started with, note that increasing values of $x$ can go with increasing as well as decreasing values of $y$ when $\rho$ is positive or negative. Hence that type of interpretation cannot be given to $\rho$ either.

**Example 14.13.** Consider the following probability function for $x$:

$$f(x) = \begin{cases} \frac{1}{2}, & x = \alpha \\ \frac{1}{2}, & x = -\alpha \\ 0, & \text{elsewhere.} \end{cases}$$

Compute $\rho$ and check the quadratic relationship between $x$ and $y$ as given in (N14.8).

**Solution 14.13.** Here, $E(x) = 0$, $E(x^2) = \alpha^2$, $E(x^4) = \alpha^4$. Then $\rho$ in (N14.9) becomes

$$\rho = \frac{b}{|b|} = \pm 1$$

but $c \neq 0$ thereby (N14.8) holds. Is there something wrong with the Cauchy–Schwarz inequality? This is left to the student to think over.

Then what is the correlation coefficient $\rho$? What does it really measure? The numerator of $\rho$, namely $\mathrm{Cov}(x, y)$, measures the joint variation of $x$ and $y$ or the scatter of the point $(x, y)$, or angular dispersion between $x$ and $y$, corresponding to the scatter in $x$, $\mathrm{Var}(x) = \mathrm{Cov}(x, x)$. Then division by $\sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}$ has the effect of making the covariance scale-free. Hence $\rho$ really measures the joint variation of $x$ and $y$ or a type of scatter in the point $(x, y)$ in the sense that when $y = x$ it becomes $\mathrm{Var}(x)$ which is the square of a measure of scatter in $x$. Thus $\rho$ is more appropriately called a scale-free covariance and this author suggested through one of the published papers to call $\rho$ *scovariance* or scale-free covariance. It should never be interpreted as measuring relationship or linearity or near linearity or anything like that. Only two points $\rho = +1$ and $\rho = -1$ are connected to linearity and no other value of $\rho$ is connected to linearity or near linearity. For postulates defining covariance or for an axiomatic definition of covariance the student may consult the book [14].

## Exercises 14.5

**14.5.1.** Verify equations (N14.4) and (N14.5) for the following partitioned matrices:

$$A = \begin{bmatrix} 1 & -1 & 0 & 2 \\ 3 & 4 & 1 & 1 \\ 2 & 1 & -1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad A_{11} = \begin{bmatrix} 1 & -1 \\ 3 & 4 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 2 & -1 & 0 \\ 1 & -1 & 3 & 1 \\ -1 & 0 & 1 & 4 \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad B_{11} = [1].$$

**14.5.2.** For real scalar random variable $x$, let $E(x) = 0$, $E(x^2) = 4$, $E(x^3) = 6$, $E(x^4) = 24$. Let $y = 50 + x + cx^2$, $c \neq 0$. Compute the correlation between $x$ and $y$ and interpret it for various values of $c$.

**14.5.3.** Let $x$ be a standard normal variable, that is, $x \sim N(0,1)$. Let $y = a + x + 2x^2 + cx^3$, $c \neq 0$. Compute the correlation between $x$ and $y$ and interpret it for various values of $c$.

**14.5.4.** Let $x$ be a type-1 beta random variable with the parameters $(\alpha = 2, \beta = 1)$. Let $y = ax^\delta$ for some parameters $a$ and $\delta$. Compute the correlation between $x$ and $y$ and interpret it for various values of $a$ and $\delta$.

**14.5.5.** Repeat the Exercise in 14.5.4 if $x$ is type-2 beta with the parameters $(\alpha = 1, \beta = 2)$.

## 14.6 Regression analysis versus correlation analysis

As mentioned earlier, for studying regression one needs only the conditional distribution of $x_1$ given $x_2, \dots, x_k$ because the regression of $x_1$ on $x_2, \dots, x_k$ is the conditional expectation of $x_1$ given $x_2, \dots, x_k$, that is, $E(x_1|x_2, \dots, x_k)$. But for correlation analysis we need the joint distribution of all the variables involved. For example, in order to compute multiple correlation coefficient $\rho_{1.(2\dots k)}$ we need the joint moments involving all the variables $x_1, x_2, \dots, x_k$ up to second-order moments. Hence the joint distribution, not just the conditional distribution, is needed. Thus regression analysis and correlation analysis are built up on two different premises and should not be mixed up.

### 14.6.1 Multiple correlation ratio

In many of our examples, it is seen that the regression function is not linear in many situations. Let $E(x_1|x_2, \dots, x_k) = M(x_2, \dots, x_k)$, may or may not be a linear function of $x_2, \dots, x_k$. Consider an arbitrary predictor $g(x_2, \dots, x_k)$ for predicting $x_1$. To start with, we are assuming that there is a joint distribution of $x_1, x_2, \dots, x_k$. Let us compute the correlation between $x_1$ and an arbitrary predictor $g(x_2, \dots, x_k)$ for $x_1$.

$$\operatorname{Cov}(x_1, g) = E\{[x_1 - E(x_1)][g - E(g)]\} = E\{x_1[(g - E(g)]\} \tag{14.23}$$

as explained earlier since $E\{E(x_1)[g - E(g)]\} = E(x_1)E[g - E(g)]\} = 0$. Let us convert the expected value in (14.21) into an expectation of the conditional expectation through Lemma 14.1. Then

$$\begin{aligned}\operatorname{Cov}(x_1, g) &= E\{E[x_1(g - E(g)]|x_2, \dots, x_k]\}\\ &= E\{(g - E(g))E(x_1|x_2, \dots, x_k)\}, \quad \text{since } g \text{ is free of } x_1\\ &= E\{(g - E(g))M(x_2, \dots, x_k)\}\\ &= \operatorname{Cov}(g, M) \le \sqrt{\operatorname{Var}(g)\operatorname{Var}(M)},\end{aligned}$$

the last inequality follows from the fact that the correlation $\rho \le 1$. Then the correlation between $x_1$ and an arbitrary predictor, which includes linear predictors also, denoted by $\eta$, is given by the following:

$$\eta = \frac{\operatorname{Cov}(x_1, g)}{\sqrt{\operatorname{Var}(g)}\sqrt{\operatorname{Var}(x_1)}} \le \frac{\sqrt{\operatorname{Var}(g)}\sqrt{\operatorname{Var}(M)}}{\sqrt{\operatorname{Var}(g)}\sqrt{\operatorname{Var}(x_1)}} = \frac{\sqrt{\operatorname{Var}(M)}}{\sqrt{\operatorname{Var}(x_1)}}. \tag{14.24}$$

**Definition 14.4** (Multiple correlation ratio)**.** The maximum correlation between $x_1$ and an arbitrary predictor of $x_1$ by using $x_2, \dots, x_k$ is given by the following:

$$\max_g \eta = \frac{\sqrt{\operatorname{Var}(M)}}{\sqrt{\operatorname{Var}(x_1)}} = \eta_{1.(2\dots k)}. \tag{14.25}$$

This maximum correlation between $x_1$ and an arbitrary predictor of $x_1$ by using $x_2, \dots, x_k$ is given by $\sqrt{\frac{\operatorname{Var}(M)}{\operatorname{Var}(x_1)}}$ and it is defined as the *multiple correlation ratio* $\eta_{1.(2\dots k)}$ and the maximum is attained when the arbitrary predictor is the regression of $x_1$ on $x_2, \dots, x_k$, namely, $E(x_1|x_2, \dots, x_k) = M(x_2, \dots, x_k)$.

Note that when $M(x_2, \dots, x_k)$ is linear in $x_2, \dots, x_k$ we have the multiple correlation coefficient given in (14.21). Thus when $g$ is confined to linear predictors or in the class of linear predictors

$$\eta^2_{1.(2\dots k)} = \rho^2_{1.(2\dots k)} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}. \tag{14.26}$$

Some further properties of $\rho^2_{1.(2\dots k)}$ can be seen easily. Note that from (N14.7)

$$1 - \rho^2_{1.(2\dots k)} = \frac{\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}} = \frac{(\sigma^{11})^{-1}}{\sigma_{11}} = \frac{1}{\sigma_{11}\sigma^{11}}. \tag{14.27}$$

**Example 14.14.** Check whether the following matrix $\Sigma$ can represent the covariance matrix of $X' = (x_1, x_2, x_3)$ where $x_1, x_2, x_3$ are real scalar random variables. If so, evaluate $\rho_{1.(2.3)}$ and verify (14.27):

$$\Sigma = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 3 \end{bmatrix}.$$

**Solution 14.14.**

$$2 > 0, \quad \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4 > 0, \quad |\Sigma| = 8 > 0$$

and hence $\Sigma = \Sigma' > 0$ (positive definite) and hence it can represent the covariance matrix of $X$. For being a covariance matrix one needs only symmetry plus at least positive semi-definiteness. As per our notation,

$$\sigma_{11} = 2, \quad \Sigma_{12} = [0, 1], \quad \Sigma_{22} = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}, \quad \Sigma_{21} = \Sigma_{12}'$$

and, therefore,

$$\begin{aligned} \Sigma_{22}^{-1} &= \frac{1}{5}\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}, \\ \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}} &= \frac{1}{(5)(2)}[0, 1]\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \frac{1}{5} = \rho_{1.(2.3)}^2 \\ 1 - \rho_{1.(2.3)}^2 &= 1 - \frac{1}{5} = \frac{4}{5}; \\ \frac{\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}} &= \frac{1}{2}\left[2 - \frac{2}{5}\right] = \frac{4}{5}; \\ \frac{1}{\sigma_{11}\sigma^{11}} &= \frac{1}{2}\left(\frac{8}{5}\right) = \frac{4}{5}. \end{aligned}$$

Thus (14.27) is verified.

### 14.6.2 Multiple correlation as a function of the number of regressed variables

Let $X' = (x_1, \dots, x_k)$ and let the variance-covariance matrix in $X$ be denoted by $\Sigma = (\sigma_{ij})$. Our general notation for the multiple correlation coefficient is

$$\rho_{1.(2\dots k)} = \sqrt{\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}}.$$

For $k = 2$,

$$\rho_{1.2} = \sqrt{\frac{\sigma_{12}\sigma_{12}}{\sigma_{22}\sigma_{11}}} = \sqrt{\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}} = \rho_{12}$$

is the correlation between $x_1$ and $x_2$. For $k = 3$,

$$\rho_{1.(2.3)}^2 = \frac{1}{\sigma_{11}}[\sigma_{12}, \sigma_{13}] \begin{bmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{21} \\ \sigma_{31} \end{bmatrix}.$$

Converting everything on the right in terms of the correlations, that is,

$$\begin{aligned} &\sigma_{21} = \sigma_{12} = \rho_{12}\sigma_1\sigma_2, \quad \sigma_{11} = \sigma_1^2, \quad \sigma_{22} = \sigma_2^2, \\ &\sigma_{31} = \sigma_{13} = \rho_{13}\sigma_1\sigma_3, \quad \sigma_{23} = \rho_{23}\sigma_2\sigma_3, \end{aligned}$$

we have the following:

$$\begin{aligned} \rho_{1.(2.3)}^2 &= \frac{1}{\sigma_1^2}[\rho_{12}\sigma_1\sigma_2, \rho_{13}\sigma_1\sigma_3] \begin{bmatrix} \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{12}\sigma_1\sigma_2 \\ \rho_{13}\sigma_1\sigma_3 \end{bmatrix} \\ &= [\rho_{12}\sigma_2, \rho_{13}\sigma_3] \begin{bmatrix} \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{12}\sigma_2 \\ \rho_{13}\sigma_3 \end{bmatrix} \\ &= [\rho_{12}, \rho_{13}] \begin{bmatrix} \sigma_2 & 0 \\ 0 & \sigma_3 \end{bmatrix} \begin{bmatrix} \sigma_2 & 0 \\ 0 & \sigma_3 \end{bmatrix}^{-1} \\ &\quad \times \begin{bmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_2 & 0 \\ 0 & \sigma_3 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_2 & 0 \\ 0 & \sigma_3 \end{bmatrix} \begin{bmatrix} \rho_{12} \\ \rho_{13} \end{bmatrix} \\ &= [\rho_{12}, \rho_{13}] \begin{bmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{12} \\ \rho_{13} \end{bmatrix} \\ &= \frac{1}{1 - \rho_{23}^2}[\rho_{12}, \rho_{13}] \begin{bmatrix} 1 & -\rho_{23} \\ -\rho_{23} & 1 \end{bmatrix} \begin{bmatrix} \rho_{12} \\ \rho_{13} \end{bmatrix}, \quad 1 - \rho_{23}^2 > 0, \\ &= \frac{\rho_{12}^2 + \rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23}}{1 - \rho_{23}^2}. \end{aligned}$$

Then

$$\begin{aligned} \rho_{1.(2.3)}^2 - \rho_{12}^2 &= \frac{\rho_{12}^2 + \rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23}}{1 - \rho_{23}^2} - \rho_{12}^2 \\ &= \frac{\rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23} + \rho_{12}^2\rho_{23}^2}{1 - \rho_{23}^2} \\ &= \frac{(\rho_{13} - \rho_{12}\rho_{23})^2}{1 - \rho_{23}^2} \geq 0 \end{aligned}$$

which is equal to zero only when $\rho_{13} = \rho_{12}\rho_{23}$. Thus, in general,

$$\rho^2_{1.(2.3)} - \rho^2_{12} \geq 0 \quad \Rightarrow \quad \rho^2_{1.(2.3)} \geq \rho^2_{12}.$$

In other words, the multiple correlation coefficient increased when we incorporated one more variable $x_3$ in the regressed set. It is not difficult to show (left as an exercise to the student) that

$$\rho^2_{12} \leq \rho^2_{1.(2.3)} \leq \rho^2_{1.(2.3.4)} \leq \dots \tag{14.28}$$

This indicates that $\rho^2_{1.(2\dots k)}$ is an increasing function of $k$, the number of variables involved in the regressed set. There is a tendency among applied statisticians to use the sample multiple correlation coefficient as an indicator of how good is a linear regression function by looking at the value of the multiple correlation coefficient, in the sense, bigger the value better the model. From (14.28), it is evident that this is a fallacious approach. Also this approach comes from the tendency to look at the correlation coefficient as a measure of relationship, which again is a fallacious concept.

## Exercises 14.6

**14.6.1.** Show that $\rho^2_{1.(2.3)} \leq \rho^2_{1.(2.3.4)}$ with the standard notation for the multiple correlation coefficient $\rho_{1.(2\dots k)}$.

**14.6.2.** (i) Show that the following matrix $V$ can be a covariance matrix:

$$V = \begin{bmatrix} 1 & 1 & 0 & -1 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ -1 & 0 & 1 & 2 \end{bmatrix}.$$

(ii) Compute $\rho^2_{1.2}, \rho^2_{1(2.3)}, \rho^2_{1.(2.3.4)}$.
(iii) Verify that $\rho^2_{1.2} \leq \rho^2_{1.(2.3)} \leq \rho^2_{1.(2.3.4)}$.

**14.6.3.** Let the conditional density of $x_1$ given $x_2$ be Gaussian with mean value $1 + 2x_2 + x_2^2$ and variance 1 and let the marginal density of $x_2$ be uniform over $[0, 1]$. Compute the square of the correlation ratio of $x_1$ to $x_2$, that is,

$$\eta^2_{1.2} = \frac{\operatorname{Var}(M)}{\operatorname{Var}(x_1)}, \quad M = E(x_1|x_2)$$

and

$$\operatorname{Var}(x_1) = \operatorname{Var}[E(x_1|x_2)] + E[\operatorname{Var}(x_1|x_2)].$$

**14.6.4.** Let the conditional density of $x_1$ given $x_2$, $x_3$ be exponential with mean value $1+x_2+x_3+x_2x_3$ and let the joint density of $x_2$ and $x_3$ be

$$f(x_2,x_3)=x_2+x_3,\quad 0\le x_2\le 1,\ 0\le x_3\le 1$$

and zero elsewhere. Compute the square of the correlation ratio

$$\eta^2_{1.(2.3)}=\frac{\mathrm{Var}(M)}{\mathrm{Var}(x_1)}$$

where

$$M=E(x_1|x_2,x_3)$$

and

$$\mathrm{Var}(x_1)=\mathrm{Var}[E(x_1|x_2,x_3)]+E[\mathrm{Var}(x_1|x_2,x_3)].$$

**14.6.5.** Let the conditional density of $x_1$ given $x_2$, $x_3$ be Gaussian, $N(x_2+x_3+x_2x_3,1)$, where let $x_2$, $x_3$ have a joint density as in Exercise 14.6.4. Evaluate the square of the correlation ratio $\eta^2_{1.(2.3)}$.

There are other concepts of partial correlation coefficient, partial correlation ratio, etc., which fall in the general category of residual analysis in regression problems. We will not go into these aspects here. These will be covered in a module on model building. We will conclude this section with a note on variances and covariances of linear functions of random variables. These were already discussed in Module 6, which will be recalled here for ready reference.

**Note 14.6** (Variances and covariances of linear functions). Let $x_1,\dots,x_p$ be real scalar variables with $E(x_j)=\mu_j$, $\mathrm{Var}(x_j)=\sigma_{jj}$, $\mathrm{Cov}(x_i,x_j)=\sigma_{ij}, i,j=1,\dots,p$. Let

$$X=\begin{bmatrix}x_1\\ \vdots\\ x_p\end{bmatrix},\quad a=\begin{bmatrix}a_1\\ \vdots\\ a_p\end{bmatrix},\quad b=\begin{bmatrix}b_1\\ \vdots\\ b_p\end{bmatrix},\quad \Sigma=(\sigma_{ij})=\begin{bmatrix}\sigma_{11} & \dots & \sigma_{1p}\\ \sigma_{21} & \dots & \sigma_{2p}\\ \vdots & \dots & \vdots\\ \sigma_{p1} & \dots & \sigma_{pp}\end{bmatrix}$$

where $a$ and $b$ are constant vectors, a prime denotes the transpose, $E$ denotes the expected value and let $\mu'=(\mu_1,\dots,\mu_p)$, $\mu_j=E(x_j), j=1,\dots,p$. As per the definition,

$$\mathrm{Var}(x_j)=E[(x_j-E(x_j))^2],$$
$$\mathrm{Cov}(x_i,x_j)=E[(x_i-E(x_i))(x_j-E(x_j))];\quad \mathrm{Cov}(x_j,x_j)=\mathrm{Var}(x_j).$$

Then

$$u=a_1x_1+\dots+a_px_p\quad\Rightarrow\quad u=a'X=X'a$$

$$v = b_1x_1 + \cdots + b_px_p = b'X = X'b; \quad E(a'X) = a'E(X) = a'\mu$$
$$E(b'X) = b'E(X) = b'\mu; \quad \mathrm{Var}(a'X) = E[a'X - a'\mu]^2 = E[a'(X - \mu)]^2.$$

From elementary theory of matrices, it follows that if we have a $1 \times 1$ matrix $c$ then it is a scalar and its transpose is itself, that is, $c' = c$. Being a linear function, $a'(X - \mu)$ is a $1 \times 1$ matrix and hence it is equal to its transpose, which is, $(X - \mu)'a$. Hence we may write

$$E[a'(X - \mu)]^2 = E[a'(x - \mu)(x - \mu)'a] = a'E[(X - \mu)(X - \mu)']a$$

since $a$ is a constant the expected value can be taken inside. But

$$\begin{aligned} &E[(X - \mu)(X - \mu)'] \\ &= E\begin{bmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) & \dots & (x_1 - \mu_1)(x_p - \mu_p) \\ (x_2 - \mu_2)(x_1 - \mu_1) & (x_2 - \mu_2)^2 & \dots & (x_2 - \mu_2)(x_p - \mu_p) \\ \vdots & \vdots & \dots & \vdots \\ (x_p - \mu_p)(x_1 - \mu_1) & (x_p - \mu_p)(x_2 - \mu_2) & \dots & (x_p - \mu_p)^2 \end{bmatrix}. \end{aligned}$$

Taking expectations inside the matrix, we have

$$\begin{aligned} E[(X - \mu)(X - \mu)'] &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \\ &= \Sigma = \text{covariance matrix in } X. \end{aligned}$$

Therefore,

$$\mathrm{Var}(a'X) = a'\Sigma a, \quad \mathrm{Var}(b'X) = b'\Sigma b, \quad \mathrm{Cov}(a'X, b'X) = a'\Sigma b = b'\Sigma a$$

since $\Sigma = \Sigma'$.

Details on variances of linear functions and covariance between two linear functions are needed to deal with the area of *Canonical Correlation Analysis*. This is an area of predicting one set of variables by using another set of variables. In the regression problem that we considered in Sections 14.3–14.5, we were predicting one scalar variable by using one or more or one set of other scalar variables. We can generalize this idea and try to predict one set of scalar variables by using another set of scalar variables. Since individual variables are contained in linear functions, what is usually done is to maximize the correlation between one arbitrary linear function of one set of variables and another arbitrary linear function of the other set of variables. By maximizing the correlations, we construct the optimal linear functions, which are called pairs of canonical variables. This aspect will be dealt with in detail in the module on model building.

Another useful area is vector and matrix differential operators and their uses in multivariate statistical analysis. When estimating or testing hypotheses on the parameters in a multivariate statistical density, these operators will come in handy. Since the detailed discussion is beyond the scope of this book, we will just indicate the main ideas here for the benefit of curious students.

**Note 14.7** (Vector and matrix derivatives)**.** Consider the following vector of partial differential operators. Let

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \frac{\partial}{\partial Y} = \begin{bmatrix} \frac{\partial}{\partial y_1} \\ \vdots \\ \frac{\partial}{\partial y_n} \end{bmatrix}, \quad \frac{\partial}{\partial Y}[f] = \begin{bmatrix} \frac{\partial f}{\partial y_1} \\ \vdots \\ \frac{\partial f}{\partial y_n} \end{bmatrix}$$

where $f$ is a real-valued scalar function of $Y$. For example,

$$f_1(Y) = a_1y_1 + \cdots + a_ny_n = a'Y, \quad a' = (a_1, \dots, a_n) \tag{i}$$

is such a function, where $a$ is a constant vector. Here, $f_1$ is a linear function of $Y$, something like

$$2y_1 - y_2 + y_3; \quad y_1 + y_2 + \cdots + y_n; \quad y_1 + 3y_2 - y_3 + 2y_4$$

etc.

$$f_2(y_1, \dots, y_n) = y_1^2 + y_2^2 + \cdots + y_n^2 \tag{ii}$$

which is the sum of squares or a simple quadratic form or a general quadratic form in its canonical form.

$$f_3(y_1, \dots, y_n) = Y'AY, \quad A = (a_{ij}) = A' \tag{iii}$$

is a general quadratic form where $A$ is a known constant matrix, which can be taken to be symmetric without loss of generality. A few basic properties that we are going to use will be listed here as lemmas.

**Lemma 14.4.**

$$f_1 = a'Y \quad \Rightarrow \quad \frac{\partial f_1}{\partial Y} = a.$$

Note that the partial derivative of the linear function $a'Y = a_1y_1 + \cdots + a_ny_n$, with respect to $y_j$ gives $a_j$ for $j = 1, \dots, n$, and hence the column vector

$$\frac{\partial f_1}{\partial Y} = \frac{\partial(a'Y)}{\partial Y} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a.$$

For example, if $a'Y = y_1 - y_2 + 2y_3$ then

$$\frac{\partial}{\partial Y}(a'Y) = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}.$$

**Lemma 14.5.**

$$f_2 = y_1^2 + \cdots + y_n^2 = Y'Y \quad \Rightarrow \quad \frac{\partial f_2}{\partial Y} = 2Y = 2\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Note that $Y'Y$ is a scalar function of $Y$ whereas $YY'$ is a $n \times n$ matrix, and hence it is a matrix function of $Y$. Note also that when $Y'Y$ is differentiated with respect to $Y$ the $Y'$ disappears and a 2 comes in. We get a column vector because our differential operator is a column vector.

**Lemma 14.6.**

$$f_3 = Y'AY, \quad A = A' \quad \Rightarrow \quad \frac{\partial f_3}{\partial Y} = 2AY.$$

Here, it can be seen that if $A$ is not taken as symmetric then instead of $2AY$ we will end up with $(A+A')Y$. As an illustration of $f_3$, we can consider

$$\begin{aligned} f_3 &= 2y_1^2 + y_2^2 + y_3^2 - 2y_1y_2 + 5y_2y_3 \\ &= [y_1, y_2, y_3]\begin{bmatrix} 2 & -1 & 0 \\ -1 & 1 & \frac{5}{2} \\ 0 & \frac{5}{2} & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = Y'AY, \quad A = A' \\ &= [y_1, y_2, y_3]\begin{bmatrix} 2 & -2 & 0 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = Y'BY, \quad B \neq B'. \end{aligned}$$

In the first representation, the matrix $A$ is symmetric whereas in the second representation of the same quadratic form the matrix $B$ is not symmetric. By straight differentiation,

$$\frac{\partial f_3}{\partial y_1} = 4y_1 - 2y_2, \quad \frac{\partial f_3}{\partial y_2} = 2y_2 - 2y_1 + 5y_3, \quad \frac{\partial f_3}{\partial y_3} = 2y_3 + 5y_2$$

Therefore,

$$\begin{aligned} \frac{\partial f_3}{\partial Y} &= \begin{bmatrix} \frac{\partial f_3}{\partial y_1} \\ \frac{\partial f_3}{\partial y_2} \\ \frac{\partial f_3}{\partial y_3} \end{bmatrix} = \begin{bmatrix} 4y_1 - 2y_2 \\ 2y_2 - 2y_1 + 5y_3 \\ 2y_3 + 5y_2 \end{bmatrix} \\ &= 2\begin{bmatrix} 2 & -1 & 0 \\ -1 & 1 & \frac{5}{2} \\ 0 & \frac{5}{2} & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = 2AY. \end{aligned}$$

But

$$
\begin{aligned}
B+B' &= \begin{bmatrix} 2 & -2 & 0 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 5 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 4 & -2 & 0 \\ -2 & 2 & 5 \\ 0 & 5 & 2 \end{bmatrix} = 2\begin{bmatrix} 2 & -1 & 0 \\ -1 & 1 & \frac{5}{2} \\ 0 & \frac{5}{2} & 1 \end{bmatrix} = 2A.
\end{aligned}
$$

Note that when applying Lemma 14.5 write the matrix in the quadratic form as a symmetric matrix. This can be done without any loss of generality since for any square matrix $B$, $\frac{1}{2}(B+B')$ is a symmetric matrix. Then when operating with the partial differential operator $\frac{\partial}{\partial Y}$ on $Y'AY, A = A'$ the net result is to delete $Y'$ (not $Y$) and premultiply by 2 or write $2AY$.

With the help of Note 14.7, one can now evaluate the pairs of canonical variables by using the vector and matrix differential operators. When we consider linear functions $u = a_1x_1 + \cdots + a_mx_m = a'X$, $v = b_1y_1 + \cdots + b_ny_n = b'Y$, where $a' = (a_1,\dots,a_m)$, $X' = (x_1,\dots,x_n)$, $b' = (b_1,\dots,b_n)$, $Y' = (y_1,\dots,y_n)$. Then $\mathrm{Var}(u) = a'\Sigma_1 a$, $\mathrm{Var}(v) = b'\Sigma_2 b$ where $\Sigma_1$ and $\Sigma_2$ are the covariance matrices in $X$ and $Y$, respectively. Since $a$ and $b$ are arbitrary, $\mathrm{Var}(u)$ and $\mathrm{Var}(v)$ can be arbitrarily large and hence when maximizing the covariance between $u$ and $v$ confine to unit hyperspheres or put the conditions $\mathrm{Var}(u) = 1$ and $\mathrm{Var}(v) = 1$. Construction of canonical variables is left as an exercise to the student.

## 14.7 Estimation of the regression function

In the earlier sections, we looked at prediction functions and “best predictors”, best in the minimum mean square sense. We found that in this case the “best” predictor of a dependent real scalar variable $y$ at preassigned values of the real scalar variables $x_1,\dots,x_k$ would be the conditional expectation of $y$ given $x_1,\dots,x_k$. For computing this conditional expectation, so that we have a good predictor function, we need at least the conditional distribution of $y$ given $x_1,\dots,x_k$. If the joint distribution of $y$ and $x_1,\dots,x_k$ is available that is also fine, but in a joint distribution, there is more information than what we need. In most of the practical situations, we may have some idea about the conditional expectation but we may not know the conditional distribution. In this case, we cannot explicitly evaluate the regression function analytically. Hence we will consider various scenarios in this section.

The problem that we will consider in this section is the situation that it is known that there exists the conditional expectation but we do not know the conditional distribution but a general idea is available about the nature of the conditional expectation or the regression function such as that the regression function is linear in the regressed variables or a polynomial type or some such known functions. Then the procedure is to collect observations on the variables, estimate the regression function and then use this estimated regression function to estimate the value of the dependent variable $y$ at preassigned values of $x_1, \dots, x_k$, the regressed variables. We will start with $k = 1$, namely one real scalar variable to be predicted by preassigning one independent real scalar variable. Let us start with the linear regression function, here "linear" means linear in the regressed variable or the so-called "independent variable".

### 14.7.1 Estimation of linear regression of $y$ on $x$

This means that the regression of the real scalar random variable $y$ on the real scalar random variable $x$ is believed to be of the form:

$$E(y|x) = \beta_0 + \beta_1 x \tag{14.29}$$

where $\beta_0$ and $\beta_1$ are unknown because the conditional distribution is not available and the only information available is that the conditional expectation, or the regression of $y$ on $x$, is linear of the type (14.29). In order to estimate $\beta_0$ and $\beta_1$, we will start with the model

$$y = a + bx \tag{14.30}$$

and try to take observations on the pair $(y, x)$. Let there be $n$ data points $(y_1, x_1), \dots, (y_n, x_n)$. Then as per the model in (14.30) when $x = x_j$ the estimated value, as per the model (14.30), is $a + bx_j$ but this estimated value need not be equal to the observed $y_j$ of $y$. Hence the error in estimating $y$ by using $a + bx_j$ is the following, denoting it by $e_j$:

$$e_j = y_j - (a + bx_j). \tag{14.31}$$

When the model is written, the following conventions are used. We write the model as $y = a + bx$ or $y_j = a + bx_j + e_j$, $j = 1, \dots, n$. The error $e_j$ can be positive for some $j$, negative for some other $j$ and zero for some other $j$. Then trying to minimize the errors by minimizing the sum of the errors is not a proper procedure to be used because the sum of $e_j$'s may be zero but this does not mean that there is no error. Here, the negative and positive values may sum up to zero. Hence a proper quantity to be used is a measure of mathematical "distance" between $y_j$ and $a + bx_j$ or a norm in $e_j$'s. The sum of squares of the errors, namely $\sum_{j=1}^n e_j^2$ is a squared norm or the square of the Euclidean distance between $y_j$'s and $a + bx_j$'s. For a real quantity if the square is zero, then the quantity itself is zero and if the square attains a minimum then we can say that the distance between the observed $y$ and the estimated $y$, estimated by the model $y = a + bx$, is minimized. For the model in (14.31),

$$\sum_{j=1}^n e_j^2 = \sum_{j=1}^n (y_j - a - bx_j)^2. \tag{14.32}$$

If the model corresponding to (14.31) is a general function $g(a_1,\dots,a_r,x_1,\dots,x_k)$, for some $g$ where $a_1,\dots,a_r$ are unknown constants in the model, $x_1,\dots,x_k$ are the regressed variables, then the $j$-th observation on $(x_1,\dots,x_k)$ is denoted by $(x_{1j},x_{2j},\dots,x_{kj})$, $j=1,\dots,n$ and then the error sum of squares can be written as

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} [y_j - g(a_1,\dots,a_r,x_{1j},\dots,x_{kj})]^2. \tag{14.33}$$

The unknown quantities in (14.33) are $a_1,\dots,a_r$. If the unknown quantities $a_1,\dots,a_r$, which are also called the *parameters in the model*, are estimated by minimizing the error sum of squares then the method is known as the *method of least squares*, introduced originally by Gauss. For our simple model in (14.32), there are two parameters $a,b$ and the minimization is to be done with respect to $a$ and $b$. Observe that in (14.33) the functional form of $g$ on $x_1,\dots,x_k$ is unimportant because some observations on these variables only appear in (14.33) but the nature of the parameters in (14.33) is important or (14.33) is a function of the unknown quantities $a_1,\dots,a_r$. Thus when we say that a model is linear it means linear in the unknowns, namely linear in the parameters. If we say that the model is a quadratic model, then it is a quadratic function in the unknown parameters. Note the subtle difference. When we say that we have a linear regression, then we are talking about the linearity in the regressed variables where the coefficients are known quantities, available from the conditional distribution. But when we set up a model to estimate a regression function then the unknown quantities in the model are the parameters to be estimated, and hence the degrees go with the degrees of the parameters.

Let us look at the minimization of the sum of squares of the errors in (14.32). This can be done either by using purely algebraic procedures or by using calculus. If we use calculus, then we differentiate partially with respect to the parameters $a$ and $b$ and equate to zero and solve the resulting equations.

$$\frac{\partial}{\partial a}\left[\sum_{j=1}^{n} e_j^2\right]=0, \quad \frac{\partial}{\partial b}\left[\sum_{j=1}^{n} e_j^2\right]=0 \quad \Rightarrow$$

$$-2\sum_{j=1}^{n}(y_j - a - bx_j) = 0, \quad \Rightarrow \quad \sum_{j=1}^{n}(y_j - \hat{a} - \hat{b}x_j) = 0. \tag{14.34}$$

$$-2\sum_{j=1}^{n}x_j(y_j - a - bx_j) = 0 \quad \Rightarrow \quad \sum_{j=1}^{n}x_j(y_j - \hat{a} - \hat{b}x_j) = 0. \tag{14.35}$$

Equations (14.34) and (14.35) do not hold universally for all values of the parameters $a$ and $b$. They hold only at the critical points. The critical points are denoted by $\hat{a}$ and $\hat{b}$, respectively. Taking the sum over all terms and over $j$, one has the following:

$$\sum_{j=1}^{n} y_j - n\hat{a} - \hat{b}\sum_{j=1}^{n} x_j = 0 \quad \text{and} \quad \sum_{j=1}^{n} x_j y_j - \hat{a}\sum_{j=1}^{n} x_j - \hat{b}\sum_{j=1}^{n} x_j^2 = 0. \tag{14.36}$$

In order to simplify the equations in (14.36), we will use the following convenient notations. [These are also standard notations.]

$$\bar{y} = \sum_{j=1}^{n} \frac{y_j}{n}, \quad \bar{x} = \sum_{j=1}^{n} \frac{x_j}{n}, \quad s_x^2 = \sum_{j=1}^{n} \frac{(x_j - \bar{x})^2}{n}$$

$$s_y^2 = \sum_{j=1}^{n} \frac{(y_j - \bar{y})^2}{n}, \quad s_{xy} = \sum_{j=1}^{n} \frac{(x_j - \bar{x})(y_j - \bar{y})}{n} = s_{yx}.$$

These are the sample means, sample variances and the sample covariance. Under these notations, the first equation in (14.36) reduces to the following, by dividing by $n$.

$$\bar{y} - \hat{a} - \hat{b}\bar{x} = 0.$$

Substituting for $\hat{a}$ in the second equation in (14.36), and dividing by $n$, we have

$$\sum_{j=1}^{n} \frac{x_j y_j}{n} - [\bar{y} - \hat{b}\bar{x}]\bar{x} - \hat{b}\sum_{j=1}^{n} \frac{x_j^2}{n} = 0.$$

Therefore,

$$\begin{aligned} \hat{b} &= \frac{\sum_{j=1}^{n} \frac{x_j y_j}{n} - (\bar{x})(\bar{y})}{\sum_{j=1}^{n} \frac{x_j^2}{n} - (\bar{x})^2} \\ &= \frac{s_{xy}}{s_x^2} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \quad \text{and} \quad \hat{a} = \bar{y} - \hat{b}\bar{x}. \end{aligned} \tag{14.37}$$

The simplifications are done by using the following formulae. For any set of real numbers $(x_1, y_1), \dots, (x_n, y_n)$,

$$\sum_{j=1}^{n}(x_j - \bar{x}) = 0, \quad \sum_{j=1}^{n}(y_j - \bar{y}) = 0, \quad \sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - n(\bar{x})^2$$

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n} y_j^2 - n(\bar{y})^2, \quad \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \sum_{j=1}^{n}(x_j y_j) - n(\bar{x}\bar{y}).$$

When we used calculus to obtain (14.36), we have noted that there is only one critical point $(\hat{a}, \hat{b})$ for our problem under consideration. Does this point $(\hat{a}, \hat{b})$ in the parameter space $\Omega = \{(a, b) \mid -\infty < a < \infty, -\infty < b < \infty\}$ correspond to a maximum or minimum? Note that since (14.32) is the sum of squares of real numbers the maximum for $\sum_{j=1}^{n} e_j^2$ for all $a$ and $b$, is at $+\infty$. Hence the only critical point $(\hat{a}, \hat{b})$ in fact corresponds to a minimum. Thus our estimated regression function, under the assumption that the regression was of the form $E(y|x) = \beta_0 + \beta_1 x$, and then estimating it by using the method of least squares, is

$$y = \hat{a} + \hat{b}x, \quad \hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{s_{xy}}{s_x^2}. \tag{14.38}$$

Hence (14.38) is to be used to estimate the values of $y$ at preassigned values of $x$.

**Example 14.15.** The growth of a certain plant $y$, growth measured in terms of its height in centimeters, is guessed to have a linear regression on $x$ the time measured in the units of weeks. Here, $x = 0$ means the starting of the observations, $x = 1$ means at the end of the first week, $x = 2$ means at the end of the second week and so on. The following observations are made:

| $x$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $y$ | 2.0 | 4.5 | 5.5 | 7.5 | 10.5 |

Estimate the regression function and then estimate $y$ at $x = 3.5$, $x = 7$.

**Solution 14.15.** As per our notation, $n = 5$,

$$\bar{x} = \frac{0+1+2+3+4}{5} = 2, \quad \bar{y} = \frac{2.0+4.5+5.5+7.5+10.5}{5} = 6.$$

If you are using a computer with a built-in or loaded program for "regression", then by feeding the observations $(x,y) = (0,2),(1,4.5),(2,5.5),(3,7.5),\ (4,10.5)$ the estimated linear function is readily printed. The same thing is achieved if you have a programmable calculator. If nothing is available to you readily and if you have to do the problem by hand, then for doing the computations fast, form the following table:

| $y$ | $x$ | $y-\bar{y}$ | $x-\bar{x}$ | $(x-\bar{x})^2$ | $(y-\bar{y})(x-\bar{x})$ | $\hat{y}$ | $y-\hat{y}$ | $(y-\hat{y})^2$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0 | −4.0 | −2 | 4 | 8.0 | 2 | 0.0 | 0.00 |
| 4.5 | 1 | −1.5 | −1 | 1 | 1.5 | 4 | 0.5 | 0.25 |
| 5.5 | 2 | −0.5 | 0 | 0 | 0 | 6 | −0.5 | 0.25 |
| 7.5 | 3 | 1.5 | 1 | 1 | 1.5 | 7 | 0.5 | 0.25 |
| 10.5 | 4 | 4.5 | 2 | 4 | 9.0 | 10 | 0.5 | 0.25 |
| | | | | 10 | 20.0 | | | 1.00 |

Therefore,

$$\hat{b} = \frac{\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y})}{\sum_{j=1}^{n}(x_j-\bar{x})^2} = \frac{20}{10} = 2$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 6 - (2)(2) = 2.$$

**Note 14.8.** Do not round up the estimated values. If an estimated value is 2.1, leave it as 2.1 and do not round it up to 2. Similarly, when averages are taken, then also do not round up the values of $\bar{x}$ and $\bar{y}$. If you are filling up sacks with coconuts and if 4 020 coconuts are filled in 100 sacks, then the average number in each sack is $\frac{4\,020}{100} = 40.2$ and it is not 40 because $40 \times 100 \neq 4\,020$.

Hence the estimated regression function is

$$y = 2 + 2x.$$

Then the estimated value $\hat{y}$ of $y$ at $x = 3.5$ is given by $\hat{y} = 2 + 2(3.5) = 9$.

**Note 14.9.** The point $x = 7$ is far outside the range of the data points. In the observations, the range of $x$ is only $0 \leq x \leq 4$ whereas we are asked to estimate $y$ at $x = 7$, which is far out from 4. The estimated function $y = 2 + 2x$ can be used for this purpose if we are 100% sure that the underlying regression is the same function $2 + 2x$ for all values of $x$ then we can use $x = 7$ and obtain the estimated $y$ as $\hat{y} = 2+2(7) = 16$. If there is any doubt as to the nature of the function at $x = 7$ then $y$ should not be estimated at a point for $x$ which is far out of the observational range for $x$.

**Note 14.10.** In the above table for carrying out computations, the last 3 columns, namely, $\hat{y}$, $y - \hat{y}$, $(y - \hat{y})^2$ are constructed for making some other calculations later on. The least square minimum is given by the last column sum and in this example it is 1.

Before proceeding further, let us introduce some more technical terms. If we apply calculus on the error sum of squares under the general model in (14.33), we obtain the following equations for evaluating the critical points:

$$\frac{\partial}{\partial a_1}\left[\sum_{j=1}^{n} e_j^2\right] = 0, \dots, \frac{\partial}{\partial a_r}\left[\sum_{j=1}^{n} e_j^2\right] = 0. \tag{14.39}$$

These minimizing equations in (14.39) under the least square analysis, are often called *normal equations*. This is another awkward technical term in statistics and it has nothing to do with normality or Gaussian distribution or it does not mean that other equations have some abnormalities. The nature of the equations in (14.39) will depend upon the nature of the involvement of the parameters $a_1, \dots, a_r$ with the regressed variables $x_1, \dots, x_k$.

**Note 14.11.** What should be the size of $n$ or how many observations are needed to carry out the estimation process? If $g(a_1, \dots, a_r, x_1, \dots, x_k)$ is a linear function of the form,

$$a_0 + a_1x_1 + \cdots + a_kx_k$$

then there are $k + 1$ parameters $a_0, \dots, a_k$ and (14.39) leads to $k + 1$ linear equations in $k + 1$ parameters. This means, in order to estimate $a_0, \dots, a_k$ we need at least $k + 1$ observation points if the system of linear equations is consistent or have at least one solution. Hence in this case $n \geq k + 1$. In a non-linear situation, the number of observations needed may be plenty more in order to estimate all parameters successfully. Hence the minimum condition needed on $n$ is $n \geq k + 1$ where $k + 1$ is the total number of parameters in a model and the model is linear in these $k+1$ parameters. Since it is not a mathematical problem of solving a system of linear equations, the practical advice is to take $n$ as large as feasible under the given situation so that a wide range of observational points will be involved in the model.

**Note 14.12.** As a reasonable criterion for estimating $y$ based on $g(a_1,\dots,a_r,x_1,\dots,x_k)$, we used the error sum of squares, namely

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} [y_j - g(a_1,\dots,a_r,x_{1j},\dots,x_{kj})]^2. \tag{14.40}$$

This is the square of a mathematical distance between $y_j$ and $g$. We could have used other measures of distance between $y_j$ and $g$, for example,

$$\sum_{j=1}^{n} |y_j - g(a_1,\dots,a_r,x_{1j},\dots,x_{kj})|. \tag{14.41}$$

Then minimization of this distance and estimation of the parameters $a_1,\dots,a_r$ thereby estimating the function $g$ is a valid and reasonable procedure. Then why did we choose the squared distance as in (14.40) rather than any other distance such as the one in (14.41)? This is done only for mathematical convenience. For example, if we try to use calculus then differentiation of (14.41) will be rather difficult compared to (14.40).

### 14.7.2 Inference on the parameters of a simple linear model

Consider the linear model

$$y_j = a + bx_j + e_j, \quad j = 1,\dots,n$$

where we wish to test hypotheses on the parameters $a$ and $b$ as well as construct confidence intervals for these. These can be done by making some assumptions on the error variable $e_j$, $j = 1,\dots,n$. Note that $x_j$'s are constants or preassigned numbers and the only variables on the right are the $e_j$'s, thereby $y_j$'s are also random variables. Let us assume that $e_j$'s are such that $E(e_j) = 0$, $\mathrm{Var}(e_j) = \sigma^2$, $j = 1,\dots,n$ and mutually noncorrelated. Then we can examine the least square estimators for $a$ and $b$. We have seen that

$$\begin{aligned}
y_j &= a + bx_j + e_j \quad \Rightarrow \quad \bar{y} = a + b\bar{x} + \bar{e} \\
\hat{b} &= \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \sum_{j=1}^{n} d_j(y_j - \bar{y}), \\
&= b + \sum_{j=1}^{n} d_j e_j, \quad d_j = \frac{(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \sum_j d_j = 0
\end{aligned} \tag{a}$$

$$
\begin{aligned}
E(\hat{b}) &= b \quad \text{since } E(e_j) = 0, \quad E(\bar{e}) = 0 \\
\operatorname{Var}(\hat{b}) &= E[\hat{b} - b]^2 = E\left[\sum_{j=1}^{n} d_j e_j\right]^2 \quad \text{from (a)} \\
&= \sum_{j=1}^{n} d_j^2 E(e_j^2) + 0 = \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} && \text{(b)} \\
\hat{a} &= \bar{y} - \hat{b}\bar{x} = [a + b\bar{x} + \bar{e}] - \hat{b}\bar{x} = a + \bar{x}[b - \hat{b}] + \bar{e} \\
E[\hat{a}] &= a \quad \text{since } E(\bar{e}) = 0, \quad E(\hat{b}) = b && \text{(c)} \\
\operatorname{Var}(\hat{a}) &= E[\hat{a} - a]^2 = E[\bar{x}(\hat{b} - b) + \bar{e}]^2 \\
&= (\bar{x})^2 \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} + \frac{\sigma^2}{n} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right] && \text{(d)}
\end{aligned}
$$

If we assume further that $e_j \sim N(0, \sigma^2)$, $j = 1, \ldots, n$, that is, iid $N(0, \sigma^2)$, then both $\hat{b}$ and $\hat{a}$ will the normally distributed, being linear functions of normal variables, since the $x_j$'s are constants. In this case,

$$
\begin{aligned}
u &= \frac{\hat{b} - b}{\sigma\sqrt{\frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \\
&= \sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\left[\frac{\hat{b} - b}{\sigma}\right] \sim N(0, 1)
\end{aligned}
$$

and

$$
v = \frac{\hat{a} - a}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \sim N(0, 1).
$$

But usually $\sigma^2$ is unknown. Hence if we replace $\sigma^2$ by an unbiased estimator of $\sigma^2$ then we should get a Student-$t$ statistic. We can show that the least square minimum, denoted by $s^2$, divided by $n - 2$ is an unbiased estimator for $\sigma^2$ for $n > 2$. [We will show this later for the general linear model in Section 14.7.4]. Hence

$$
u_1 = \sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\left[\frac{\hat{b} - b}{\hat{\sigma}}\right] \sim t_{n-2} \tag{14.42}
$$

and

$$
v_1 = \frac{\hat{a} - a}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}} \sim t_{n-2} \tag{14.43}
$$

where $t_{n-2}$ is a Student-$t$ with $n - 2$ degrees of freedom, and

$$
\hat{\sigma}^2 = \frac{\text{least square minimum}}{n - 2} = \frac{s^2}{n - 2}.
$$

Hence we can construct confidence intervals as well as test hypotheses on $a$ and $b$ by using (14.42) and (14.43). A $100(1-\alpha)\%$ confidence interval for $a$ is

$$\hat{a} \mp t_{n-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}} \tag{14.44}$$

and that for $b$ is

$$\hat{b} \mp t_{n-2,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{\sum_{j=1}^{n}(x_j-\bar{x})^2}}. \tag{14.45}$$

Details may be seen from Chapter 12, and illustration is as in Figure 12.3.

The usual hypothesis that we would like to test is $H_0 : b = 0$, or in other words, there is no effect of $x$ in estimating $y$ or $x$ is not relevant as far as the prediction of $y$ is concerned. We will consider general hypotheses of the types $H_o : b = b_0$ (given) and $H_0 : a = a_0$ (given), against the natural alternates. The test criterion will reduce to the following:

$H_0 : b = b_0$ (given), $H_1 : b \neq b_0$; criterion: reject $H_0$ if the observed value of

$$\left|\sqrt{\sum_{j=1}^{n}(x_j-\bar{x})^2}\left[\frac{\hat{b}-b_0}{\hat{\sigma}}\right]\right| \geq t_{n-2,\frac{\alpha}{2}}. \tag{14.46}$$

For testing $H_0 : a = a_0$ (given), $H_1 : a \neq a_0$; criterion: reject $H_0$ if the observed value of

$$\left|\frac{\hat{a}-a_0}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}}}\right| \geq t_{n-2,\frac{\alpha}{2}}, \tag{14.47}$$

where in both $\hat{\sigma}^2 = \frac{s^2}{n-2}, n > 2$ with $s^2$ being the least square minimum. If hypotheses of the type $H_0 : b \leq b_0$ or $H_0 : b \geq b_0$ the procedure is described in the section on testing hypotheses by using Student-$t$ statistic in Section 13.3.2.

**Example 14.16.** By using the data and linear model in Example 14.15, construct 95% confidence intervals for $a$ and $b$ and test the hypotheses $H_0 : b = 0$ and $H_0 : a = 3$ at a 5% level of rejection.

**Solution 14.16.** We have made all the computations in the solution of Example 14.15. Here, $n = 5$, which means the degrees of freedom $n - 2 = 3$. We want 95% confidence interval, which means our $\alpha = 0.05$ or $\frac{\alpha}{2} = 0.025$. The tabled value of $t_{3,0.025} = 3.182$. Observed value of $\hat{b} = 2$ and observed value of $\hat{a} = 2$. From our data, $\sum_{j=1}^{n}(x_j-\bar{x})^2 = 10$ and least square minimum $s^2 = 1$. A 95% confidence interval for $b$ is given by

$$\begin{aligned}\hat{b} \mp t_{n-2,\frac{\alpha}{2}}\sqrt{\frac{\text{least square minimum}}{3\sum_{j=1}^{n}(x_j-\bar{x})^2}} &= 2 \mp (3.182)\sqrt{\frac{1}{3\times 10}}\\ &\approx [-15.43, 19.43].\end{aligned}$$

A 95% confidence interval for $a$ is given by

$$\hat{a} \mp t_{n-2,\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}} = 2 \mp (3.182)\sqrt{\frac{1}{3}}\sqrt{\frac{1}{5} + \frac{4}{10}}$$

$$\approx [0.58, 3.42].$$

For testing $H_0 : b = 0$, we reject if the observed value of

$$\left|\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\left[\frac{\hat{b} - 0}{\hat{\sigma}}\right]\right| = \sqrt{10}\left|\sqrt{3}(2 - 0)\right|$$

$$\approx 10.95 \geq t_{n-2,\alpha/2} = t_{3,0.025} = 3.182.$$

Hence the hypothesis is rejected at the 5% level of rejection. For $H_0 : a = 3$, we reject the hypothesis, at a 5% level of rejection, if the observed value of

$$|\hat{a} - a_0|\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}} = |2 - 3|\sqrt{\frac{1}{3}}\sqrt{\frac{1}{5} + \frac{4}{10}}$$

$$\approx 0.45 \geq t_{3,0.025} = 3.182.$$

Hence this hypothesis is not rejected at the 5% level of rejection.

## Exercises 14.7

**14.7.1.** The weight gain $y$ in grams of an experimental cow under a certain diet $x$ in kilograms is the following:

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $y$ | 2 | 6 | 10 | 18 | 30 | 40 |

(i) Fit the model $y = a + bx$ to this data; (ii) compute the least square minimum; (iii) estimate the weight gain at $x = 3.5, x = 2.6$.

**14.7.2.** For the same data in Exercise 14.7.1, fit the model $y = a + bx + cx^2$, $c \neq 0$. (i) Compute the least square minimum; (ii) by comparing the least square minima in Exercises 14.7.1 and 14.7.2 check to see which model can be taken as a better fit to the data.

**14.7.3.** For the data and model in Exercise 14.7.1, construct a 99% confidence interval for $a$ as well as for $b$, and test, at 1% level of rejection, the hypotheses $H_0 : b = 0$ and $H_0 : a = 5$.

### 14.7.3 Linear regression of $y$ on $x_1, \dots, x_k$

Suppose that the regression of $y$ on $x_1, \dots, x_k$ is suspected to be linear in $x_1, \dots, x_k$, that is, of the form:

$$E(y|x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Suppose that $n$ data points $(y_j, x_{1j}, \dots, x_{kj})$, $j = 1, \dots, n$ are available. Since the regression is suspected to be linear and if we want to estimate the regression function, then we will start with the model

$$y = a_0 + a_1 x_1 + \dots + a_k x_k.$$

Hence at the $j$-th data point if the error in estimating $y$ is denoted by $e_j$ then

$$e_j = y_j - [a_0 + a_1 x_{1j} + \dots + a_k x_{kj}], \quad j = 1, \dots, n$$

and the error sum of squares is then

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} [y_j - a_0 - a_1 x_{1j} - \dots - a_k x_{kj}]^2. \tag{14.48}$$

We obtain the normal equations by differentiating partially with respect to $a_0$, $a_1, \dots, a_k$ and equating to zeros. That is,

$$\frac{\partial}{\partial a_0}\left[\sum_{j=1}^{n} e_j^2\right] = 0 \quad \Rightarrow \quad -2\left[\sum_{j=1}^{n} y_j - n\hat{a}_0 - \hat{a}_1 \sum_{j=1}^{n} x_{1j} - \dots - \hat{a}_k \sum_{j=1}^{n} a_{kj}\right] = 0.$$

We can delete $-2$ and divide by $n$. Then

$$\begin{aligned} \bar{y} &= \hat{a}_0 + \hat{a}_1 \bar{x}_1 + \dots + \hat{a}_k \bar{x}_k \quad \text{or} \\ \hat{a}_0 &= \bar{y} - \hat{a}_1 \bar{x}_1 - \dots - \hat{a}_k \bar{x}_k \end{aligned} \tag{14.49}$$

where

$$\bar{y} = \sum_{j=1}^{n} \frac{y_j}{n}, \quad \bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}, \quad i = 1, \dots, k$$

and $\hat{a}_0$, $\hat{a}_i$, $i = 1, \dots, k$ indicate the critical point $(\hat{a}_0, \dots, \hat{a}_k)$ or the point at which the equations hold. Differentiating with respect to $a_i$, $i = 1, \dots, k$, we have

$$\begin{aligned} &\frac{\partial}{\partial a_i}\left[\sum_{j=1}^{n} e_j^2\right] = 0 \quad \Rightarrow \quad -2\sum_{j=1}^{n} x_{ij}[y_j - \hat{a}_0 - \hat{a}_1 x_{1j} - \dots - \hat{a}_k x_{kj}] = 0 \\ \Rightarrow \quad &\sum_{j=1}^{n} x_{ij} y_j = \hat{a}_0 \sum_{j=1}^{n} x_{ij} + \hat{a}_1 \sum_{j=1}^{n} x_{ij} x_{1j} + \dots + \hat{a}_k \sum_{j=1}^{n} x_{ij} x_{kj}. \end{aligned} \tag{14.50}$$

Substituting the value of $\hat{a}_0$ from (14.49) into (14.50) and rearranging and then dividing by $n$, we have the following:

$$s_{iy} = \hat{a}_1 s_{1i} + \hat{a}_2 s_{2i} + \dots + \hat{a}_k s_{ki}, \quad i = 1, \dots, k \tag{14.51}$$

where

$$s_{ij} = \sum_{k=1}^{n} \frac{(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{n} = s_{ji},$$

$$s_{iy} = s_{yi} = \sum_{k=1}^{n} \frac{(y_k - \bar{y})(x_{ik} - \bar{x}_i)}{n}$$

or the corresponding sample variances and covariances. If we do not wish to substitute for $\hat{a}_0$ from (14.49) into (14.50), then we may solve (14.49) and (14.50) together to obtain a solution for $(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_k)$. But from (14.51) we get only $(\hat{a}_1, \dots, \hat{a}_k)$ and then this has to be used in (14.49) to obtain $\hat{a}_0$. From (14.51), we have the following matrix equation:

$$\begin{aligned} s_{1y} &= \hat{a}_1 s_{11} + \hat{a}_2 s_{12} + \cdots + \hat{a}_k s_{1k} \\ s_{2y} &= \hat{a}_1 s_{21} + \hat{a}_2 s_{22} + \cdots + \hat{a}_k s_{2k} \\ \vdots\ &\ \vdots \\ s_{ky} &= \hat{a}_1 s_{k1} + \hat{a}_2 s_{k2} + \cdots + \hat{a}_k s_{kk} \quad \text{or} \\ S_y &= S\hat{a} \end{aligned} \tag{14.52}$$

where

$$S_y = \begin{bmatrix} s_{1y} \\ \vdots \\ s_{ky} \end{bmatrix}, \quad \hat{a} = \begin{bmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_k \end{bmatrix}, \quad S = (s_{ij}),$$

$$s_{ij} = \sum_{k=1}^{n} \frac{(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{n} = s_{ji}.$$

From (14.52),

$$\hat{a} = \begin{bmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_k \end{bmatrix} = S^{-1} S_y, \quad \text{for } |S| \neq 0. \tag{14.53}$$

From (14.53),

$$\hat{a}_0 = \bar{y} - \hat{a}'\bar{x} = \bar{y} - S_y' S^{-1} \bar{x}, \quad \bar{x} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{bmatrix}. \tag{14.54}$$

**Note 14.13.** When observations on $(x_1, \dots, x_k)$ are involved, even if we take extreme care sometimes near singularity may occur in $S$. In general, one has to solve the system of linear equations in (14.52) for which many standard methods are available whether the coefficient matrix, in our case $S$, is non-singular or not. In a regression-type model, as in our case above, the points $(x_{1j}, x_{2j}, \dots, x_{kj})$, $j = 1, \dots, n$ are preassigned, and hence while preassigning, make sure that data points for $(x_1, \dots, x_k)$, which are linear functions of other points which are already included, are not taken as a new data point. If linear functions are taken, then this will result in $S$ being singular.

**Example 14.17.** In a feeding experiment on cows, it is suspected that the increase in weight $y$ has a linear regression on the amount of green fodder $x_1$ and the amount of marketed cattle feed $x_2$ consumed. The following observations are available; all observations on $x_1$ and $x_2$ are in kilograms and the observations on $y$ are in grams:

| $x_1$ | 1 | 1.5 | 2 | 1 | 2.5 | 1 |
|---|---|---|---|---|---|---|
| $x_2$ | 2 | 1.5 | 1 | 1.5 | 2 | 4 |
| $y$ | 5 | 5 | 6 | 4.5 | 7.5 | 8 |

Construct the estimating function and then estimate $y$ at the points (i) $(x_1,x_2) = (1,0),(1,3),(5,8)$.

**Solution 14.17.** As per our notation $n = 6$,

$$\bar{x}_1 = \frac{(1.0 + 1.5 + 2.0 + 1.0 + 2.5 + 1.0)}{6} = 1.5,$$
$$\bar{x}_2 = \frac{(2.0 + 1.5 + 1.0 + 1.5 + 2.0 + 4.0)}{6} = 2,$$
$$\bar{y} = \frac{(5.0 + 5.0 + 6.0 + 4.5 + 7.5 + 8.0)}{6} = 6.$$

Again, if we are using a computer or programmable calculator then regression problems are there in the computer and the results are instantly available by feeding in the data. For the calculations by hand, the following table will be handy:

| $y$ | $x_1$ | $x_2$ | $y-\bar{y}$ | $x_1-\bar{x}_1$ | $x_2-\bar{x}_2$ | $(y-\bar{y})(x_1-\bar{x}_1)$ |
|---|---|---|---|---|---|---|
| 5.0 | 1.0 | 2.0 | −1 | −0.5 | 0 | 0.5 |
| 5.0 | 1.5 | 1.5 | −1 | 0 | −0.5 | 0 |
| 6.0 | 2.0 | 1.0 | 0 | 0.5 | −1 | 0 |
| 4.5 | 1.0 | 1.5 | −1.5 | −0.5 | −0.5 | 0.75 |
| 7.5 | 2.5 | 2.0 | 1.5 | 1.0 | 0 | 1.5 |
| 8.0 | 1.0 | 4.0 | 2.0 | −0.5 | 2.0 | −1.0 |
| | | | | | | 1.75 |

| $(y-\bar{y})(x_2-\bar{x}_2)$ | $(x_1-\bar{x}_1)^2$ | $(x_2-\bar{x}_2)^2$ | $(x_1-\bar{x}_1)(x_2-\bar{x}_2)$ |
|---|---|---|---|
| 0 | 0.25 | 0 | 0 |
| 0.5 | 0 | 0.25 | 0 |
| 0 | 0.25 | 1.0 | −0.5 |
| 0.75 | 0.25 | 0.25 | 0.25 |
| 0 | 1.0 | 0 | 0 |
| 4.0 | 0.25 | 4.0 | −1 |
| 5.25 | 2.0 | 5.5 | −1.25 |

The equations corresponding to (14.52), without the dividing factor $n = 6$, are the following:

$$\begin{aligned} 1.75 &= 2\hat{a}_1 - 1.25\hat{a}_2 \\ 5.25 &= -1.25\hat{a}_1 + 5.5\hat{a}_2 \quad \Rightarrow \quad \hat{a}_1 \approx 1.72, \hat{a}_2 \approx 1.34. \end{aligned}$$

Then

$$\begin{aligned} \hat{a}_0 &= \bar{y} - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2 \\ &\approx 6 - (1.72)(1.5) - (1.34)(2) = 0.74. \end{aligned}$$

Hence the estimated function is given by

$$y = 0.74 + 1.725x_1 + 1.34x_2.$$

The estimated value of $y$ at $(x_1, x_2) = (1, 3)$ is $\hat{y} = 6.48$. The point $(x_1, x_2) = (5, 8)$ is too far out of the observational range, and hence we may estimate $y$ only if we are sure that the conditional expectation is linear for all possible $(x_1, x_2)$. If the regression is sure to hold for all $(x_1, x_2)$, then the estimated $y$ at $(x_1, x_2) = (5, 8)$ is

$$\hat{y} = 0.74 + (1.72)(5) + (1.34)(8) = 20.06.$$

For example,

$$\begin{aligned} &\text{at } (x_1, x_2) = (1, 2), \quad \hat{y} = 5.14; \quad \text{at } (x_1, x_2) = (1.5, 1.5), \quad \hat{y} = 5.33; \\ &\text{at } (x_1, x_2) = (2, 1), \quad \hat{y} = 5.52; \quad \text{at } (x_1, x_2) = (1, 4), \quad \hat{y} = 7.82. \end{aligned}$$

Hence we can construct the following table:

| $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|
| 5 | 5.14 | −0.14 | 0.096 |
| 5 | 5.33 | −0.33 | 0.1089 |
| 6 | 5.52 | 0.48 | 0.2304 |
| 4.5 | 4.4 | −0.03 | 0.0009 |
| 7.5 | 7.72 | −0.22 | 0.0484 |
| 8 | 7.82 | 0.18 | 0.0324 |
| | | | 0.4406 |

An estimate of the error sum of squares as well as the least square minimum is 0.4406 in this model.

## Exercises 14.7

**14.7.4.** If the yield $y$ of corn in a test plot is expected to be a linear function of $x_1$ = amount of water supplied, in addition to the normal rain and $x_2$ = amount of organic fertilizer (cow dung), in addition to the fertility of the soil. The following is the data available:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 2 | 1.5 | 2.5 | 3 |
| $x_2$ | 0 | 1 | 1 | 1.5 | 2 | 2 | 3 |
| $y$ | 2 | 2 | 5 | 8 | 7 | 9 | 10 |

(i) Fit a linear model $y = a_0 + a_1x_1 + a_2x_2$ by the method of least squares.

(ii) Estimate $y$ at the points

$$(x_1, x_2) = (90.5, 1.5), (3.5, 2.5).$$

(iii) Compute the least square minimum.

### 14.7.4 General linear model

If we use matrix notation, then the material in Section 14.7.3 can be simplified and can be written in a nice form. Consider a general linear model of the following type: Suppose that the real scalar variable $y$ is to be estimated by using a linear function of $x_1, \dots, x_n$. Then we may write the model as

$$y_j = a_0 + a_1x_{1j} + a_2x_{2j} + \cdots + a_px_{pj} + e_j, \quad j = 1, \dots, n. \tag{14.55}$$

This can be written as

$$Y = X\beta + e,$$
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}, \quad \beta = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}. \tag{14.56}$$

Then the error sum of squares is given by

$$e'e = (Y - X\beta)'(Y - X\beta). \tag{14.57}$$

Minimization by using vector derivative (see Note 14.7) gives

$$\frac{\partial}{\partial\beta}e'e = O \quad \Rightarrow \quad -2X'(Y - X\beta) = O \tag{14.58}$$

$$\Rightarrow \quad \beta = (X'X)^{-1}X'Y \quad \text{for } |X'X| \neq 0. \tag{14.59}$$

Since $X$ is under our control (these are preassigned values), we can assume $X'X$ to be non-singular in regression-type linear models. Such will not be the situation in design models where $X$ is determined by the design used. This aspect will be discussed in the next chapter. The least square minimum, again denoted by $s^2$, is given by

$$\begin{aligned} s^2 &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'(Y - X\hat{\beta}) \quad \text{due to (14.58)} \\ &= Y'Y - Y'X(X'X)^{-1}X'Y = Y'[I - X(X'X)^{-1}X']Y = Y'[I - B]Y \end{aligned}$$

where

$$B = X(X'X)^{-1}X' = B^2.$$

This shows that $B$ is idempotent and of

$$\begin{aligned}\text{rank} &= \operatorname{tr}[X(X'X)^{-1}X'] = \operatorname{tr}[(X'X)^{-1}X'X] \\ &= \operatorname{tr}[I_{p+1}] = p+1.\end{aligned} \tag{14.60}$$

Hence the rank of $I - X(X'X)^{-1}X'$ is $n - (p+1)$. Note further that $I - B = (I-B)^2$, $(I-B)B = O$ and hence from Section 10.5 of Chapter 10 we have $u = s^2 = Y'[I - X(X'X)^{-1}X']Y$ and $v = Y'X(X'X)^{-1}X'Y$ are independently distributed when $e \sim N_n(O, \sigma^2 I_n)$ or $Y \sim N_n(X\beta, \sigma^2 I_n)$. Further,

$$\begin{aligned}\frac{1}{\sigma^2}Y[I - X(X'X)^{-1}X']Y &= \frac{1}{\sigma^2}(Y - X\beta)[I - X(X'X)^{-1}X'](Y - X\beta) \\ &\sim \chi^2_{n-(p+1)}\end{aligned}$$

where $\chi^2_\nu$ denotes a central chi-square with $\nu$ degrees of freedom, and

$$\frac{1}{\sigma^2}Y'X(X'X)^{-1}X'Y \sim \chi^2_{p+1}(\lambda)$$

where $\chi^2_{p+1}(\lambda)$ is a non-central chi-square with $p+1$ degrees of freedom and non-centrality parameter $\lambda = \frac{1}{2}\beta'(X'X)\beta$. [See the discussion of non-central chi-square in Example 10.9 of Chapter 10.] Hence we can test the hypothesis that $\beta = O$ (a null vector), by using a $F$-statistic, under the hypothesis:

$$\begin{aligned}F_{p+1,n-(p+1)} &= \frac{v/(p+1)}{u/(n-(p+1))}, \\ v &= Y'X(X'X)^{-1}X'Y, \\ u &= Y'[I - X(X'X)^{-1}X']Y\end{aligned} \tag{14.61}$$

and we reject the hypothesis for large values of the observed $F$-statistic.

Note that the individual parameters are estimated by the equation

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{14.62}$$

the various column elements of the right side gives the individual estimates. What is the variance–covariance matrix of this vector of estimators $\hat{\beta}$? Let us denote the covariance matrix by $\operatorname{Cov}(\hat{\beta})$. Then from the definition of covariance matrix,

$$\begin{aligned}\operatorname{Cov}(\hat{\beta}) &= (X'X)^{-1}X' \operatorname{Cov}(Y)X(X'X)^{-1} = (X'X)^{-1}\sigma^2 I(X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}.\end{aligned} \tag{14.63}$$

How do we construct confidence interval for the parameter $a_j$ in $\beta$ and test hypotheses on $a_j$, $j = 0,1,\dots,p$? Let $\hat{a}_j$ be the $(j+1)$-th element in the right side column in (14.62) and let $b_{j+1,j+1}$ be the $(j+1,j+1)$-th diagonal element in $(X'X)^{-1}$, $j = 0,1,\dots,p$. Then

$$\frac{\hat{a}_j - a_j}{\hat{\sigma}^2 b_{j+1,j+1}} \sim t_{n-(p+1)}$$

or a Student-$t$ with $n-(p+1)$ degrees of freedom, where

$$\hat{\sigma}^2 = \frac{s^2}{n-(p+1)} = \frac{\text{least square minimum}}{n-(p+1)}. \tag{14.64}$$

Then use this Student-$t$ to construct confidence intervals and test hypotheses. Since it will take up too much space, we will not do a numerical example here.

Sometimes we may want to separate $a_0$ from the parameters $a_1,\dots,a_p$. In this case, we modify the model as

$$y_j - \bar{y} = a_1(x_{1j} - \bar{x}_1) + a_2(x_{2j} - \bar{x}_2) + \cdots + a_p(x_{pj} - \bar{x}_p) + e_j - \bar{e}. \tag{14.65}$$

Now, proceed exactly as before. Let the resulting matrices be denoted by $\tilde{Y}$, $\tilde{X}$, $\tilde{\beta}$, $\tilde{e}$ where

$$\tilde{Y} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \quad \tilde{\beta} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}, \quad \tilde{e} = \begin{bmatrix} e_1 - \bar{e} \\ \vdots \\ e_n - \bar{e} \end{bmatrix},$$

$$\tilde{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & \dots & x_{p1} - \bar{x}_p \\ \vdots & \vdots & \vdots \\ x_{1n} - \bar{x}_1 & \dots & x_{pn} - \bar{x}_p \end{bmatrix} \tag{14.66}$$

Then the estimator, covariance matrix of the estimator, least square minimum, etc. are given by the following, where the estimators are denoted by a star:

$$\begin{aligned} \tilde{\beta}_* &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \\ \mathrm{Cov}(\tilde{\beta}_*) &= \sigma^2(\tilde{X}'\tilde{X})^{-1} \\ \frac{s_*^2}{n-1-p} &= \hat{\sigma}^2 = \frac{\text{least square minimum}}{n-1-p} \\ &= \frac{\tilde{Y}'[I - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}']\tilde{Y}}{n-1-p}. \end{aligned}$$

Under normality assumption for $e \sim N(O, \sigma^2 I_n)$ we have

$$\begin{aligned} u &= \frac{s_*^2}{\sigma^2} \sim \chi^2_{n-1-p}, \\ v &= \frac{\tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}}{\sigma^2} \sim \chi^2_p(\lambda_1) \end{aligned}$$

where $u$ and $v$ are independently distributed, and the non-centrality parameter $\lambda_1 = \frac{1}{2}\tilde{\beta}'(\tilde{X}'\tilde{X})\tilde{\beta}$. Note that $u$ and $v$ are independently distributed.

**Note 14.14.** Models are classified as linear or non-linear depending upon the linearity or non-linearity of the parameters in the model. All linear models can be handled by the procedure in Section 14.7.4 by renaming the coefficients of the parameters. For example, (1) $y = a_0 + a_1x + a_2x^2 + \cdots + a_kx^k$ (write $x = u_1$, $x^2 = u_2$, ..., $x^k = u_k$) and apply the techniques in Section 14.7.4, (2) $y = a_0 + a_1x_1x_2 + a_2x_1^2 + a_3x_2^2$ (Write $u_1 = x_1x_2, u_2 = x_1^2, u_3 = x_2^2$), (3) $y = a_0 + a_1x_1x_2 + a_2x_2x_3 + a_3x_1^2x_2$ (write $u_1 = x_1x_2$, $u_2 = x_2x_3$, $u_3 = x_1^2x_2$), are all linear models, whereas (4) $y = ab^x$, (5) $y = 3^{a+bx}$ are non-linear models. There are non-linear least square techniques available for handling non-linear models. That area is known as non-linear least square analysis.

**Note 14.15.** In some books, the student may find a statement of the type, asking to take logarithms and use linear least square analysis for handling a model of the type $y = ab^x$. This is wrong unless the error $e$ is always positive and enters into the model as a product or unless the model is of the form $y_j = ab^{x_j}e_j$, with $e_j > 0$, which is a very unlikely scenario. We are dealing with real variables here and then the logarithm cannot be taken when $e_j$ is negative or zero. If $ab^x$ is taken to predict $y$, then the model should be constructed as $y_j = ab^{x_j} + e_j$, $j = 1, \dots, n$. Then the error sum of squares will become

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - ab^{x_j})^2 \tag{a}$$

and it will be difficult to handle this situation. The analytic solution will not be available for the normal equations coming out of this equation (a) here. There are several methods available for handling non-linear least square problems. A non-linear least square analysis is to be conducted for analyzing models such as $y = ab^x$. The most frequently used non-linear least square analysis technique is Marquardt's method. For a very efficient algorithm for non-linear least squares, which usually never fails, may be seen from [9].

## Exercises 14.7

**14.7.5.** For the linear model and data in Exercise 14.7.4, construct 95% confidence intervals for (1) $a_1$; (2) $a_2$, and test the hypotheses, at 5% level of rejection, (3) $H_0 : a_1 = 0$; (4) $H_0 : a_2 = 0$; (5) $H_0 : (a_1, a_2) = (0, 0)$.

# 15 Design of experiments and analysis of variance

## 15.1 Introduction

In Chapter 14, we have looked into regression and regression-type models in the area of model building. Here, we consider a slightly different type of model known as design type models. All the technical terms used in this area are connected with agricultural experiments because, originally, the whole area was developed for agricultural experimentation of crop yield, methods of planting, the effects of various types of factors on the yields, etc. The main technical terms are the following: *A plot* means an experimental unit. If different methods of teaching are compared and the students are subjected to various methods of teaching, then the basic unit on which the experiment is carried out is a student. Then a student in this case is a plot. If the breaking strength of an alloy is under study and if 15 units of the same alloy are being tested, then each unit of the alloy is a plot. If the gain in weight of experimental animals is under study, then a plot here is an experimental animal. If experimental plots of land are there where tapioca is planted and the experiment is conducted to study the effect of different types of fertilizers on the yield, then an experimental plot is a plot of land. The basic unit which is subjected to experimentation is called a plot. A group of such plots is called a *block* or block of plots. If a piece of land is divided into 10 plots and experimentation is done on these 10 plots, then this block contains 10 plots. If another piece of land is divided into 8 plots for experimentation, then that block contains 8 plots. The item or factor under study in the experimentation is called a *treatment*. In the case of students being subjected to 3 different methods of teaching, there are 3 treatments. In the case of 6 different fertilizers being studied with reference to yield of tapioca, then there are 6 treatments, etc. For a proper experimentation, all the plots must be homogeneous, within and between, as far as variations with respect to all factors are concerned, which are not under study. For example, if 3 different methods of teaching are to be compared, then the students selected for this study must have the same background, same exposure to the subject matter, or must be the same as far as all other factors are concerned, which may have some relevance to the performance of the students, performance may be measured by computing the grades obtained in an examination at the end of subjecting the student with a particular method of teaching.

Hence planning of an experiment means to make sure that all experimental plots are fully homogeneous within and between with respect to all other factors, other than the factors under study, which may have some effect on the experimental outcome. If one variety of corn is planted on 10 different plots of land, where the plots have different natural soil fertility level, different water drainage, different exposure to sun, etc., then the plots are not homogeneous within or between them. If we are trying to study the effect of different fertilizers on the yield of corn, then

https://doi.org/10.1515/9783110562545-015

the pieces of land (experimental plots) taken must be of the same type and fully homogeneous with respect to all factors of variation other than the effect of fertilizers used.

## 15.2 Fully randomized experiments

In this experiment, suppose that we have selected $n_1$ plots to try the first treatment, $n_2$ plots to try the second treatment, ..., $n_k$ plots to try the $k$-th treatment, then all $n_1 + \cdots + n_k = n$ plots must be fully homogeneous within and between and with respect to all factors of variation. If it is an agricultural experiment involving the study of 5 different method of planting of one type of rubber trees, then the plots of land selected must be of the same shape and size, of the same basic fertility of the soil, of the same type of elevation, same type of drainage, same type of precipitation, etc., or in short, identical with respect to all known factors which may have some effect on the yield of rubber latex. It may not be possible to find such $n$ plots of land at one place but we may get identical plots at different locations, if not at the same place. There should not be any effect of the location on the yield. Suppose we have $n_1$ plots at one place, $n_2$ plots at another place, ..., $n_k$ plots at the $k$-th place but all the $n = n_1 + \cdots + n_k$ plots are identical in every respect. Then take one of the methods at random and subject the $n_1$ plots to this method, take another method at random, etc. or assign the methods at random to the $k$ sets of plots. If it is an experiment involving 4 different methods of teaching and if all $n$ homogeneous students can be found in one school, then divide them into different groups according to convenience of the teachers and subject them to these 4 different methods of teaching $n_1 + \cdots + n_4 = n$. It is not necessary to divide them into groups of different numbers. If equal numbers can be grouped, then it is well and good. In most of the situations, we may find 50 students in one school with the same background, 30 students with the same background within the group as well as between the groups in another school, etc., and thus the numbers in the groups may be different. The 50 students may be subjected to one method of teaching, the 30 another method of teaching, etc. Let $x_{ij}$ be the grade obtained by the $j$-th student ($j$-th plot) under the $i$-th method of teaching ($i$-th treatment). Then the possibility is that $x_{ij}$ may contain a general effect, call it $\mu$, an effect due to the $i$-th method of teaching, call it $\alpha_i$. The general effect can be given the following interpretation. Suppose that the student is given a test without subjecting the student to any particular method of teaching. We cannot expect the student to get a zero grade. There will be some general effect. Then $\alpha_i$ can be interpreted as the deviation from the general effect due to the $i$-th treatment. In the experimentation, we have only controlled all known factors of variation. But there may be still unknown factors which may be contributing towards $x_{ij}$. The sum total effect of all unknown factors is known as the random effect $e_{ij}$. Thus, $x_{ij}$ in this fully randomized experiment is a function of $\mu$, $\alpha_i$ and $e_{ij}$. What is the functional form or which way these effects enter into $x_{ij}$? The simplest model that we can come up with is a simple linear additive model or we assume that

$$x_{ij} = \mu + \alpha_i + e_{ij}, \quad j = 1, \dots, n_i,\ i = 1, \dots, k,\ e_{ij} \sim N(0, \sigma^2). \tag{15.1}$$

That is, for $i = 1$,

$$\begin{aligned} x_{11} &= \mu + \alpha_1 + e_{11} \\ x_{12} &= \mu + \alpha_1 + e_{12} \\ \vdots &= \vdots \\ x_{1n_1} &= \mu + \alpha_1 + e_{1n_1} \end{aligned}$$

and similar equations for $i = 2, \dots, k$. If the $\alpha_i$'s are assumed to be some unknown constants, then the model in (15.1) is called a *simple additive fixed effect one-way classification model*. Here, one-way classification means that only one set of treatments are studied here, one set of methods of teaching, one set of fertilizers, one set of varieties of corn, etc. There is a possibility that $\alpha_i$'s could be random variables then the model will be a *random effect* model. We will start with a fixed effect model.

The first step in the analysis of any model is to estimate the effects and then try to test some hypotheses by putting some assumptions on the random part $e_{ij}$'s. For estimating $\mu$, $\alpha_i$, $i = 1, \dots, k$, we will use the method of least squares because we do not have any assumption of any distribution on $e_{ij}$'s or $x_{ij}$'s. The error sum of squares is given by the following, where we can use calculus for minimization, observing that the maximum is at $+\infty$ and hence the critical point will correspond to a minimum.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i,j}(x_{ij} - \mu - \alpha_i)^2 \tag{a}$$

$$\frac{\partial}{\partial \mu}\sum_{ij} e_{ij}^2 = 0 \quad \Rightarrow \quad -2\sum_{ij}(x_{ij} - \hat{\mu} - \hat{\alpha}_i) = 0 \tag{b}$$

$$\Rightarrow \quad x_{..} - n_{.}\hat{\mu} - \sum_{i=1}^{k} n_i\hat{\alpha}_i = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{x_{..}}{n_{.}} - \frac{\sum_{i=1}^{k} n_i\hat{\alpha}_i}{n_{.}} \tag{c}$$

because when we sum up with respect to $j$ we get $n_i$ and then summation with respect to $i$ is denoted by $n_{.}$. Also the standard notation $\sum_j x_{ij} = x_{i.}$, $\sum_i x_{ij} = x_{.j}$, when $j$ is free of $i$, will be used, where the summation with respect to a subscript is denoted by a dot. Hence $x_{..} = \sum_{ij} x_{ij}$. Differentiation of (a) with respect to $\alpha_i$ for a specific $i$ such as $\alpha_1$ will yield the following:

$$\begin{aligned} \frac{\partial}{\partial \alpha_i}\sum_{ij} e_{ij}^2 = 0 \quad &\Rightarrow \quad \sum_{j=1}^{n_i}(x_{ij} - \hat{\mu} - \hat{\alpha}_i) = 0 \\ &\Rightarrow \quad \hat{\alpha}_i = \frac{x_{i.}}{n_i} - \hat{\mu}. \end{aligned} \tag{d}$$

Note that in (c), without loss of generality, $\sum_i n_i \alpha_i$ can be taken as zero because our $\alpha_i$'s are defined as deviations from the general effect due to the $i$-th treatment, then the sum of the deviations is zero, because for any set of numbers $y_1, \dots, y_n$, $(y_1 - \bar{y}) + \dots + (y_n - \bar{y}) = 0$. Our analysis in equations (a) to (d) will go through even if we do not wish to use this condition. The least square minimum, denoted by $s^2$, is available from (a) by substituting the least square estimates $\hat{\mu}$ and $\hat{\alpha}_i$. That is,

$$\begin{aligned} s^2 &= \sum_{ij} (x_{ij} - \hat{\mu} - \hat{\alpha}_i)^2 = \sum_{ij} \left(x_{ij} - \hat{\mu} - \left[\frac{x_{i.}}{n_i} - \hat{\mu}\right]\right)^2 \\ &= \sum_{ij} \left(x_{ij} - \frac{x_{i.}}{n_i}\right)^2 \qquad (15.2) \\ &= \sum_{ij} \left(x_{ij} - \frac{x_{..}}{n_.}\right)^2 - \sum_i \left(\frac{x_{i.}}{n_i} - \frac{x_{..}}{n_.}\right)^2 \qquad (15.3) \\ &= \left(\sum_{ij} x_{ij}^2 - \frac{x_{..}^2}{n_.}\right) - \left(\sum_i \frac{x_{i.}^2}{n_i} - \frac{x_{..}^2}{n_.}\right). \qquad (15.4) \end{aligned}$$

This $s^2$ is called the residual sum of squares. All the different representations in (15.2) to (15.4) will be made use of later. The derivations are left to the student. If we have a hypothesis of the type $H_0 : \alpha_1 = 0 = \alpha_2 = \dots = \alpha_k$, then under this $H_0$ the model becomes $x_{ij} = \mu + e_{ij}$ and if we proceed as before then the least square minimum, denoted by $s_0^2$, is given by the following:

$$s_0^2 = \sum_{ij} \left(x_{ij} - \frac{x_{..}}{n_.}\right)^2 = \sum_{ij} x_{ij}^2 - \frac{x_{..}^2}{n_.}. \qquad (15.5)$$

Then

$$s_0^2 - s^2 = \sum_{ij} \left(\frac{x_{i.}}{n_i} - \frac{x_{..}}{n_.}\right)^2 = \sum_i \frac{x_{i.}^2}{n_i} - \frac{x_{..}^2}{n_.}. \qquad (15.6)$$

can be called the sum of squares due to the hypothesis or the sum of squares due to the $\alpha_i$'s. Thus we have the following identity:

$$\begin{aligned} s_0^2 &\equiv [s_0^2 - s^2] + [s^2] \\ &= \text{sum of squares due to the treatments} + \text{residual sum of squares} \\ &= \text{between treatment sum of squares} \\ &\quad + \text{within treatment sum of squares} \end{aligned}$$

**Definition 15.1** (Analysis of variance principle)**.** The principle of splitting the total variation in the data into the sum of variations due to different components is known as the analysis of the variance principle or the ANOVA principle.

Since we are not dividing by the sample sizes and making it per unit variation or we are not taking sample variances, the principle is more appropriately called the analysis of variation principle. For a one-way classification model as in (15.1), there is only one component of variation, namely one set of treatments. More elaborate designs will have more components of variation.

In order to test hypotheses of the type $H_0 : \alpha_1 = \cdots = \alpha_k = 0$, which is the same as saying $\mu_1 = \cdots = \mu_k$ where $\mu_j = \mu + \alpha_j$, we will make use of the two results from Chapter 10, namely Result 10.14 on the chi-squaredness of quadratic form and Result 10.15 on the independence of two quadratic forms in standard normal variables. The chi-squaredness, in effect, says that if the $m \times 1$ vector $Y$ has a $m$-variate normal distribution $Y \sim N_m(O, \sigma^2 I_m)$, where $\sigma^2$ is a scalar quantity and $I_m$ is the identity matrix of order $m$, then $\frac{1}{\sigma^2} Y'AY$, $A = A'$, is a chi-square with $\nu$ degrees of freedom if and only if $A$ is idempotent and of rank $\nu$. Result 10.15 says that two such quadratic forms $Y'AY$ and $Y'BY$ are independently distributed if and only if $AB = O$, where $O$ is a null matrix. We will make use of these two results throughout the whole discussion of Design of Experiments and Analysis of Variance. Derivations of each item will take up too much space. One item will be illustrated here and the rest of the derivations are left to the student. For illustrative purposes, let us consider

$$\begin{aligned} s^2 &= \sum_{ij}\left(x_{ij} - \frac{x_{i.}}{n_i}\right)^2 \\ &= \sum_{ij}\left([\mu + \alpha_i + e_{ij}] - \left[\mu + \alpha_i + \frac{e_{i.}}{n_i}\right]\right)^2 = \sum_{ij}\left(e_{ij} - \frac{e_{i.}}{n_i}\right)^2 \end{aligned}$$

If we write $x_{(1)}$ and $e_{(1)}$ for the subvectors,

$$x_{(1)} = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n_1} \end{bmatrix}, \quad e_{(1)} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n_1} \end{bmatrix} \quad \text{then} \quad \begin{bmatrix} e_{11} - \frac{e_{1.}}{n_1} \\ \vdots \\ e_{1n_1} - \frac{e_{1.}}{n_1} \end{bmatrix} \tag{15.7}$$

which can be written in matrix notation as $(I_{n_1} - B_1)e_{(1)}$ where $B_1 = \frac{1}{n_1} J_1 J_1'$ with $J_1' = (1, 1, \ldots, 1)$. We note that $B_1 = B_1^2$ and hence $B_1$ is idempotent. Further, $(I - B_1)^2 = (I - B_1)$ or $I - B_1$ is also idempotent. Then we can write

$$\sum_{ij}\left(x_{ij} - \frac{x_{i.}}{n_i}\right)^2 = e'(I - B)e$$

where $e' = (e_{11}, \ldots, e_{1n_1}, e_{21}, \ldots, e_{2n_2}, \ldots, e_{k1}, \ldots, e_{kn_k})$ and $B$ will be a block diagonal matrix with the diagonal blocks being the matrices $B_1, \ldots, B_k$ where $B_i = \frac{1}{n_i} J_i J_i'$. Further, we note that $I - B$ is idempotent and of rank $n_. - k$. Therefore, from Result 10.13,

$$\frac{s^2}{\sigma^2} \sim \chi^2_{n_.-k} \tag{15.8}$$

that is, the least square minimum divided by $\sigma^2$ is a chi-square with $n_. - k$ degrees of freedom. In a similar fashion, we can show that

$$\begin{aligned} s_0^2 &= \sum_{ij}\left(x_{ij} - \frac{x_{..}}{n_.}\right)^2 \sim \sigma^2\chi^2_{n_.-1} \quad \text{under } H_0 \\ s_0^2 - s^2 &= \sum_{ij}\left(\frac{x_{i.}}{n_i} - \frac{x_{..}}{n_.}\right)^2 \sim \sigma^2\chi^2_{k-1} \quad \text{under } H_0 \\ s^2 &= \sum_{ij}\left(x_{ij} - \frac{x_{i.}}{n_i}\right)^2 \sim \sigma^2\chi^2_{n.-k} \end{aligned} \tag{15.9}$$

and that $s^2$ and $(s_0^2 - s^2)$ are independently distributed. Here, the decomposition is of the following form:

$$\chi^2_{n_.-1} \equiv \chi^2_{k-1} + \chi^2_{n_.-k}. \tag{15.10}$$

From (15.8), (15.9) and Result 10.15, it follows that

$$\frac{(s_0^2 - s^2)/(k-1)}{s^2/(n.-k)} \sim F_{k-1,n.-k} \quad \text{under } H_0. \tag{15.11}$$

[If $H_0$ is not true then the left side of (15.11) will be a non-central $F$ with the numerator chi-square being non-central.] The test criterion will be to reject for large values of this $F$-statistic or reject $H_0$, at the level $\alpha$, if the observed value of $F_{k-1,n_.-k} \geq F_{k-1,n_.-k,\alpha}$.

### 15.2.1 One-way classification model as a general linear model

The model in (15.1), which is a linear fixed effect one-way classification model, can be put in matrix notation as a general linear model of the type $Y = X\beta + e$ of Chapter 14. Here, $Y$ is the $n_. \times 1 = (n_1 + \cdots + n_k) \times 1$ vector of the observations $x_{ij}$'s or $Y' = (x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{2n_2}, \dots, x_{k1}, \dots, x_{kn_k})$, $e$ is the corresponding vector of $e_{ij}$'s. $\beta$ is the $(k+1) \times 1$ vector of parameters or $\beta' = (\mu, \alpha_1, \dots, \alpha_k)$. Here, $X$ is the design matrix. It is $n_. \times (k+1)$ or $(n_1 + \cdots + n_k) \times (k+1)$ matrix with the first column all ones, the second column is all ones for the first $n_1$ rows only, the third column is all ones from $(n_1+1)$-th row to $(n_1+n_2)$-th row, and so on. In other words, the sum of the second to $(k+1)$-th column is equal to the first column. If we delete the first column, then all the remaining columns are linearly independent and thus the column rank of $X$ is $k$. But $n_. \geq (k+1)$, and hence the rank of the design matrix in this case is $k$, and thus $X$ is a *less than full rank* matrix and therefore $X'X$ is singular. If we use the notations of (15.7) then the model in (15.1) can be written as follows, as a general linear model $Y = X\beta + e$, where.

$$Y = \begin{bmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(k)} \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix}, \quad e = \begin{bmatrix} e_{(1)} \\ e_{(2)} \\ \vdots \\ e_{(k)} \end{bmatrix},$$

$$X = \begin{bmatrix} J_1 & J_1 & O & O & \dots & O \\ J_2 & O & J_2 & O & \dots & O \\ J_3 & O & O & J_3 & \dots & O \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ J_k & O & O & O & \dots & J_k \end{bmatrix} \tag{15.12}$$

where $J_m$ is an $m \times 1$ column vector of ones, $m = n_1, n_2, \dots, n_k$, and $O$ denotes a null matrix. In this general linear model if we wish to estimate the parameter vector, then the minimization of the error sum of squares $e'e$ leads to the normal equation:

$$X'X\hat{\beta} = X'Y \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-}X'Y. \tag{15.13}$$

In the light of the discussion in (15.12), note that (15.13) is a singular system of normal equations. Hence there is no unique solution since $(X'X)^{-1}$ does not exist. A solution can be written in terms of a g-inverse $(X'X)^{-}$ of $X'X$ as indicated in (15.13). Thus, if we use matrix-methods in analyzing a one-way classification model, or other design models, then the procedure will be complicated. It will be seen that the procedure adopted in Section 15.2 of separating the sum of squares is the simplest method, and we will be using the same type of procedures in other design models also.

### 15.2.2 Analysis of variance table or ANOVA table

In data analysis connected with Design of Experiments, usually the final analysis is put in a nice tabular format, known as the *Analysis of Variance Table or the ANOVA Table*. For a one-way classification linear fixed effect model, which is applicable in a completely randomized experiment, the following is the format of the ANOVA table, where d.f = degrees of freedom, S.S = sum of squares, M.S = mean squares.

#### ANOVA table for a one-way classification

| Variation due to | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) = (3)/(2) | (5) |
| Between treatments | $k-1$ | $\sum_i \frac{x_{i.}^2}{n_i} - \text{C.F}$ | $T$ | $T/E \sim (F_{k-1,n_{.}-k})$ |
| Within treatments | $n_{.} - k$ | (subtract) | $E$ | |
| Total | $n_{.} - 1$ | $\sum_{ij} x_{ij} - \text{C.F}$ | | |

where C.F stands for “correction factor”, which is C.F $= x_{..}^2/n_{.}$. In this ANOVA table, there is a (6)-th column called “Inference”. Due to lack of space, this column is not listed above. In this column, write “significant” or “not significant” as the case may be. Here, “significant” means that the observed $F$-value is significantly high or we reject the null hypothesis of the effects being zero, or in the one-way case the hypothesis is that $\alpha_i$'s are zeros or the treatment effect are zeros and this hypothesis is rejected. Otherwise write "not significant". The residual sum of squares can be obtained by subtraction of the sum of squares due to treatments, namely, $s_0^2 - s^2 = \sum_i \frac{x_{i.}^2}{n_i} - \text{C.F}$ from the total sum of squares $\sum_{ij} x_{ij}^2 - \text{C.F}$. Similarly, the degrees of freedom corresponding to the residual sum of squares $s^2$ is available from total degrees of freedom, namely $n_. - 1$, minus the degrees of freedom for the treatment sum of squares, namely $k - 1$, which gives $(n_. - 1) - (k - 1) = n_. - k$.

**Example 15.1.** The following table gives the yield of wheat per test plot under three different fertilizers. These fertilizers are denoted by $A, B, C$.

Yield of wheat under fertilizers $A, B, C$

| | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $A$ | 50 | 60 | 60 | 65 | 70 | 80 | 75 | 80 | 85 | 75 | 700 |
| $B$ | 60 | 60 | 65 | 70 | 75 | 80 | 70 | 75 | 85 | 80 | 720 |
| $C$ | 40 | 50 | 50 | 60 | 60 | 60 | 65 | 75 | 70 | 70 | 600 |

Assume that a one-way classification fixed effect model is appropriate. Test the hypothesis, at a 5% level of rejection, that the fertilizer effects are the same, assuming $e_{ij} \sim N(0, \sigma^2)$ and mutually independently distributed.

**Solution 15.1.** Let $x_{ij}$ be the $j$-th observation under the $i$-th fertilizer, $i = 1, 2, 3$ and here all the sample sizes are equal to 10, and hence $j = 1, \dots, 10$.

$$\sum_{ij} x_{ij} = x_{..} = 600 + 720 + 700 = 2\,020;$$

$$\text{C.F} = \frac{x_{..}^2}{n_.} = \frac{(2\,020)^2}{30} = 136\,013.33;$$

$$\sum_i \frac{x_{i.}^2}{n_i} = \frac{1}{10}[600^2 + 720^2 + 700^2] = 136\,840;$$

$$\sum_{ij} x_{ij}^2 - \text{C.F} = 50^2 + \cdots + 70^2 - \text{C.F} = 3\,636.67;$$

$$\sum_i \frac{x_{i.}^2}{n_i} - \text{C.F} = 827.67.$$

Now, we can set up the ANOVA table

| Variation due to | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| Between fertilizers | $k - 1 = 2$ | 826.67 | 413.33 | $\frac{413.33}{104.08} > 3.25$ |
| Within fertilizers | 27 | 2 810.00 | 104.08 | |
| Total | $n_. - 1 = 29$ | 3 636.67 | | |

The tabulated point $F_{2,27,0.05} = 3.25$ and our observed $F_{2,27} > 3.25$, and hence we reject the hypothesis that the effects of the fertilizers are equal. In the column on inference, which is not listed above, we will write as "significant" or the $F$-value is significantly high.

### 15.2.3 Analysis of individual differences

If the hypotheses of no effect of the treatments is rejected, that is, if the observed $F$-value is significantly high, then there is a possibility that this high value may be contributed by some of the differences $\alpha_i - \alpha_j$, $i \neq j$ being not equal to zero or some of the individual differences may not be zeros. If the hypothesis of no effect is not rejected, then we stop the analysis here and we do not proceed further. We may proceed further only when the hypothesis is rejected or when the $F$-value is found to be significantly high. Individual hypotheses of the types $H_0 : \alpha_i - \alpha_j = 0$ for $i \neq j$ can be tested by using a Student-$t$ test when we assume that $e_{ij} \sim N(0, \sigma^2)$ for all $i$ and $j$ and mutually independently distributed. Note that the least square estimate is

$$\hat{\alpha}_i - \hat{\alpha}_j = \frac{x_{i.}}{n_i} - \frac{x_{j.}}{n_j}$$

with variance

$$\mathrm{Var}(\hat{\alpha}_i - \hat{\alpha}_j) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

and under the hypothesis $H_0 : \alpha_i - \alpha_j = \delta$ (given)

$$\frac{\hat{\alpha}_i - \hat{\alpha}_j - \delta}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{n_.-k}$$

where

$$\hat{\sigma}^2 = \frac{\text{Least square minimum}}{n_. - k} = \frac{s^2}{n_. - k}$$

Hence the test criterion is the following: Reject $H_0 : \alpha_i - \alpha_j = \delta$ (given), if the observed value of

$$\left|\frac{\frac{x_{i.}}{n_i} - \frac{x_{j.}}{n_j} - \delta}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}\right| \geq t_{n.-k,\frac{\alpha}{2}}$$

where $\Pr\{t_{n.-k} \geq t_{n.-k,\frac{\alpha}{2}}\} = \frac{\alpha}{2}$. A $100(1-\alpha)\%$ confidence interval for $\alpha_i - \alpha_j$ is then

$$\left(\frac{x_{i.}}{n_i}-\frac{x_{j.}}{n_j}\right)\mp t_{n.-k,\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}.$$

**Note 15.1.** In order to see which difference or for which $i$ and $j$, $\alpha_i-\alpha_j$ is contributing towards the significant sum of squares due to the $\alpha_j$'s, we should be considering all differences $\hat{\alpha}_i-\hat{\alpha}_j$. Hence a practical procedure is the following: Take the largest absolute difference $|\frac{x_{i.}}{n_i}-\frac{x_{j.}}{n_j}|$, then take the next largest difference, and so on, and test the corresponding hypotheses $\alpha_i-\alpha_j=0$ until the difference is found to be not significant and then stop. If $\hat{\alpha}_r-\hat{\alpha}_t$ is found to be significant or the hypothesis $\alpha_r-\alpha_t=0$ is rejected, then an estimate of $\alpha_r-\alpha_t$ is $\frac{x_{r.}}{n_r}-\frac{x_{t.}}{n_t}$. By using the same procedure one can test hypotheses and construct confidence intervals on linear functions $c_1\alpha_1+\cdots+c_k\alpha_k$, for specific $c_1,\dots,c_k$. Then take $c_j$'s such that $c_1+\cdots+c_k=0$ so that the contribution from $\mu$, the general effect, is canceled. Such linear functions are often called *cosets*.

**Example 15.2.** In Example 15.1, test the hypotheses on individual differences or hypotheses of the type $H_0:\alpha_i-\alpha_j=0$ and see which differences are contributing towards significant contribution due to the $\alpha_j$'s. Test at a 5% level of rejection.

**Solution 15.2.** The individual differences to be considered are $\alpha_1-\alpha_2$, $\alpha_1-\alpha_3$, $\alpha_2-\alpha_3$. Consider the following computations:

$$\hat{\sigma}^2=\frac{s^2}{n_.-k}=\frac{2\,810}{27};\quad \sqrt{\frac{1}{n_i}+\frac{1}{n_j}}=\sqrt{\frac{1}{5}}$$
$$\frac{x_{1.}}{n_1}=\frac{700}{10}=70;\quad \frac{x_{2.}}{n_2}=\frac{720}{10}=72;\quad \frac{x_{3.}}{n_3}=\frac{600}{10}=60;$$
$$t_{27,0.025}=2.052;\quad \hat{\sigma}\sqrt{\frac{1}{n_2}+\frac{1}{n_3}}=4.56.$$

The largest absolute difference between estimates is $\hat{\alpha}_2-\hat{\alpha}_3=72-60=12$. Hence $\frac{12}{4.56}=2.63>2.052$. This hypothesis is rejected. $\hat{\alpha}_1-\hat{\alpha}_3=10$. $\frac{10}{4.56}=2.19>2.052$. This is also rejected but $H_0:\alpha_1-\alpha_2=0$ is not rejected. Hence the differences $\alpha_1-\alpha_3$ and $\alpha_2-\alpha_3$ are contributing significantly towards the treatment sum of squares.

## Exercises 15.2

**15.2.1.** Under the assumption that the errors in the one-way classification fixed effect model $e_{ij}\sim N(0,\sigma^2)$, are mutually independently distributed prove, by examining the corresponding quadratic forms or otherwise, that

$$s_0^2\sim\sigma^2\chi^2_{n_.-1}\quad\text{under }H_0\tag{i}$$
$$s_0^2-s^2\sim\sigma^2\chi^2_{k-1}\quad\text{under }H_0\tag{ii}$$

and that $s^2$ and $s_0^2-s^2$ are independently distributed.

**15.2.2.** Show that in Exercise 15.2.1, for $s^2$ to be $\sigma^2\chi^2_{n.-k}$ the null hypothesis need not hold or for the chi-squaredness of the least square minimum divided by $\sigma^2$ no hypotheses need to hold, and that $s_0^2 - s^2$, divided by $\sigma^2$, will be a non-central chi-square in general and a central chi-square when $H_0$ holds.

**15.2.3.** Write the model in (15.1) as a general linear model of the form $Y = X\beta + e$ and show that the rank of the design matrix $X$ is $k$, thereby $X'X$ is singular.

**15.2.4.** Set up the ANOVA table if the group sizes are equal to $m$ and if there are $k$ groups in a completely randomized experiment.

**15.2.5.** Analyze fully the following one-way classification fixed effect data, which means to test the hypothesis that the treatment effects are equal and if this hypothesis is rejected, then test for individual differences and set up confidence intervals for the individual differences. Test at 5% level of rejection and set up 95% confidence intervals. Data:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group 1 | 10 | 12 | 15 | 10 | 25 | 13 | | |
| Group 2 | 20 | 25 | 32 | 33 | 28 | 34 | 30 | 32 |
| Group 3 | 5 | 8 | 2 | 4 | 6 | 8 | 4 | |

## 15.3 Randomized block design and two-way classifications

If it is an agricultural experiment for checking the effectiveness of 10 different fertilizers on the yield of sunflower seeds, then it may be very difficult to get $n_1 + \cdots + n_{10} = n_{.}$ identical experimental plots. A few plots may be available in one locality where all the plots are homogeneous within the plots and between the plots. There may be a few other plots available in a second locality but between these two localities there may be differences due to the difference in the fertility of the soil in these two localities. Then we have two blocks of plots which are fully homogeneous within each block but there may be differences between blocks. Then there will be the treatment effect due to the fertilizers and a block effect due to the different blocking of experimental plots. If it is an experiment involving method of teaching, then it is usually difficult to come up with a large number of students having exactly the same backgrounds and intellectual capacities. In one school, the students of the same class may have the same background but if we take students from two different schools, then there may be differences in their backgrounds. Here, the schools will act as blocks. In the above two cases, we have two different types of effects, one due to the treatments and the other due to the blocks. If $m$ blocks of $n$ plots each are taken, where the plots are homogeneous within each block and if the $n$ treatments are assigned at random to these $n$ plots in each block then such an experiment is called *a randomized block experiment*. If $x_{ij}$ is the observation on the $j$-th treatment from the $i$-th block, then we may write the simplest model in the following format:

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \dots, m,\ j = 1, \dots, n,\ e_{ij} \sim N(0, \sigma^2) \tag{15.14}$$

where for $i = 1$ we have

$$\begin{aligned} x_{11} &= \mu + \alpha_1 + \beta_1 + e_{11} \\ x_{12} &= \mu + \alpha_1 + \beta_2 + e_{12} \\ \vdots\ &\quad \vdots \\ x_{1n} &= \mu + \alpha_1 + \beta_n + e_{1n} \end{aligned}$$

and similar equations for $i = 2, \dots, m$, where $\mu$ is a general effect, $\alpha_i$ is the deviation from the general effect due to the $i$-th block and $\beta_j$ being the deviation from the general effect due to the $j$-th treatment. The model in (15.14) is called the linear, fixed effect, two-way classification model without interaction, with one observation per cell. Here, "fixed effect" means that the $\alpha_i$'s and $\beta_j$'s are taken as fixed unknown quantities and not as random quantities. The word "interaction" will be explained with a simple example. In a drug testing experiment, suppose that 5 different drugs are tried on 4 different age group of patients. Here, the age groups are acting as blocks and the drugs as treatment. The effect of the drug may be different with different age groups. In other words, if $\beta_j$ is the effect of the $j$-th drug, then $\beta_j$ may vary with the age group. There is a possibility of an effect due to the combination of the $i$-th block and $j$-th treatment, something like $\gamma_{ij}$, an effect depending on $i$ and $j$. If a joint effect is possible then the model will change to the following:

$$x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}, \quad e_{ijk} \sim N(0, \sigma^2). \tag{15.15}$$

Since both $\gamma_{ij}$ and $e_{ij}$ have both the subscripts $i$ and $j$, and since all quantities on the right are unknown, there is no way of estimating the joint effect $\gamma_{ij}$ because it cannot be separated from $e_{ij}$. Such joint effects are called *interactions*. If the interaction is to be estimated, then the experiment has to be repeated a number of times, say, $r$ times. In this case, we say that the experiment is *replicated $r$ times*. In that case, the $k$-th observation in the $i$-th block corresponding to the $j$-th treatment can be denoted by $x_{ijk}$ or the model can be written as

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad i = 1, \dots, m,\ j = 1, \dots, n,\ k = 1, \dots, r,\ e_{ijk} \sim N(0, \sigma^2). \tag{15.16}$$

In this case, we can estimate $\gamma_{ij}$ and test hypotheses on the interaction $\gamma_{ij}$ also. The model in (15.16) is called the *two-way classification model with interaction* and (15.14) is the two-way classification model without interaction. A randomized block experiment is conducted in such a way that there is no possibility of interaction between blocks and treatments so that the model in (15.14) is appropriate. When there is possibility of interaction, then the experiment has to be replicated so that one can use the model in (15.16) and complete the analysis. First, we will start with the model in (15.14).

### 15.3.1 Two-way classification model without interaction

Consider a randomized block experiment where there is no possibility of interaction between blocks and treatments so that one can use the model in (15.14), which is an *additive, fixed effect, two-way classification model without interaction*. Suppose that we have one observation per cell or one observation corresponding to each combination of *i* and *j*. For estimating the parameters, we will use the method of least squares. That is, we consider the error sum of squares

$$\sum_{ij} e_{ij}^2 = \sum_{ij} (x_{ij} - \mu - \alpha_i - \beta_j)^2$$

differentiate with respect to $\mu$, $\alpha_i$, $\beta_j$, equate to zero and solve. [This part is left as an exercise to the student.] We get the estimates as follows:

$$\begin{aligned} \hat{\alpha}_i &= \frac{x_{i.}}{n} - \hat{\mu}, \quad \hat{\beta}_j = \frac{x_{.j}}{m} - \hat{\mu} \\ \hat{\mu} &= \frac{x_{..}}{mn} - n\alpha_{.} - m\beta_{.} = \frac{x_{..}}{mn}. \end{aligned} \tag{15.17}$$

Since we have defined $\alpha_i$ as the deviation from the general effect due to the $i$-th treatment, without loss of generality, we can take $\alpha_{.} = \alpha_1 + \cdots + \alpha_m = 0$, and similarly $\beta_{.} = 0$ so that $\hat{\mu} = \frac{x_{..}}{mn}$. The least square minimum, denoted by $s^2$, is given by the following where $\mathrm{C.F} = \frac{x_{..}^2}{mn}$:

$$\begin{aligned} s^2 &= \sum_{ij} (x_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 \\ &= \sum_{ij} \left( x_{ij} - \frac{x_{..}}{mn} \right)^2 - \sum_{ij} \left( \frac{x_{i.}}{n} - \frac{x_{..}}{mn} \right)^2 - \sum_{ij} \left( \frac{x_{.j}}{m} - \frac{x_{..}}{mn} \right)^2 \\ &= \left[ \sum_{ij} x_{ij}^2 - \mathrm{C.F} \right] - \left[ \sum_{i} \frac{x_{i.}^2}{n} - \mathrm{C.F} \right] - \left[ \sum_{j} \frac{x_{.j}^2}{m} - \mathrm{C.F} \right]. \end{aligned} \tag{15.18}$$

The simplifications in (15.18) are given as exercises to the student. If we put the hypothesis that $\alpha_1 = 0 = \cdots = \alpha_m$, then the least square minimum, denoted by $s_0^2$ will be the same $s^2$ as in (15.18), excluding the term

$$\sum_{ij} \left( \frac{x_{i.}}{n} - \frac{x_{..}}{mn} \right)^2 = \sum_{i} \frac{x_{i.}^2}{n} - \mathrm{C.F.} \tag{15.19}$$

Hence the sum of squares due to the $\alpha_i$'s is given by (15.19). Similarly, the sum of squares due to $\beta_j$'s is given by

$$\sum_{ij} \left( \frac{x_{.j}}{m} - \frac{x_{..}}{mn} \right)^2 = \sum_{j} \frac{x_{.j}^2}{m} - \mathrm{C.F.} \tag{15.20}$$

If we assume $e_{ij} \sim N(0, \sigma^2)$ for all $i$ and $j$ and mutually independently distributed, then we can establish the following results by examining the corresponding quadratic forms:

$$\begin{aligned}
&\text{Sum of squares due to } \alpha_i\text{'s}\\
&= \sum_i \frac{x_{i.}^2}{n} - \text{C.F} = \sigma^2\chi^2_{m-1}(\lambda_1), \quad \lambda_1 = \text{non-centrality parameter}\\
&= \sigma^2\chi^2_{m-1} \quad \text{when } \alpha_1 = 0 = \cdots = \alpha_m\\
&\text{Sum of squares due to the } \beta_j\text{'s}\\
&= \sum_j \frac{x_{.j}^2}{m} - \text{C.F} = \sigma^2\chi^2_{n-1}(\lambda_2), \quad \lambda_2 = \text{non-centrality parameter}\\
&= \sigma^2\chi^2_{n-1} \quad \text{when } \beta_1 = 0 = \cdots = \beta_n\\
&\text{Total sum of squares} = \sum_{ij} x_{ij}^2 - \text{C.F} = \sigma^2\chi^2_{mn-1}\\
&\text{Least square minimum} = s^2 = \sigma^2\chi^2_{(m-1)(n-1)}.
\end{aligned} \tag{15.21}$$

Further, it can be shown that the least square minimum or the residual sum of squares $s^2$ and the sum of squares due to the $\alpha_i$'s are independently distributed. Similarly, $s^2$ and the sum of squares due to the $\beta_j$'s are independently distributed. But sum of squares due the $\alpha_i$'s and sum of squares due to $\beta_j$'s are not independently distributed. Thus the total sum of squares (S.S) can be split into the sum of the sum of squares due to $\alpha_i$'s, sum of squares due to the $\beta_j$'s and the residual sum of squares. That is,

$$\begin{aligned}
\text{Total S.S} &= \text{S.S due to } \alpha_i\text{'s}\\
&\quad + \text{S.S due to } \beta_j\text{'s} + \text{residual S.S}\\
\sum_{ij}\left(x_{ij} - \frac{x_{..}}{mn}\right)^2 &= \left[\sum_{ij}\left(\frac{x_{i.}}{n} - \frac{x_{..}}{mn}\right)^2\right]\\
&\quad + \left[\sum_{ij}\left(\frac{x_{.j}}{m} - \frac{x_{..}}{mn}\right)^2\right] + s^2
\end{aligned} \tag{15.22}$$

Then, under the hypothesis $\alpha_1 = 0 = \cdots = \alpha_m$ we have

$$\frac{(\text{S.S due to } \alpha_i\text{'s})/(m-1)}{s^2/[(m-1)(n-1)]} \sim F_{m-1,(m-1)(n-1)} \tag{15.23}$$

and under the hypothesis $\beta_1 = 0 = \cdots = \beta_n$ we have

$$\frac{(\text{S.S due to } \beta_j\text{'s})/(n-1)}{s^2/[(m-1)(n-1)]} \sim F_{n-1,(m-1)(n-1)}. \tag{15.24}$$

We can use (15.23) and (15.24) to test the hypotheses on $\alpha_i$'s and $\beta_j$'s, respectively. The degrees of freedom for the residual sum of squares is obtained by the formula:

$$mn - 1 - (m-1) - (n-1) = mn - m - n + 1 = (m-1)(n-1).$$

The residual sum of squares can also be obtained in a similar fashion as the total sum of squares minus the sum of squares due to $\alpha_i$'s minus the sum of squares due to $\beta_j$'s. Then the analysis of variance table or ANOVA table for a two-way classification with one observation per cell can be set up as follows, where d.f = degrees of freedom, S.S = sum of squares, M.S = mean square, C.F = correction factor = $\frac{x_{..}^2}{mn}$, $\nu = (m-1)(n-1)$:

### ANOVA table for a randomized block experiment

| Variation due to | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) = (3)/(2) | |
| Blocks | $m-1$ | $\sum_i \frac{x_{i.}^2}{n} - \text{C.F}$ | $A$ | $\frac{A}{C} = F_{m-1,\nu}$ |
| Treatments | $n-1$ | $\sum_j \frac{x_{.j}^2}{m} - \text{C.F}$ | $B$ | $\frac{B}{C} = F_{n-1,\nu}$ |
| Residual | $\nu$ | (obtained by<br>subtraction) | $C$ | |
| Total | $mn-1$ | $\sum_{ij} x_{ij}^2 - \text{C.F}$ | | |

The last column in the ANOVA table is "Inference", which is not shown in the above table due to lack of space. In the column on "Inference" write "significant" or "not significant". Here, "significant" means the hypothesis of no effect of the corresponding treatments is rejected, or we are saying that the contribution corresponding to the effects is significantly high compared to the residual sum of squares or the $F$-value is above the critical point. Similar is the inference on the significance of the block sum of squares also.

As in the one-way classification case, we can test for individual differences among $\alpha_j$'s as well as individual differences among $\beta_j$'s. This should be done only when the corresponding hypothesis is rejected or when the corresponding effect is found to be significantly high. If the block effect is found to be significantly high or if the hypothesis of no effect of the blocks is rejected, then test individual hypotheses of the type

$$H_0 : \alpha_i - \alpha_j = 0$$

by using the fact that the estimates of the effects $\alpha_i$ and $\alpha_j$ are linear functions of $e_{ij}$'s and, therefore, normally distributed under normality assumption for the $e_{ij}$'s. Hence we can use a Student-$t$ test since the population variance $\sigma^2$ is unknown. Note that

$$\begin{aligned}\hat{\alpha}_i - \hat{\alpha}_j &= \frac{x_{i.}}{n} - \frac{x_{j.}}{n}\\ &\sim N\left(\alpha_i - \alpha_j, \sigma^2\left(\frac{1}{n} + \frac{1}{n}\right) = 2\frac{\sigma^2}{n}\right)\end{aligned}$$

and hence under the hypothesis $H_0 : \alpha_i - \alpha_j = 0$

$$\frac{(\hat{\alpha}_i - \hat{\alpha}_j) - 0}{\sqrt{\hat{\sigma}^2(\frac{2}{n})}} = \frac{(\frac{x_{i.}}{n} - \frac{x_{j.}}{n})}{\hat{\sigma}\sqrt{\frac{2}{n}}} = \frac{x_{i.} - x_{j.}}{\sqrt{2n\hat{\sigma}^2}}$$
$$\sim t_{(m-1)(n-1)}.$$

Hence the criterion will be to reject the hypothesis $H_0 : \alpha_i - \alpha_j = 0$ if the observed value of

$$\left|\frac{\frac{x_{i.}}{n} - \frac{x_{j.}}{n} - 0}{\hat{\sigma}\sqrt{\frac{2}{n}}}\right| \geq t_{(m-1)(n-1),\frac{\alpha}{2}} \tag{15.25}$$

where

$$\hat{\sigma}^2 = \frac{\text{Least square minimum}}{(m-1)(n-1)} = \frac{s^2}{(m-1)(n-1)} \tag{15.26}$$

and

$$\Pr\{t_{(m-1)(n-1)} \geq t_{(m-1)(n-1),\frac{\alpha}{2}}\} = \frac{\alpha}{2}.$$

Start with the biggest absolute difference of $\hat{\alpha}_i - \hat{\alpha}_j$ and continue until the hypothesis is not rejected. Until that stage, all the differences are contributing towards the significant contribution due to $\alpha_i$'s. Similar procedure can be adopted for testing individual hypotheses on $\beta_i - \beta_j$. This is to be done only when the original hypothesis $\beta_1 = 0 = \cdots = \beta_n$ is rejected. Construction of confidence intervals can also be done as in the case of one-way classification. A $100(1-\alpha)\%$ confidence interval for $\alpha_i - \alpha_j$ as well as for $\beta_i - \beta_j$ are the following:

$$\frac{x_{i.}}{n} - \frac{x_{j.}}{n} \mp t_{(m-1)(n-1),\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{2}{n}} \tag{15.27}$$
$$\frac{x_{.i}}{m} - \frac{x_{.j}}{m} \mp t_{(m-1)(n-1),\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{2}{m}} \tag{15.28}$$

where $\hat{\sigma}^2$ is given in (15.26).

**Note 15.2.** Suppose that the hypothesis $\alpha_1 = 0 = \cdots = \alpha_m$ is not rejected but suppose that someone tries to test individual differences $\alpha_i - \alpha_j = 0$. Is it possible that some of the individual differences are significantly high or a hypothesis on individual difference being zero is rejected? It is possible and there is no inconsistency because, locally, some differences may be significantly high but overall contribution may not be significantly high. The same note is applicable to $\beta_j$'s also.

**Example 15.3.** The following is the data collected from a randomized block design without replication and it is assumed that blocks and treatments do not interact with each other. The data is collected and classified according to blocks $B_1, B_2, B_3, B_4$ and treatments $T_1, T_2, T_3$. Do the first stage analysis of block effects and treatment effects on this data.

|  | $T_1$ | $T_2$ | $T_3$ |
|---|---|---|---|
| $B_1$ | 1 | 5 | 8 |
| $B_2$ | 6 | 4 | 2 |
| $B_3$ | 2 | 4 | 4 |
| $B_4$ | 5 | 4 | 5 |

**Solution 15.3.** The number of rows $m = 4$ and the number of columns $n = 3$. The marginal sums are the following: $x_{1.} = 14$, $x_{2.} = 12$, $x_{3.} = 10$, $x_{4.} = 14$. $x_{.1} = 14$, $x_{.2} = 17$, $x_{.3} = 19$. $x_{..} = 14 + 17 + 19 = 50$. Then the correction factor

$$\text{C.F} = \frac{x_{..}^2}{mn} = \frac{50^2}{12} \approx 208.33.$$

Then the sum of squares due to rows

$$\sum_i \frac{x_{i.}^2}{n} - \text{C.F} = \frac{1}{3}[(14)^2 + (12)^2 + (10)^2 + (14)^2] - \text{C.F} \approx 3.67.$$

Sum of squares due to treatments

$$\sum_j \frac{x_{.j}^2}{m} - \text{C.F} = \frac{1}{4}[(14)^2 + (17)^2 + (19)^2] - \text{C.F} \approx 3.17.$$

Total sum of squares

$$\sum_{ij} x_{ij}^2 - \text{C.F} = (1)^2 + (5)^2 + (8)^2 + \cdots + (5)^2 - \text{C.F} \approx 39.67$$

Residual sum of squares

$$s^2 = 39.67 - 3.17 - 3.67 \approx 32.83.$$

Then the analysis of variance table can be set up as follows:

| Variation due to (1) | d.f (2) | S.S (3) | M.S (3)/(2) = (4) | F-ratio |
|---|---|---|---|---|
| Blocks | 3 | 3.67 | 1.22 | $0.22 = F_{3,6}$ |
| Treatments | 2 | 3.17 | 1.59 | $0.29 = F_{2,6}$ |
| Residual | 6 | 32.83 | 5.47 |  |
| Total | 11 | 39.67 |  |  |

Let us test at a 5% level of rejection. The tabulated values are the following:

$$F_{2,6,0.05} = 5.14, \quad F_{3,6,0.05} = 4.76.$$

Hence the hypothesis that the block effects are the same is not rejected. The hypothesis that the treatment effects are the same is also not rejected. Hence the contributions due to $\alpha_i$'s as well as due to $\beta_j$'s are not significant. Hence no further analysis of the differences between individual effects will be done here.

### 15.3.2 Two-way classification model with interaction

As explained earlier, there is a possibility that there may be a joint effect when two sets of treatments are tried in an experiment, such as variety of corn (one set of treatments) and fertilizers (second set of treatments). Certain variety may interact with certain fertilizers. In such a situation, a simple randomized block experiment with one observation per cell is not suitable for the analysis of the data. We need to replicate the design so that we have $r$ observations each in each cell. We may design a randomized block experiment to replicate $r$ times. Suppose that it is an experiment involving planting of tapioca. Animals like to eat the tapioca plant. Suppose that in some of the replicates a few plots are eaten up by animals and the final set of observations is of the form of $n_{ij}$ observations in the $(i,j)$-th cell, where the $n_{ij}$'s need not be equal. This is the general situation of a two-way classification model with multiple observations per cell. We will consider here only the simplest situation of equal numbers of observations per cell, and the general case will not be discussed here. The students are advised to read books on design of experiments for getting information on the general case as well as for other designs and also see the paper [1].

Consider the model of (15.16) where the $k$-th observation on $(i,j)$-th combination of the two types of treatments be $x_{ijk}$, $i = 1, \dots, m$, $j = 1, \dots, n$, $k = 1, \dots, r$. Then a linear, fixed effect, additive model with interaction is of the following type:

$$\begin{aligned} x_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\ &\qquad i = 1, \dots, m,\ j = 1, \dots, n,\ k = 1, \dots, r,\ e_{ijk} \sim N(0, \sigma^2) \qquad (15.29) \\ &= \mu_{ij} + e_{ijk}, \quad \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (15.30) \end{aligned}$$

where $\mu$ is a general effect, $\alpha_i$ is the deviation from the general effect due to the $i$-th treatment of the first set, $\beta_j$ is the deviation from the general effect due to the $j$-th treatment of the second set, $\gamma_{ij}$ is the interaction effect and $e_{ijk}$ is the random part. Here again, without loss of generality, we may assume $\alpha_. = 0$, $\beta_. = 0$, $\gamma_{..} = 0$. From the model (15.30), one can easily compute the residual sum of squares. Consider the error sum of squares as

$$\sum_{ijk} e_{ijk}^2 = \sum_{ijk} (x_{ijk} - \mu_{ij})^2. \qquad (15.31)$$

By differentiating the left side of (15.31) with respect to $\mu_{ij}$ and equating to zero, one has

$$\hat{\mu}_{ij} = \frac{\sum_k x_{ijk}}{r} = \frac{x_{ij.}}{r}. \tag{15.32}$$

Hence the least square minimum, $s^2$, is given by

$$\begin{aligned} s^2 &= \sum_{ijk}\left(x_{ijk} - \frac{x_{ij.}}{r}\right)^2 \\ &= \sum_{ijk}\left(x_{ijk} - \frac{x_{...}}{mnr}\right)^2 - \sum_{ijk}\left(\frac{x_{ij.}}{r} - \frac{x_{...}}{mnr}\right)^2 \\ &= \left(\sum_{ijk} x_{ijk}^2 - \frac{x_{...}^2}{mnr}\right) - \left(\sum_{ij}\frac{x_{ij.}^2}{r} - \frac{x_{...}^2}{mnr}\right). \end{aligned} \tag{15.33}$$

The first hypothesis to be tested is that $H_0 : \gamma_{ij} = 0$ or there is no interaction. Then the model is

$$x_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

and the least square minimum, denoted by $s_0^2$, is given by

$$\begin{aligned} s_0^2 = \sum_{ijk}\left(x_{ijk} - \frac{x_{...}}{mnr}\right)^2 - \sum_{ijk}\left(\frac{x_{i..}}{nr} - \frac{x_{...}}{mnr}\right)^2 \\ - \sum_{ijk}\left(\frac{x_{.j.}}{mr} - \frac{x_{...}}{mnr}\right)^2. \end{aligned} \tag{15.34}$$

Therefore, the sum of squares due to interaction is given by

$$\begin{aligned} s_0^2 - s^2 &= \left[\sum_{ijk}\left(x_{ijk} - \frac{x_{...}}{mnr}\right)^2 - \sum_{ijk}\left(\frac{x_{i..}}{nr} - \frac{x_{...}}{mnr}\right)^2\right. \\ &\quad \left. - \sum_{ijk}\left(\frac{x_{.j.}}{mr} - \frac{x_{...}}{mnr}\right)^2\right] - \left[\sum_{ijk}\left(x_{ijk} - \frac{x_{...}}{mnr}\right)^2 - \sum_{ijk}\left(\frac{x_{ij.}}{r} - \frac{x_{...}}{mnr}\right)^2\right] \\ &= \sum_{ijk}\left(\frac{x_{ij.}}{r} - \frac{x_{...}}{mnr}\right)^2 - \sum_{ijk}\left(\frac{x_{i..}}{nr} - \frac{x_{...}}{mnr}\right)^2 \\ &\quad - \sum_{ijk}\left(\frac{x_{.j.}}{mr} - \frac{x_{...}}{mnr}\right)^2 \\ &= \left[\sum_{ij}\frac{x_{ij.}^2}{r} - \mathrm{C.F}\right] - \left[\sum_{i}\frac{x_{i..}^2}{nr} - \mathrm{C.F}\right] - \left[\sum_{j}\frac{x_{.j.}^2}{mr} - \mathrm{C.F}\right] \end{aligned} \tag{15.35}$$

where C.F $= \frac{x_{...}^2}{mnr}$. When the interaction $\gamma_{ij} = 0$, then there is meaning in estimating $\alpha_i$'s and $\beta_j$'s separately. When the interaction effect $\gamma_{ij}$ is present, then part of the effect due to the $i$-th treatment of the first set is mixed up with $\gamma_{ij}$. Similarly, part of the effect of the $j$-th treatment of the second set is also mixed up with $\gamma_{ij,}$ and hence one should not try to estimate $\alpha_i$ and $\beta_j$ separately when $\gamma_{ij}$ is present. Hence, testing of hypotheses on $\alpha_i$'s should be done only if the hypothesis $\gamma_{ij} = 0$ is not rejected; similar is the case for testing hypotheses on $\beta_j$'s.

Now we can check the degrees of freedom for the various chi-squares when $e_{ijk} \sim N(0, \sigma^2)$ for all $i, j, k$, and mutually independently distributed. The sum of squares due to interaction is denoted by S.S.(int) and it is given by

$$\text{S.S(int)} = \left[\sum_{ij} \frac{x_{ij.}^2}{r} - \text{C.F}\right] - \left[\sum_{i} \frac{x_{i..}^2}{nr} - \text{C.F}\right] - \left[\sum_{j} \frac{x_{.j.}^2}{mr} - \text{C.F}\right] \tag{15.36}$$

with degrees of freedom $[mn-1] - [m-1] - [n-1] = (m-1)(n-1)$. The residual sum of squares is $[\sum_{ijk} x_{ijk}^2 - \text{C.F}] - [\sum_{ij} \frac{x_{ij.}^2}{r} - \text{C.F}]$ with degrees of freedom $[mnr-1] - [mn-1] = mn(r-1)$. Once $\gamma_{ij} = 0$, then the sum of squares due to $\alpha_i$ is $[\sum_i \frac{x_{i..}^2}{nr} - \text{C.F}]$ with degrees of freedom $[m-1]$. Similarly, once $\gamma_{ij} = 0$ then the sum of squares due to $\beta_j$ is $[\sum_j \frac{x_{.j.}^2}{mr} - \text{C.F}]$ with degrees of freedom $[n-1]$. Now, we can set up analysis of variance table.

In the following table, Set A means the first set of treatments and Set B means the second set of treatments, C.F = correction factor $= \frac{x_{...}^2}{mnr}$, $\rho = mn(r-1)$, is the degrees of freedom for the residual sum of squares, $\nu = (m-1)(n-1)$ is the degrees of freedom for the interaction sum of squares, and interaction sum of squares, which is given above in (15.36), is denoted by S.S(int).

**ANOVA for two-way classification with *r* observations per cell**

| Variation | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| due to | | | $\frac{(3)}{(2)}$ | |
| (1) | (2) | (3) | = (4) | |
| Set A | $m-1$ | $\sum_i \frac{x_{i..}^2}{nr} - \text{C.F}$ | $A$ | $\frac{A}{D} = F_{m-1,\rho}$ |
| Set B | $n-1$ | $\sum_j \frac{x_{.j.}^2}{mr} - \text{C.F}$ | $B$ | $\frac{B}{D} = F_{n-1,\rho}$ |
| Interaction | $(m-1)(n-1)$ | S.S(int) | $C$ | $\frac{C}{D} = F_{\nu,\rho}$ |
| Between cells | $mn-1$ | $\sum_{ij} \frac{x_{ij.}^2}{r} - \text{C.F}$ | | |
| Residual | $\rho$ | (by subtraction) | $D$ | |
| Total | $mnr-1$ | $\sum_{ijk} x_{ijk}^2 - \text{C.F}$ | | |

**Note 15.3.** There is a school of thought that when the interaction effect is found to be insignificant or the hypothesis $H_0 : \gamma_{ij} = 0$ is not rejected, then add up the sum of squares due to interaction along with the residual sum of squares, with the corresponding degrees of freedoms added up, and then test the main effects $\alpha_i$'s and $\beta_j$'s against this new residual sum of squares. We will not adopt that procedure because, even though the interaction sum of squares is not significantly high it does not mean that there is no contribution from interaction. Hence we will not add up the interaction sum of squares to the residual sum of squares. We will treat them separately, even if the interaction effect is not significant. If insignificant then we will proceed to test hypotheses on the main effects $\alpha_i$'s and $\beta_j$'s, otherwise we will not test hypotheses on $\alpha_i$'s and $\beta_j$'s.

Individual hypotheses on the main effects $\alpha_i$'s and $\beta_j$'s can be tested only if the interaction effect is found to be insignificant. In this case,

$$\frac{\hat{\alpha}_s - \hat{\alpha}_t}{\hat{\sigma}\sqrt{\frac{2}{nr}}} = \frac{\frac{x_{s..}}{nr} - \frac{x_{t..}}{nr}}{\hat{\sigma}\sqrt{\frac{2}{nr}}} \sim t_{mn(r-1)} \qquad (15.37)$$

under the hypothesis $\alpha_s - \alpha_t = 0$, where

$$\hat{\sigma}^2 = \frac{\text{Least square minimum}}{mn(r-1)} = \frac{s^2}{mn(r-1)}.$$

A similar result can be used for testing the hypothesis $\beta_s - \beta_t = 0$. The dividing factor in this case is $mr$ instead of $nr$. The confidence interval can also be set up by using the result (15.37) and the corresponding result for $\hat{\beta}_s - \hat{\beta}_t$.

**Example 15.4.** A randomized block design is done on 3 blocks and 4 treatments and then it is replicated 3 times. The final data are collected and classified into the following format. Do an analysis of the following data:

| | $T_1$ | $T_2$ | $T_3$ | | $T_1$ | $T_2$ | $T_3$ | | $T_1$ | $T_2$ | $T_3$ | | $T_1$ | $T_2$ | $T_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $B_1$ | 1 | 4 | 2 | $B_2$ | 5 | 6 | 4 | $B_3$ | 8 | 6 | 4 | $B_4$ | 1 | 2 | 2 |
| | 2 | 2 | 3’ | | 6 | 8 | 9’ | | 5 | 5 | 6’ | | 2 | 1 | 1 |
| | 3 | 4 | 4 | | 4 | 5 | 5 | | 6 | 5 | 4 | | 1 | 3 | 2 |

**Solution 15.4.** Here, $i = 1, 2, 3, 4$ or $m = 4$, $j = 1, 2, 3$ or $n = 3$, $k = 1, 2, 3$ or $r = 3$.

$$x_{11.} = 6, \quad x_{12.} = 10, \quad x_{13.} = 8, x_{21.} = 15, \quad x_{22.} = 19, \quad x_{23.} = 18$$
$$x_{31.} = 19, \quad x_{32.} = 16, \quad x_{33.} = 14x_{41.} = 4, \quad x_{42.} = 6, \quad x_{43.} = 5$$

Total $= \sum_{ijk} x_{ijk} = 140$. Hence C.F $= \frac{x_{...}^2}{mnr} = \frac{(140)^2}{36} \approx 544.44$

$$\text{Block S.S} = \sum_i \frac{x_{i..}^2}{nr} - \text{C.F}$$

$$= \frac{1}{9}[(24)^2 + (52)^2 + (49)^2 + (15)^2] - \text{C.F} \approx 111.78$$

$$\text{Treatment S.S} = \sum_j \frac{x_{.j.}^2}{mr} - \text{C.F}$$

$$= \frac{1}{12}[(44)^2 + (51)^2 + (45)^2] - \text{C.F} \approx 2.39$$

$$\sum_{ij} \frac{x_{ij.}^2}{r} - \text{C.F} = \frac{1}{3}[(6)^2 + (10)^2 + \cdots + (5)^2] - \text{C.F} \approx 122.23$$

$$\text{Residual S.S.} = 166.56 - 122.23 = 44.33$$

$$\text{Total S.S.} = \sum_{ijk} x_{ijk}^2 - \text{C.F} \approx 166.56$$

Then the analysis of variance table is the following:

| Variation due to (1) | d.f (2) | S.S (3) | M.S (3)/(2) = (4) | F-ratio |
|---|---|---|---|---|
| Blocks | 3 | 111.78 | 37.26 | $20.14 = F_{3,24}$ |
| Treatments | 2 | 2.5 | 1.25 | $0.68 = F_{2,24}$ |
| Interaction | 6 | 7.95 | 1.34 | $4.30 = F_{6,24}$ |
| Between cells | 11 | 122.23 | | |
| Residual | 24 | 44.33 | 1.85 | |
| Total | 35 | 166.56 | | |

Let us test at $\alpha = 0.05$ level of rejection. Then $F_{6,24,0.05} = 2.51 < 4.30$ from tables, and hence the hypothesis of no interaction is rejected. Since this hypothesis is rejected, there is possibility of interaction, and hence we cannot test any hypothesis on the main effects or on $\alpha_i$'s and $\beta_j$'s, and hence we stop the analysis here.

## Exercises 15.3

**15.3.1.** When $e_{ij} \sim N(0, \sigma^2)$ and independently distributed in a two-way classification linear fixed effect model without interaction, with $i = 1, \dots, m$, $j = 1, \dots, n$ prove the following:

(i) The block sum of squares $= \sum_{ij}\left(\frac{x_{i.}}{n} - \frac{x_{..}}{mn}\right)^2 = \sum_i \frac{x_{i.}^2}{n} - \frac{x_{..}^2}{mn}$

$$= A \sim \sigma^2 \chi^2_{m-1}$$

under $H_0 : \alpha_1 = 0 = \cdots = \alpha_m$

(ii) Treatment sum of squares

$$= \sum_{ij}\left(\frac{x_{.j}}{m} - \frac{x_{..}}{mn}\right)^2 = \sum_j \frac{x_{.j}^2}{m} - \frac{x_{..}^2}{mn}$$

$$= B \sim \sigma^2\chi^2_{n-1}$$

$$\text{under } H_0 : \beta_1 = 0 = \cdots = \beta_n$$

(iii) Residual sum of squares

$$= C = \sum_{ij}\left(x_{ij} - \frac{x_{..}}{mn}\right)^2 - A - B$$

$$= \sum_{ij} x_{ij}^2 - \frac{x_{..}^2}{mn} - A - B$$

$$\sim \sigma^2\chi^2_{(m-1)(n-1)}$$

(iv) $A$ and $C$ are independently distributed, and

(v) $B$ and $C$ are independently distributed.

Hint: Write in terms of $e_{ij}$'s and then look at the matrices of the corresponding quadratic forms.

**15.3.2.** Consider the two-way classification fixed effect model with interaction as in equation (15.16). Let $e_{ijk} \sim N(0,\sigma^2)$, $i = 1,\dots,m$, $j = 1,\dots,n$, $k = 1,\dots,r$, and independently distributed. By expressing various quantities in terms of $e_{ijk}$'s and then studying the properties of the corresponding quadratic forms establish the following results:

(i) Residual sum of squares

$$= \sum_{ijk}\left(x_{ijk} - \frac{x_{ij.}}{r}\right)^2$$

$$= \left[\sum_{ijk}\left(x_{ijk} - \frac{x_{...}}{mnr}\right)^2\right] - \left[\sum_{ijk}\left(\frac{x_{ij.}}{r} - \frac{x_{...}}{mnr}\right)^2\right]$$

$$= D \sim \sigma^2\chi^2_{mn(r-1)}$$

(ii) Total sum of squares

$$= \sum_{ijk}\left(x_{ijk} - \frac{x_{...}}{mnr}\right)^2 \sim \sigma^2\chi^2_{mnr-1}$$

(iii) Interaction sum of squares

$$= \left(\sum_{ij}\frac{x_{ij.}^2}{r} - \frac{x_{...}^2}{mnr}\right) - \left(\sum_i \frac{x_{i..}^2}{nr} - \frac{x_{...}^2}{mnr}\right)$$

$$- \left(\sum_j \frac{x_{.k.}^2}{mr} - \frac{x_{...}^2}{mnr}\right) = C$$

$$\sim \sigma^2\chi^2_{(m-1)(n-1)}$$

when $\gamma_{ij} = 0$ for all $i$ and $j$.

(iv) when $\gamma_{ij} = 0$ for all $i$ and $j$ then $A$ and $D$ as well as $B$ and $D$ are independently distributed, where

$$A = \left(\sum_i \frac{x_{i..}^2}{nr} - \frac{x_{...}^2}{mnr}\right), \quad B = \left(\sum_j \frac{x_{.j.}^2}{mr} - \frac{x_{...}^2}{mnr}\right).$$

**15.3.3.** Do a complete analysis of the following data on a randomized block experiment where $B_1, B_2, B_3$ denote the blocks and $T_1, T_2, T_3, T_4$ denote the treatments. The experiment is to study the yield $x_{ij}$ of beans where the 3 blocks are the 3 locations and the 4 treatments are the 4 varieties. If the block effect is found to be significant then check for individual differences in $\alpha_i$'s. Similarly, if the treatment effect is found to be significant then check for individual differences, to complete the analysis.

| | $T_1$ | $T_2$ | $T_3$ | $T_4$ |
|---|---|---|---|---|
| $B_1$ | 8 | 9 | 9 | 8 |
| $B_2$ | 5 | 6 | 5 | 4 |
| $B_3$ | 1 | 0 | 2 | 3 |

**15.3.4.** Do a complete analysis of the following data on a two-way classification with possibility of interaction. The first set of treatments are denoted by $A_1, A_2, A_3$ and the second set of treatments are denoted by $B_1, B_2, B_3, B_4$ and there are 3 replications (3 data points in each cell). If interaction is found to be insignificant, then test for the main effects. If the main effects are found to be contributing significantly, then check for individual differences.

| | $B_1$ | $B_2$ | $B_3$ | $B_4$ | | $B_1$ | $B_2$ | $B_3$ | $B_4$ | | $B_1$ | $B_2$ | $B_3$ | $B_4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $A_1$ | 12 | 11 | 10 | 11 | $A_2$ | 5 | 6 | 5 | 5 | $A_3$ | 1 | 0 | 2 | 3 |
| | 11 | 10 | 10 | 11 | | 6 | 6 | 5 | 5 | | 2 | 1 | 0 | 1 |
| | 10 | 12 | 12 | 12 | | 5 | 6 | 6 | 6 | | 3 | 1 | 1 | 2 |

## 15.4 Latin square designs

**Definition 15.2** (A Latin Square). A Latin Square is a square arrangement of $m$ Latin letters into $m$ rows and $m$ columns so that each letter appears in each row and each column once and only once.

Consider the following arrangements of 3 Latin letters, 3 Greek letters and 3 numbers in to 3 rows and 3 columns:

$$M_1 = \begin{array}{ccc} A & B & C \\ B & C & A, \\ C & A & B \end{array} \qquad M_2 = \begin{array}{ccc} \alpha & \beta & \gamma \\ \gamma & \alpha & \beta, \\ \beta & \gamma & \alpha \end{array} \tag{15.38}$$

$$M_3 = \begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1, \\ 3 & 1 & 2 \end{array} \qquad M_{12} = \begin{array}{ccc} A\alpha & B\beta & C\gamma \\ B\gamma & C\alpha & A\beta \\ C\beta & A\gamma & B\alpha \end{array} \tag{15.39}$$

Note that $M_1, M_2, M_3$ are all Latin squares, $M_2$ has Greek letters and $M_3$ has numbers in the cells but all satisfy the conditions in the Definition 15.2. Note that $M_{12}$ is in the form of $M_2$ superimposed on $M_1$. In this superimposed structure, every Greek letter comes with every Latin letter once and only once. If there are two Latin squares, which when superimposed, have the property that every letter in one square comes with every letter in the other square, once and only once, then such squares are called *orthogonal Latin squares* or Greek-Latin squares. There are some results on the maximum number of such orthogonal squares possible for a given $m$. The maximum possible is evidently $m-1$ but for every given $m$ these $m-1$ orthogonal squares may not exist. Construction of all orthogonal squares for a given $m$ is an open problem. Once in a while people come up with all squares for a new number $m$.

Here, we are concerned about using a Latin square for constructing designs called *Latin square designs*. In a Latin square design, we will assign one set of treatments corresponding to rows, one set of treatments corresponding to columns, one set of treatments corresponding to Latin letters. If orthogonal squares are available, then additional sets of treatments corresponding to the letters in each orthogonal square can be tried. The total number of cells is only $m^2$ or by using $m^2$ experimental plots one will be able to test hypotheses on different sets of $m$ treatments each. This is the main advantage of a Latin square design. The trade-off is that all the treatment sets must have equal number of treatments or $m$ treatments in each set. Another drawback is that in the analysis, there is no provision for interaction, and hence do not conduct an experiment with the help of a Latin square design if the different sets of treatments are likely to interact with each other, or if effects due to combination of treatments is likely to be present then do not use a Latin square design. Let us start with one Latin square with $m$ sides, something like $M_1$ in (15.38). A model that we can take is the following:

$$x_{ij(k)} = \mu + \alpha_i + \beta_j + \gamma_{(k)} + e_{ij(k)} \tag{15.40}$$

where $x_{ij(k)}$ is the observation in the $(i,j)$-th cell if the $k$-th letter is present in the $(i,j)$-th cell. For example, in the illustrative design $M_1$ the letter $A$ or the first letter appears, for example, in the first row first column cell. Hence $x_{111}$ is there whereas $x_{112}$ and $x_{113}$ are not there since the letters $B$ and $C$ do not appear in the $(1,1)$-th cell. This is the meaning of the subscript $k$ put in brackets. Since every letter is present in every row and every column, when we sum up with respect to $i$, row subscript, $k$ is automatically summed up. Similarly, when we sum up with respect to column subscript $j$ the subscript $k$ is also automatically summed up. Let us use the following notations.

$$R_i = i\text{-th row sum}; \quad C_j = j\text{-th column sum}; \quad T_k = k\text{-th letter sum}.$$

For calculating $R_i$, sum up all observations in the $i$-th row. Similarly, sum of all observations in the $j$-th column to obtain $C_j$. But $T_k$ is obtained by searching for the $k$-th treatment in each row and then summing up the corresponding observations. By computing the least square estimates of the effects and then substituting in the error sum of squares, we obtain the least square minimum $s^2$. Then put the hypothesis that the first set of treatment effects are zeros or $H_0 : \alpha_1 = 0 = \cdots = \alpha_m$. Compute the least square minimum under this hypothesis, $s_0^2$. Then take $s_0^2 - s^2$ to obtain the sum of squares due to the rows or due to $\alpha_i$'s. Similarly, by putting $H_0 : \beta_1 = 0 = \cdots = \beta_m$ compute the least square minimum under this hypothesis. Call it $s_{00}^2$. Then $s_{00}^2 - s^2$ is the sum of squares due to the second set of treatments or columns. By putting $H_0 : \gamma_1 = 0 = \cdots = \gamma_m$ and taking the difference between the least square minima, one under the hypothesis and one without any hypothesis, we obtain the sum of squares due to $\gamma_k$'s or the third set of treatments. The sum of squares can be simplified to the following, where the degrees of freedom corresponds to the degrees of freedom (d.f) associated with the corresponding chi-squares when it is assumed that the $e_{ij}$'s are independently distributed as $e_{ij} \sim N(0, \sigma^2)$. Here, $\text{C.F} = \frac{x_{..}^2}{m^2}$ = correction factor and S.S = sum of squares.

$$\begin{aligned} &\sum_i \frac{R_i^2}{m} - \text{C.F} = \text{S.S due to rows}, \quad \text{with d.f} = m - 1 \\ &\sum_j \frac{C_j^2}{m} - \text{C.F} = \text{S.S due to columns}, \quad \text{with d.f} = m - 1 \\ &\sum_k \frac{T_k^2}{m} - \text{C.F} = \text{S.S due to letters}, \quad \text{with d.f} = m - 1. \end{aligned} \tag{15.41}$$

The total sum of squares is $\sum_{ij} x_{ij}^2 - \text{C.F}$ with d.f $= m^2 - 1$ and the residual sum of squares $s^2$ = the total sum of squares minus the sum of squares due to rows, columns, and letters or the three sets of treatments, with degrees of freedom $\nu = (m^2 - 1) - 3(m - 1) = (m - 1)(m - 2)$ for $m \geq 3$. By using the above observations, one can set up the analysis of variance or ANOVA table for a simple Latin square design, where the three sets of treatments, one corresponding to the rows, one corresponding to the columns and one corresponding to the letters, are such that there is no possibility of interactions among any two of them.

One more set of treatments can be tried if we have a pair of orthogonal designs or if we have a Greek-Latin square as $M_{12}$ of (15.39). In this case, the model will be of the form:

$$x_{ij(kt)} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_t + e_{ij} \tag{15.42}$$

where $i = 1, \ldots, m$, $j = 1, \ldots, m$, $k = 1, \ldots, m$, $t = 1, \ldots, m$. When we sum up with respect to $i$ or $j$, automatically $k$ and $t$ are summed up. In this case, corresponding to (15.41), there will be one more sum of squares due to the fourth set of treatments, denoted by $\sum_t \frac{U_t^2}{m} - \text{C.F}$ with degrees of freedom $m - 1$ again. The correction factor remains the same as above. In this case, the degrees of freedom for the residual sum of squares is $\nu_1 = m^2 - 1 - 4(m - 1) = (m - 1)(m - 3)$ for $m \geq 4$. The analysis of variance table for the simple Latin square design and the Greek-Latin square designs are the following:

### ANOVA table for a simple Latin square design

| Variation due to | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) = (3)/(2) | |
| Rows | $m-1$ | $\sum_i \frac{R_i^2}{m} - \text{C.F}$ | $A$ | $\frac{A}{D} = F_{m-1,v}$ |
| Columns | $m-1$ | $\sum_j \frac{C_j^2}{m} - \text{C.F}$ | $B$ | $\frac{B}{D} = F_{m-1,v}$ |
| Letters | $m-1$ | $\sum_k \frac{T_k^2}{m} - \text{C.F}$ | $C$ | $\frac{C}{D} = F_{m-1,v}$ |
| Residue | $v$ | $s^2$ | $D$ | |
| Total | $m^2-1$ | $\sum_{ij} x_{ij}^2 - \text{C.F}$ | | |

where $v = (m^2-1) - 3(m-1) = (m-1)(m-2)$.

### ANOVA table for a Greek-Latin square design

| Variation due to | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) = (3)/(2) | |
| Rows | $m-1$ | $\sum_i \frac{R_i^2}{m} - \text{C.F}$ | $A$ | $\frac{A}{E} = F_{m-1,v_1}$ |
| Columns | $m-1$ | $\sum_j \frac{C_j^2}{m} - \text{C.F}$ | $B$ | $\frac{B}{E} = F_{m-1,v_1}$ |
| Latin letters | $m-1$ | $\sum_k \frac{T_k^2}{m} - \text{C.F}$ | $C$ | $\frac{C}{E} = F_{m-1,v_1}$ |
| Greek letters | $m-1$ | $\sum_t \frac{U_t^2}{m} - \text{C.F}$ | $D$ | $\frac{D}{E} = F_{m-1,v_1}$ |
| Residue | $v_1$ | $s^2$ | $E$ | |
| Total | $m^2-1$ | $\sum_{ij} x_{ij}^2 - \text{C.F}$ | | |

where $v_1 = (m-1)(m-3)$, $m \geq 4$. If we have more orthogonal designs or a set of $n$ orthogonal designs, then by using one set of $n$ orthogonal designs we can try $(n+2)$ sets of treatments by using $m^2$ test plots for $m \geq n+2$. The procedure is exactly the same. The residual degrees of freedom in this case will be $v_2 = (m^2-1) - (n+2)(m-1) = (m-1)(m-n-1)$. One illustrative example on a simple Latin square design will be given here.

**Example 15.5.** The following is the design and the data on a simple Latin square design. Do the analysis of the data.

| | | | | | | | Sum |
|---|---|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | | 1 | 5 | 4 | 10 |
| $B$ | $C$ | $A$ | | 2 | 6 | 7 | 15 |
| $C$ | $A$ | $B$ | | 5 | 2 | 4 | 11 |
| | | | Sum | 8 | 13 | 15 | 36 |

**Solution 15.5.** According to our notation the row sums, denoted by $R_1, R_2, R_3$, column sums, denoted by $C_1, C_2, C_3$ and sums corresponding to letters, denoted by $T_1, T_2, T_3$ are the following:

$$R_1 = 10, \quad R_2 = 15, \quad R_3 = 11$$
$$C_1 = 8, \quad C_2 = 13, \quad C_3 = 15$$
$$T_1 = 10, \quad T_2 = 11, \quad T_3 = 15$$

Note that $x_{..} = 36$, and hence the

$$\text{C.F} = \frac{x_{..}^2}{m^2} = \frac{(36)^2}{9} = 144.$$

Also

$$\text{S.S due to rows} = \sum_i \frac{x_{i.}^2}{m} - \text{C.F} \approx 144.67$$
$$\text{S.S due to columns} = \sum_j \frac{x_{.j}^2}{m} - \text{C.F} \approx 148.67$$
$$\text{S.S. due to letters} = \sum_k \frac{T_k^2}{m} \approx 144.67$$

Hence the ANOVA table is the following:

| Variation due to | d.f | S.S | M.S | F-ratio |
|---|---|---|---|---|
| Rows | 2 | 4.67 | 2.32 | 0.17 |
| Columns | 2 | 8.67 | 4.32 | 0.31 |
| Letters | 2 | 4.67 | 2.32 | 0.17 |
| Residue | 2 | 13.99 | | |
| Total | 8 | 32 | | |

Let us test at $\alpha = 0.05$. Then the tabulated value of $F_{2,2,0.05} = 19$. Hence the hypothesis $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ is not rejected or the row effect is not significant. Similarly, it can be seen that the column effect as well as letter effect are not significant here.

## Exercises 15.4

**15.4.1.** For a Latin square design of side $m$, and with normality assumption for the errors [$e_{ij} \sim N(0, \sigma^2)$ and mutually independently distributed] show that the residual sum of squares $s^2 \sim \sigma^2\chi^2_{(m-1)(m-2)}$ and that

$$\sum_i \frac{R_i^2}{m} - \text{C.F} \sim \sigma^2\chi^2_{m-1} \quad \text{under } H_0 : \alpha_1 = 0 = \cdots = \alpha_m$$

$$\sum_j \frac{C_j^2}{m} - \text{C.F} \sim \sigma^2\chi^2_{m-1} \quad \text{under } H_0 : \beta_1 = 0 = \cdots = \beta_m$$

$$\sum_k \frac{T_k^2}{m} - \text{C.F} \sim \sigma^2\chi^2_{m-1} \quad \text{under } H_0 : \gamma_1 = 0 = \cdots = \gamma_m$$

where C.F = correction factor = $\frac{x^2_{..}}{m^2}$, $R_i$ = $i$-th row sum, $C_j$ = $j$-th column sum and $T_k$ is the sum of the observations corresponding to the $k$-th letter.

**15.4.2.** In Exercise 15.4.1, show that $s^2$ and sum of squares due to rows are independently distributed and so is the case of column sum of squares and sum of squares corresponding to letters.

**15.4.3.** Do a complete analysis of the following data where the design and the corresponding data are given:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $A$ | $B$ | $C$ | $D$ | | 5 | 8 | 2 | 6 |
| $B$ | $C$ | $D$ | $A$ | | 4 | 2 | 1 | 5 |
| $C$ | $D$ | $A$ | $B$’ | | 3 | 8 | 2 | 4 |
| $D$ | $A$ | $B$ | $C$ | | 2 | 5 | 2 | 6 |

## 15.5 Some other designs

There are several other designs in practical use, such as incomplete block designs, balanced incomplete block designs, partially balanced incomplete block designs, Youden square designs, factorial designs, etc.

### 15.5.1 Factorial designs

In drug testing experiments, usually the same drug at different doses are administered. If two drugs at 5 different doses each are tried in an experiment, then we call it two factors at 5 levels and write as $2^5$ design. If $m$ factors at $n$ levels each are tried in an experiment then the design is called a $m^n$ factorial design. If $m_1$ factors at $n_1$ levels each, ..., $m_k$ factors at $n_k$ levels each are tried in an experiment, then we call it a $m_1^{n_1} \cdots m_k^{n_k}$ factorial design. The analysis of factorial designs is completely different from what we have done so far, because there can be all sorts of effects such as linear effects, quadratic effects and so on, as well as different categories of interactions. This is an area by itself.

### 15.5.2 Incomplete block designs

In a randomized block experiment, each block has $t$ plots and only $t$ treatments are tried, or in other words, all treatments appear in every block. Most of the times it may be difficult to find sets of $t$ plots each which are fully homogeneous within each block. In an experiment involving animals, it may be difficult to find a large number of identical animals with respect to genotype and other characteristics. In such a situation, we go for incomplete block designs. In each block, there will be $s$ homogeneous plots, $s < t$, and take $b$ such blocks so that $bs$ plots are there. Then the $t$ treatments are randomly assigned to the plots so that each treatment is repeated $r$ times in $r$ different blocks or $bs = rt$. Then such a design is called an incomplete block design. We may put other restrictions such as each pair of treatments appear $\lambda$ times or the $i$-th pair is repeated $\lambda_i$ times and so on. There are different types of balancing as well as partial balancing possible and such classes of designs are called balanced and partially balanced incomplete block designs.

Sometime, in a Latin square design, one or more rows or columns may be fully lost before the experiment is completed. The remaining rows and columns, of course, do not make a Latin square. They will be incomplete Latin squares, called Youden squares. Such designs are called Youden square designs.

### 15.5.3 Response surface analysis

In all the analysis of various problems that we have considered so far, we took the model as linear additive models. For example, in a one-way classification model we have taken the model as

$$x_{ij} = \mu + \alpha_i + e_{ij} \tag{15.43}$$

where $\mu$ is the general effect, $\alpha_i$ is the deviation from the general effect due to the $i$-th treatment and $e_{ij}$ is the random part. Here, $x_{ij}$, the observation, could be called the response to the $i$-th treatment. In general, $x_{ij}$ could be some linear or non-linear function. Let us denote it by $\phi$. Then

$$x_{ij} = \phi(\mu, \alpha_i, e_{ij}). \tag{15.44}$$

Analysis of this non-linear function $\phi$ is called the response surface analysis. This is an area by itself.

### 15.5.4 Random effect models

So far,we have been considering only fixed effect models. For example, in a model such as the one in (15.43), we assumed that $\alpha_i$ is fixed unknown quantity, not another random variable. If $\alpha_i$ is a random variable with $E(\alpha_i) = 0$ and $\mathrm{Var}(\alpha_i) = \sigma_1^2$, then the final analysis will be different. For simplifying matters, one can assume both $\alpha_i$'s and $e_{ij}$'s are independently normally distributed. Such models will be called random effect models, the analysis of which will be more complicated compared to the analysis of fixed effect models. Students who are interested in this area are advised to read books on the Design of Experiments and books on Response Surface Analysis.

# 16 Questions and answers

In this chapter, some of the questions asked by students at the majors level (middle level courses) are presented. At McGill University, there are three levels of courses, namely, honors level for very bright students, majors level for average students in mathematical, physical and engineering sciences and faculty programs for below average students from physical, biological and engineering sciences, and students from social sciences. Professor Mathai had taught courses at all levels for the past 57 years from 1959 onward. Questions from honors level students are quite deep, and hence they are not included here. Questions from students in the faculty programs are already answered in the texts in Chapters 1 to 15 and in the comments, notes and remarks therein. Questions at the majors level that Professor Mathai could recollect and which are not directly answered in the texts in Chapters 1 to 15 are answered here, mostly for the benefits of teachers of probability and statistics at the basic level and curious students. [These materials are taken from Module 9 of CMSS (Author: A. M. Mathai), and hence the contexts are those of Module 9.]

## 16.1 Questions and answers on probability and random variables

**Question.** Can a given experiment be random and non-random at the same time?

**Answer.** The answer is in the affirmative. It all depend upon what the experimenter is looking for in that experiment. Take the example of throwing a stone into a pond of water. If the outcome that she is looking for is whether the stone sinks in water or not, then the experiment is not a random experiment because from physical laws we know that the stone sinks, and hence the outcome is pre-determined, whereas if she is looking for the region on the surface of the pond where the stone hits the water, then it becomes a random experiment because the location of hit is not determined beforehand.

**Question.** What are called, in general, postulates or axioms?

**Answer.** Postulates or axioms are assumptions that you make to define something. These assumptions should be consistent and mutually non-overlapping. Since they are your own assumptions, there is no question of proving or disproving these postulates or axioms.

**Question.** In some books, a coin tossed twice and two coins tossed once are taken as equivalent. Is this correct?

**Answer.** No. If the two coins are identical in every respect and the original act of tossing twice and the second act of tossing once are identical in every respect, then the two situations can be taken as one and the same, otherwise not, usually not.

https://doi.org/10.1515/9783110562545-016

**Question.** In the situation of a random cut of an interval, the probability that the cut is at a given point is zero. Does it mean that it is impossible to cut the string? The child has already cut the string! Is there any contradiction here?

**Answer.** According to our rule here, we assigned probabilities proportional to length. Since a point has zero length, the assigned probability is zero. Probability of the impossible event $\phi$ is zero but if the assigned probability of an event is zero this does not mean that the event is impossible. We can cut the string.

**Question.** Does pair-wise independence imply mutual independence? For example, if $A, B, C$ are three events in $S$ and if $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, $P(B \cap C) = P(B)P(C)$, then does this imply that $P(A \cap B \cap C) = P(A)P(B)P(C)$ also?

**Question.** The second doubt is that if $P(A \cap B \cap C) = P(A)P(B)P(C)$ will it not be sufficient for pair-wise independence also?

**Answers.** Apparently, in some books, $P(A \cap B \cap C) = P(A)P(B)P(C)$ is stated as implying that they are pair-wise independent also. This is incorrect. In the following figures, suppose that symmetry is assumed in the sample space $S$ and $S$ contains only a finite number of elementary events. Figure 16.1 (a) is an illustration showing that pair-wise independence does not imply mutual independence. Figure 16.1 (b) shows that if PPP (product probability property) holds for three events then that need not imply pair-wise independence.

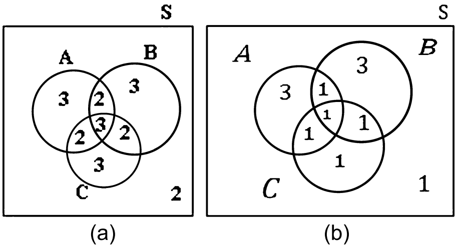


**Figure 16.1:** Pairwise and mutual independence.

In Figure 16.1 (a), $P(A) = \frac{1}{2} = P(B) = P(C)$, $P(A \cap B) = \frac{1}{4} = P(A)P(B)$, $P(A \cap C) = \frac{1}{4} = P(A)P(C)$, $P(B \cap C) = \frac{1}{4} = P(B)P(C)$, and hence pair-wise independence holds but $P(A \cap B \cap C) = \frac{1}{20} \neq P(A)P(B)P(C) = \frac{1}{8}$. Hence pair-wise independence need not imply mutual independence. In Figure 16.1 (b), $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$, $P(C) = \frac{1}{2}$ and $P(A \cap B \cap C) = \frac{1}{12} = P(A)P(B)P(C)$. But $P(B \cap C) = \frac{1}{12} \neq P(B)P(C) = \frac{1}{6}$. Hence if PPP (product probability property) holds for three events, PPP need not hold pair-wise. If there are $k$ events $A_i \subset S$, $i = 1, \dots, k$ for mutual independence to hold, one must have PPP holding for all possible intersections of different events, that is, all intersections two at a time, all intersections three at a time, ..., intersection of all or $k$ at a time.

**Question.** Will the probabilities $P(A_1|B), P(A_2|B), \dots, P(A_k|B)$ sum up to 1 when $A_1, \dots, A_k$ constitute a partitioning of the sample space and $B$ is any other event in the same sample space?

**Answer.** Note that

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)}, \quad P(B) \neq 0, \quad j = 1, \dots, k.$$

Hence by taking the sum for $P(B) \neq 0$,

$$P(A_1|B) + \dots + P(A_k|B) = \frac{1}{P(B)}[P(A_1 \cap B) + \dots + P(A_k \cap B)] = \frac{P(B)}{P(B)} = 1$$

since $A_1 \cap B, \dots, A_k \cap B$ are the mutually exclusive partitions of $B$ and their union is $B$ itself. Note that in the above results and procedures $k$ need not be finite. There can be a countably infinite number of events $A_1, A_2, \dots$ in the partition and still the results will hold.

**Question.** Are the procedures of sampling without replacement (taking one by one without putting back the one already taken) and the procedure of taking one subset at random, one and the same in the sense that both the procedures give rise to the same probability statements in whatever the computations that we are going to do?

**Answer.** Yes. Let us look at the above example. [A box contains 10 red and 8 green identical marbles and marbles are taken at random.] What is the probability of getting exactly 2 red and 1 green marbles? Consider the first procedure of taking one subset of 3. Then the 2 red can come from the 10 red in $\binom{10}{2}$ ways and the one green in $\binom{8}{1}$ ways. Let $D$ be the event of getting exactly 2 red and 1 green marbles. Then in the first situation of taking one sample of 3 is given by

$$P(D) = \frac{\binom{10}{2}\binom{8}{1}}{\binom{18}{3}} = \frac{10 \times 9}{2!}\frac{8}{1!}\frac{3!}{18 \times 17 \times 16} = 3\left[\frac{1}{18} \times \frac{9}{17} \times \frac{8}{16}\right].$$

Now, let us consider the second procedure of taking one at a time without replacement. Let $A$ be the event that the first marble is red, $B$ be the event that the second marble is red and $C$ be the event that the third marble is green. Then the intersection $A \cap B \cap C$ is the event of getting the sequence RRG (red, red, green). Then the probability of getting this sequence is given by the following by using the splitting of the intersections with the help of the definition of conditional probabilities:

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) = \frac{10}{18} \times \frac{9}{17} \times \frac{8}{16}$$

because for the first trial there are 18 marbles and out of which 10 are red and the selection is done at random and hence the probability is $\frac{10}{18}$. When one red marble is removed then there are 9 red marbles left. Hence, given $A$, the probability for $B$ or $P(B|A) = \frac{9}{17}$. When two red marbles are removed there are 16 marbles, out of which 8 are green. Then the probability of $C$, given that two red marbles are taken, or given $A \cap B$, is $P(C|A \cap B) = \frac{8}{16}$. Then from the multiplication rule of intersection and conditional statement above the probability for the sequence RRG

$$P(RRG) = P(A \cap B \cap C) = \frac{10}{18} \times \frac{9}{17} \times \frac{8}{16} = \left[\frac{10 \times 9 \times 8}{18 \times 17 \times 16}\right].$$

How many such sequences are there with two reds and one green? RRG, RGR, GRR or three such sequences are there. For each such sequence, we see that the same probability as above appears. Hence the required probability in sampling without replacement scheme is $3[\frac{10}{18} \times \frac{9}{17} \times \frac{8}{16}]$. This is the same result as in the case of taking one subset of 3 from the whole set. Hence the two procedures will lead to the same result. From the steps above, it is seen that this is true in general also.

**Question.** One doubt of the students is that why do we write “zero elsewhere” when writing a probability or density function? Is it necessary?

**Answer.** Since a real random variable is defined over the whole real line, the density should also be defined over the whole real line. Hence the non-zero part is to be mentioned as well as the zero part is to be mentioned.

**Question.** Another serious doubt is whether the boundary point, in the above case of a three-parameter gamma density the lower boundary point $x = \gamma$ is to be included in the non-zero part or in the zero part since the probability at $x = \gamma$ is zero in any case? Should we write the range for the non-zero part as $\gamma < x < \infty$ or as $\gamma \le x < \infty$?

**Answer.** Suppose that $x = \gamma$ is not included in the non-zero part. What is the maximum likelihood estimate (MLE) of the parameter $\gamma$? It does not exist if the point $x = \gamma$ is not included in the non-zero part. It exists and is equal to the smallest order statistic if $x = \gamma$ is included in the non-zero part or if the support is given as $\gamma \le x < \infty$. Since $\infty$ is not a number or a point, the upper boundary point does not arise here.

As another example consider a uniform density:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{elsewhere.} \end{cases}$$

If the non-zero part is written as for $a < x \le b$, then the MLE for the parameter $b$ exists but the MLE for $a$ does not exist. If the support is written as $a \le x < b$, then the MLE for $a$ exists but the MLE for $b$ does not exist. If the range is written as $a < x < b$, then MLE for $a$ and $b$ do not exist. If the range is written as $a \le x \le b$, then the MLE for both $a$ and $b$ exist and they are the smallest order statistic and the largest order statistic, respectively. Hence the range for the non-zero part must be written as $a \le x \le b$ even though the probabilities $x = a$ and $x = b$ are zeros. The rule to be followed is that the non-zero part of the density or probability function has to be written including the boundary points of the support or the interval where the non-zero part is defined.

**Question.** What about the condition $\alpha > 0$, where $\alpha$ is the shape parameter in a gamma density? From where is this condition coming?

**Answer.** The function $x^{\alpha-1}\mathrm{e}^{-\frac{x}{\beta}}$ is a smooth function (no singularities or the function does not become zero in the denominator at any point) if $x$ is away from 0 and $\infty$. When integrating over $[0,\infty)$ when $x \to \infty$, the integral behaves like the integral of $\mathrm{e}^{-\frac{x}{\beta}}, \beta > 0$ since the polynomial part $x^{\alpha-1}$ is dominated by the exponential part $\mathrm{e}^{-\frac{x}{\beta}}$. But when $x \to \infty$, $\mathrm{e}^{-\frac{x}{\beta}}$ goes to zero, which is finite since $\beta > 0$. Hence when $x \to \infty$ there is no problem with the integral. When $x \to 0$, $\mathrm{e}^{-\frac{x}{\beta}} \to 1$. Hence the difficulty may come only from the factor $x^{\alpha-1}$. In the integral, $x^{\alpha-1}$ behaves like $\frac{x^{\alpha}}{\alpha}$. Hence $\alpha \ne 0$. If $\alpha$ is negative, then $x^{\alpha}$ behaves like $\frac{1}{x^{y}}$, $y = -\alpha > 0$ and $\frac{1}{x^{y}}$ goes to $= \infty$ when $x \to 0_{+}$. Hence $\alpha$ is not zero or negative or $\alpha$ must be positive, when real. If $\alpha$ is a complex quantity, then the condition will be $\mathbb{R}(\alpha) > 0$ where $\mathbb{R}(\cdot)$ denotes the real part of $(\cdot)$.

**Question.** How are the conditions $\alpha > 0$ and $\beta > 0$ coming in a beta density with the parameters $\alpha$ and $\beta$?

**Answer.** Take for example a type-1 beta integral. For integrating from $a$ to $b$, $a > 0$, $b < 1$ the integral is smooth, it exists and it has no difficulties, no singularities or the denominator does not become zero at any point. There can be problems when approaching 0 or 1. When approaching zero, $1-x$ will behave like 1 and there is no problem with the second factor in the integrand. Consider the first factor $x^{\alpha-1}$. In the integral it behaves like $\frac{x^{\alpha}}{\alpha}$ and hence $\alpha \ne 0$. By using the same argument as in the gamma integral, $\alpha$ must be positive, if real, otherwise the real part of $\alpha$ must be positive. Now, change $y = 1-x$ or consider the $y$-integral for type-1 beta. Now $\beta$ appears at the place of $\alpha$, and hence $\beta > 0$. Similar arguments can be put forward to show that $\alpha > 0$, $\beta > 0$ in type-2 beta integrals also. [This is left as an exercise to the students.]

**Question.** What exactly is d$x$ the differential of $x$? Is it a small value of $\Delta x$?

**Answer.** Some teachers may have told you that it is small increment in some variable $x$. There is another notation $\Delta x$ for small increment in $x$. If it is small change, then it can be negative or positive or zero.

**Question.** Can d$x$ be zero or negative?

**Answer.** Let $x$ be an independent real variable, independent in the sense that we will be preassigning values to $x$. Let $y = f(x)$ be a dependent variable, dependent on the preassigned values of $x$, through the function $f(x)$. Let $\Delta x$ be a small increment in $x$ and let $\Delta y$ be the corresponding increment in $y = f(x)$. That is, $\Delta y = f(x + \Delta x) - f(x)$. Then $\Delta y = 0$ if $y$ is a constant function of $x$. By convention, we take $\Delta x > 0$ always ( or $\Delta x$ not negative or zero) so that we can talk about increasing and decreasing functions. For $\Delta x > 0$ if $\Delta y < 0$, then the function $y = f(x)$ is decreasing. For $\Delta x > 0$ if $\Delta y > 0$, then $y = f(x)$ is an increasing function of $x$. Hence $\Delta x > 0$ always by convention but $\Delta y$ can be negative, positive or zero. $\mathrm{d}x$ or $\mathrm{d}y$ by itself has no meaning. When $\Delta x$ goes to zero, it goes to zero and it does not by itself become $\mathrm{d}x$ or something else. Small increments are denoted by $\Delta x$ and $\Delta y$ and not by $\mathrm{d}x$ and $\mathrm{d}y$. Then what exactly is this $\mathrm{d}x$? Consider the ratio $\frac{\Delta y}{\Delta x}$. This ratio can always be written because we have assumed that $\Delta x > 0$ and $\Delta x$ is always positive. Then we can write the identity

$$\Delta y \equiv \frac{\Delta y}{\Delta x}\Delta x \tag{i}$$

Consider $\Delta x$ becoming smaller and smaller. If at any stage $\frac{\Delta y}{\Delta x}$ attains a limit, then let this limit be denoted by $f'(x)$. At this stage when the limit is attained, the value of $\Delta x$ is denoted by $\mathrm{d}x$ and the corresponding $\Delta y$ is denoted by $\mathrm{d}y$. Thus we have the identity (it is not any approximation)

$$\mathrm{d}y \equiv f'(x)\mathrm{d}x \quad \Rightarrow \quad f'(x) = \mathrm{d}y \quad \text{divided by } \mathrm{d}x \tag{ii}$$

or $f'(x)$ is a ratio of differentials. Thus, $\mathrm{d}x$, being the differential associated with the independent variable $x$, this $\mathrm{d}x > 0$ by convention. The corresponding dependent variable has the differential $\mathrm{d}y$. This $\mathrm{d}y$ can be negative, positive or zero depending upon the nature of the function.

**Question.** Which one $x$ or $y$ to be taken as the independent variable in the function $2x + 3y - 5 = 0$? This is the same as $x = \frac{1}{2}(5 - 3y)$ or it is also the same as $y = \frac{1}{3}(5 - 2x)$. Then which is the independent variable and which is the dependent variable?

**Answer.** It all depend upon whether we want to calculate $x$ at preassigned values of $y$ or vice versa. If $y$ is preassigned and $x$ is calculated from there then $y$ is the independent variable and $x$ is the dependent variable. If $x$ is preassigned and $y$ is calculated from the preassigned $x$, then $x$ is the independent variable and $y$ is the dependent variable. It will be more clear from the following case. There is a physical law $pv = c$ or pressure multiplied volume is a constant under constant temperature. In the equation, $p$ represents pressure, $v$ volume and $c$ the constant. We can write the equation as $p = \frac{c}{v}$ or as $v = \frac{c}{p}$. If we want to ask the question: what will be the pressure if the volume is 10 cubic centimeters? Then $v$ is preassigned and $p$ is calculated from the formula $p = \frac{c}{v}$. In this case, $v$ is the independent variable and $p$ is the dependent variable. If we want to calculate $v$ at preassigned values of $p$, then $p$ is the independent variable and $v$ is the dependent variable.

**Question.** In an implicit function $f(x_1, x_2, \dots, x_k) = 0$ which is the dependent variable and which are the independent variables?

**Answer.** If our aim is to calculate $x_1$ at preassigned values of $x_2, \dots, x_k$, then $x_2, \dots, x_k$ are independent variables and $x_1$ is the dependent variable. Then the differentials $\mathrm{d}x_2, \dots, \mathrm{d}x_k$ are strictly positive by convention or $\mathrm{d}x_2 > 0, \dots, \mathrm{d}x_k > 0$ and $\mathrm{d}x_1$ could be negative, zero or positive according to the nature of the function $f$.

**Question.** What is the moment problem?

**Answer.** The famous moment problem in physics and statistics is the following: We have seen that if arbitrary moments are available, then the corresponding density is uniquely determined through inverse Mellin transforms under some minor conditions. Suppose that only the integer moments are available, that is, for the single real variable case let $E(x^h)$ for $h = 0, 1, 2, \dots$ are all available, countably infinite number of integer moments are available. Is the density $f(x)$ uniquely determined by these integer moments? This is the famous moment problem in physics and statistics. The answer is: not necessarily. There can be two different density functions corresponding to a given set of integer moments. There are several sets of sufficient conditions available so that a given integer moment sequence will uniquely determine a density/probability function. One such sufficient condition is that the non-zero part of the density is defined over a finite range $[a, b]$, $-\infty < a \le x \le b < \infty$. More sets of sufficient conditions are available; see, for example, the book [15].

This is the case of a single random variable. In the multivariate case, the corresponding problem is that if all integer product moments, that is, $E(x_1^{h_1} \cdots x_k^{h_k})$ where $h_i = 0, 1, 2, \dots$, $i = 1, \dots, k$ are given, will these integer product moments uniquely determine the corresponding multivariate density/probability function? The answer is: not necessarily so!

Let there be a multivariate density/probability function $f(x_1, \dots, x_k)$ and suppose we wish to compute the expected value of a function of the variables, for example, (i) $E(x_1^h)$, (ii) $E(x_1^2 x_2^5)$.

**Question.** How do we compute these types of expected values? In (i), only $x_1$ is involved but we have a multivariate density/probability function. We can compute $E(x_1^h)$ from the marginal density/probability function of $x_1$ but if we compute it by using the multivariate density/probability function will the two procedures give the same results?

**Answer.** This is a usual doubt of the students. By using the multivariate density/probability function, for example, consider a continuous case [the procedure is parallel in the discrete case]:

$$E(x_1^h) = \int_X x_1^h f(x_1,\dots,x_k)\mathrm{d}X$$

where $\int_X = \int_{x_1=-\infty}^{\infty} \cdots \int_{x_k=-\infty}^{\infty}$ and $\mathrm{d}X = \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_k$. Since the function, for which the expected value is to be computed, contains only $x_1$, we can integrate out the other variables $x_2,\dots,x_k$. When $x_2,\dots,x_k$ are integrated out from the joint density, we get the marginal density of the remaining variables, namely the marginal density of $x_1$, denoted by $f_1(x_1)$. Then the integral to be evaluated reduces to

$$E(x_1)^h = \int_{-\infty}^{\infty} x_1^h f_1(x_1)\mathrm{d}x_1.$$

Thus both the procedures give rise to the same result. (ii) Let us take for example, $x_3,\dots,x_k$ to be discrete. In this case,

$$\sum_{x_3=-\infty}^{\infty} \cdots \sum_{x_k=-\infty}^{\infty} f(x_1,\dots,x_k) = f_{12}(x_1,x_2)$$

where $f_{12}(x_1,x_2)$ is the marginal probability or density function of $x_1,x_2$. Now

$$E(x_1^2x_2^5) = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1^2x_2^5 f_{12}(x_1,x_2)$$

if $x_1$ and $x_2$ are both discrete. If they are continuous, then integrate out both, if one is discrete and the other continuous then integrate out the continuous one and sum up the discrete one. Thus one can compute expected values of a function of $r$, $r < k$ of the original $k$ variables then those expected values can be computed either from the joint density/probability function of all the $k$ variables or from the marginal density/probability function of the $r$ variables.

**Question.** Can $\rho$ (correlation between real scalar random variables $x$ and $y$) measure relationship between $x$ and $y$?

**Answer.** The answer is no; it cannot except at the boundary points. In general, we can show that $-1 \le \rho \le 1$. This is easily proved by considering two variables $u = \frac{x}{\sigma_x} + \frac{y}{\sigma_y}$ and $v = \frac{x}{\sigma_x} - \frac{y}{\sigma_y}$ for $\sigma_x \ne 0$, $\sigma_y \ne 0$ (non-degenerate cases). Now, take the variances of $u$ and $v$ and use the fact that for any real random variables $u$ and $v$, $\mathrm{Var}(u) \ge 0$, $\mathrm{Var}(v) \ge 0$. But

$$\mathrm{Var}(u) = \frac{\mathrm{Var}(x)}{\sigma_x^2} + \frac{\mathrm{Var}(y)}{\sigma_y^2} + 2\frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y} = 1 + 1 + 2\rho = 2(1+\rho).$$

Hence $\mathrm{Var}(u) \geq 0 \Rightarrow 2(1+\rho) \geq 0 \Rightarrow \rho \geq -1$. Similarly, $\mathrm{Var}(v) = 2(1-\rho) \geq 0 \Rightarrow \rho \leq 1$. Therefore,

$$-1 \leq \rho \leq 1.$$

From the Cauchy–Schwarz inequality, it follows that $\rho = \pm 1$ or the boundary values if and only $y = ax + b$, where $a \neq 0$, $b$ are constants, *almost surely*. Because of this property for boundary values, people are tempted to interpret $\rho$ as measuring linear relationship which means if $\rho$ is near to +1 or −1, then near linearity is there and if $\rho$ is zero then no linearity is there, etc. We will show that this interpretation is also invalid. Some misuses go to the extent that some applied statisticians interpret positive values, $\rho > 0$, as "increasing values of $x$ go with increasing values of $y$" or "decreasing values of $x$ go with decreasing values of $y$", and $\rho < 0$ means "increasing values of $x$ go with decreasing values of $y$ and vice versa. We will show that this interpretation is also invalid. We will show that no value of $\rho$ in the open interval $-1 < \rho < 1$ can be given any meaningful interpretation, and no interpretation can be given as measuring relationships between $x$ and $y$. Take, for example, $y = a + bx + cx^2, c \neq 0$ and compute the correlation between $x$ and $y$. For convenience, take a symmetric variable $x$ so that all odd moments about the origin will be zeros. For example, take a standard normal variable. Then the correlation coefficient $\rho$ will be a function of $b$ and $c$ only. There are infinitely many choices for $b$ and $c$. By selecting $b$ and $c$ appropriately, all the claims about $\rho$ can be nullified except for the case when $\rho = +1$ or $\rho = -1$. This is left as an exercise to the student.

**Question.** Do the mgf (moment generating function) of type-1 beta density and the corresponding type-1 Dirichlet density exist, because these are not seen in books?

**Answer.** For type-1 beta and type-1 Dirichlet, moment generating function (mgf) exist. For type-2 beta and type-2 Dirichlet mgf do not exist but characteristic functions, $E(\mathrm{e}^{itx})$, $i = \sqrt{-1}$, exist. For the type-1 beta, the mgf, denoted by $M_x(t)$ where $t$ is an arbitrary parameter, is given by

$$M_x(t) = E[\mathrm{e}^{tx}] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \mathrm{e}^{tx} x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x.$$

Here, $\mathrm{e}^{tx}$ can be expanded since term by term integration will be valid and the resulting series is going to be uniformly convergent and the integral is also convergent.

$$\begin{aligned} M_x(t) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}\mathrm{d}x \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+\beta+k)}, \quad \Re(\alpha+k) > 0,\ k = 0, 1, 2, \ldots. \end{aligned}$$

But

$$\Gamma(\alpha+k)=(\alpha)_k\Gamma(\alpha), \quad \Gamma(\alpha+\beta+k)=\Gamma(\alpha+\beta)(\alpha+\beta)_k$$

where, for example, $(a)_k$ is the Pochhammer symbol, given by

$$(a)_k = a(a+1)\cdots(a+k-1), \quad a\neq 0, \quad (a)_0=1.$$

Hence

$$M_x(t)=\sum_{k=0}^{\infty}\frac{t^k}{k!}\frac{(\alpha)_k}{(\alpha+\beta)_k} = {}_1F_1(\alpha;\alpha+\beta;t)$$

where ${}_1F_1$ is the confluent hypergeometric series which is convergent for all finite $t$. Since it is a hypergeometric series, many books avoid the discussion of the mgf here. By using the same procedure, one can evaluate the mgf in the type-1 Dirichlet case, namely,

$$\begin{aligned} M_{x_1,\dots,x_k}(t_1,\dots,t_k) &= E[\mathrm{e}^{t_1x_1+\cdots+t_kx_k}] \\ &= \frac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{k+1})}\int_\Omega \mathrm{e}^{t_1x_1+\cdots+t_kx_k}x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1} \\ &\quad\times(1-x_1-\cdots-x_k)^{\alpha_{k+1}-1}\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_k. \end{aligned}$$

When we expand the exponential part, we can write

$$\mathrm{e}^{t_1x_1+\cdots+t_kx_k}=\sum_{r_1=0}^{\infty}\cdots\sum_{r_k=0}^{\infty}\frac{t_1^{r_1}\cdots t_k^{r_k}}{r_1!\cdots r_k!}x_1^{r_1}\cdots x_k^{r_k}.$$

Then

$$\begin{aligned} &\int_\Omega x_1^{\alpha_1+r_1-1}\cdots x_k^{\alpha_k+r_k-1}(1-x_1-\cdots-x_k)^{\alpha_{k+1}-1}\mathrm{d}x_1\wedge\cdots\wedge\mathrm{d}x_k \\ &= \frac{\Gamma(\alpha_1+r_1)}{\Gamma(\alpha_1)}\cdots\frac{\Gamma(\alpha_k+r_k)}{\Gamma(\alpha_k)} \\ &\quad\times\frac{\Gamma(\alpha_1+\cdots+\alpha_{k+1})}{\Gamma(\alpha_1+\cdots+\alpha_k+r_1+\cdots+r_k+\alpha_{k+1})} \end{aligned}$$

for $\Re(\alpha_j+r_j)>0$, $r_j=0,1,2,\dots;j=1,\dots,k$. Now, the uniformly convergent multiple series above will sum up to a Lauricella function of the type $F_D$; details may be seen from [3].

Let us see what happens in the type-2 beta case if we expand the exponential part and try to integrate term by term. Then the integral to be evaluated is the following:

$$\int_0^\infty x^{\alpha+k-1}(1+x)^{-(\alpha+\beta)}\mathrm{d}x=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\frac{\Gamma(\beta-k)}{\Gamma(\beta)}, \quad -\Re(\alpha)<k<\Re(\beta).$$

Note that $k = 0, 1, 2, \dots$ but $-\Re(\alpha)$ and $\Re(\beta)$ are fixed quantities, and hence the condition for the existence will be violated from some stage onwards. This means that the integral is not convergent or expansion of the exponential part and term by term integration is not a valid procedure here. The moment generating function does not exist for type-2 beta and type-2 Dirichlet cases.

Students have difficulty in evaluating the density $f(x)$ from arbitrary moments by using inverse Mellin transforms. If $\phi(s)$ is $E(x^{s-1})$ for a positive continuous real random variable $x$ with density $f(x)$ and if $\phi(s)$ exists for a complex $s$ and $\phi(s)$ is analytic in a strip in the complex plane then the inverse Mellin transform is given by

$$f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \phi(s)x^{-s}\mathrm{d}s, \quad i = \sqrt{-1}.$$

Detailed conditions for the existence of Mellin and inverse Mellin transforms may be seen from [2]. We will work out the inverse Mellin transform for a known special case here so that the procedure will be clear to the students.

**Example 16.1.** If $\Gamma(s)$ is the Mellin transform of some function $f(x)$ for $x > 0$, then evaluate $f(x)$ by using the formula for the inverse Mellin transform.

**Solution 16.1.** From the integral representation of a gamma function, we know that

$$\Gamma(s) = \int_0^\infty x^{s-1}\mathrm{e}^{-x}\mathrm{d}x, \quad \Re(s) > 0. \tag{a}$$

Thus we know the function $f(x)$ as $\mathrm{e}^{-x}$ from this representation. But we want to recover $f(x)$ from the inverse Mellin transform. The function to be recovered is

$$f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(s)x^{-s}\mathrm{d}s. \tag{b}$$

Since $\Gamma(s)$ has poles at $s = 0, -1, -2, \dots$, then the infinite semicircle $c - i\infty$ to $c + i\infty$ can enclose all these poles if we take any $c > 0$, for example, $c = 0.5$ or $3.8$ etc. From residue calculus, $f(x)$ in (b) is available as the sum of the residues at the poles of the integrand in (b). The residue at $s = -\nu$, denoted by $R_\nu$, is given by

$$R_\nu = \lim_{s\to-\nu}(s+\nu)\Gamma(s)x^{-s}.$$

We cannot substitute $s = -\nu$ and evaluate the limit. We will introduce a few more factors in the numerator and denominator so that the expression becomes simpler.

$$\begin{aligned}R_\nu &= \lim_{s\to-\nu}\frac{(s+\nu)(s+\nu-1)\cdots s}{(s+\nu-1)\cdots s}\Gamma(s)x^{-s}\\ &= \lim_{s\to-\nu}\frac{\Gamma(s+\nu+1)}{(s+\nu-1)\cdots s}x^{-s} = \frac{\Gamma(1)x^\nu}{(-1)^\nu\nu!} = \frac{(-1)^\nu x^\nu}{\nu!}.\end{aligned}$$

Then

$$\sum_{\nu=0}^{\infty} R_\nu = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu x^\nu}{\nu!} = \mathrm{e}^{-x} = f(x).$$

Thus $f(x)$ is recovered through the inverse Mellin transform.

**Question.** Are there other multivariate generalizations of type-1 and type-2 beta densities, other than type-1 and type-2 Dirichlet densities?

**Answer.** Yes, there are other multivariate models available in the literature, for example, see [10]. Now the question comes: for a given density, such as exponential density or normal density, is there anything called the unique multivariate analogue? Students are used to the phrase “the multivariate normal density”. Is it a unique density corresponding to the univariate normal density?

**Question.** In general, given a univariate density/probability function, is there anything called the unique multivariate analogue?

**Answer.** There is no unique analogue. There can be different types of bivariate or multivariate densities where the marginal densities are the given densities. This is evidently obvious but some students have the feeling that the multivariate models are unique.

**Question.** Is a “multivariable or multivariate” distribution the same as “vector-variate” distribution?

**Answer.** Some authors use multivariate case and vector variable case as one and the same. This is not so. There is a clear distinction between “multivariate” and “vector-variate” cases. In the multivariate case, there is no order in which the variables enter into the model or variables could be interchanged in the model with the corresponding changes in the parameters, if any. A vector variable case is a multivariate case, and in addition, the order in which the elements appear also enters into the model. If we have a matrix-variate case and if we are looking at the marginal joint density of a particular row of the matrix, say for example, the first row of the matrix then we have a vector variable case. If we have a function $f(x_1,x_2) = c_1(x_1+x_2)$, $0 \le x_1 \le 1$, $0 \le x_2 \le 1$ and zero elsewhere, where $c_1$ is the normalizing constant, then we can take the variables as $(x_1,x_2)$ or as $(x_2,x_1)$ and both will produce the same function in the same square. Suppose that our function is $c_2(x_1+x_2)$ but defined in the triangle $0 \le x_1 \le x_2 \le 1$ then $(x_1,x_2)$ is different from $(x_2,x_1)$. They cannot be freely interchanged. The ordered set $(x_1,x_2)$ is different from the ordered set $(x_2,x_1)$. Students must make a distinction between “multivariate case” and “vector variable case”. In the latter case, the order in which the variables appear also enters into the model.

**Question.** How is Pearson's $X^2$ statistic coming from a multinomial probability law?

**Answer.** The (mgf) in the multinomial case is the following:

$$\begin{aligned}
&E[e^{t_1x_1+\dots+t_{k-1}x_{k-1}}]\\
&=\sum_{x_1=0}^{n}\cdots\sum_{x_k=0}^{n}e^{t_1x_1+\dots+t_{k-1}x_{k-1}}\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}\\
&=\sum\cdots\sum\frac{n!}{x_1!\cdots x_k!}(p_1e^{t_1})^{x_1}\cdots(p_{k-1}e^{t_{k-1}})^{x_{k-1}}p_k^{x_k}\\
&=(p_1e^{t_1}+\dots+p_{k-1}e^{t_{k-1}}+p_k)^n=M(t_1,\dots,t_{k-1}) \qquad \text{(i)}
\end{aligned}$$

from a multinomial expansion. We can differentiate this multivariate mgf to obtain integer moments. For example, denoting $T=O\Rightarrow t_1=0,\dots,t_{k-1}=0$,

$$\frac{\partial M}{\partial t_j}\Big|_{T=O}=E(x_j),\quad j=1,\dots,k-1\quad\text{and}\quad E(x_k)=n-E(x_1)-\dots-E(x_{k-1})$$

$$\frac{\partial M}{\partial t_j}\Big|_{T=O}=np_j(p_1e^{t_1}+\dots+p_{k-1}e^{t_{k-1}}+p_k)^{n-1}|_{T=O}=np_j$$

for $j=1,2,\dots,k-1$ and $E(x_k)=n-np_1-\dots-np_{k-1}=n(1-p_1-\dots-p_{k-1})=np_k$.

$$\frac{\partial^2 M}{\partial t_i\partial t_j}\Big|_{T=O}=n(n-1)p_ip_j,\quad i\neq j=1,\dots,k-1=E(x_ix_j),\quad i\neq j.$$

$$\frac{\partial^2 M}{\partial t_j^2}\Big|_{T=O}=n(n-1)p_j^2+np_j=E(x_j^2).$$

Hence

$$\mathrm{Var}(x_j)=E(x_j^2)-[E(x_j)]^2=n(n-1)p_j^2+np_j-n^2p_j^2=np_j(1-p_j),$$

for $j=1,\dots,k-1$.

$$\mathrm{Cov}(x_i,x_j)=E(x_ix_j)-E(x_i)E(x_j)=n(n-1)p_ip_j-(np_i)(np_j)=-np_ip_j,$$

for $i\neq j=1,\dots,k-1$.

$$\begin{aligned}
\mathrm{Cov}(x_i,x_k)&=\mathrm{Cov}(x_i,n-x_1-\dots-x_{k-1})\\
&=-\mathrm{Cov}(x_i,x_1)-\dots-\mathrm{Cov}(x_i,x_{k-1})\\
&=np_i(p_1+\dots+p_{i-1}+p_{i+1}+\dots+p_{k-1})-np_i(1-p_i)\\
&=np_i(1-p_k-1)=-np_ip_k.\\
\mathrm{Var}(x_k)&=\mathrm{Var}(n-x_1-\dots-x_{k-1})=\mathrm{Var}(x_1+\dots+x_{k-1})\\
&=\sum_{i=1}^{k-1}\mathrm{Var}(x_i)+2\sum_{i<j=1}^{k-1}\mathrm{Cov}(x_i,x_j)\\
&=np_1(1-p_1)+\dots+np_{k-1}(1-p_{k-1})-2\sum_{i<j=1}^{k-1}np_ip_j.
\end{aligned}$$

But

$$-2\sum_{i<j=1}^{k-1} np_ip_j = -np_1(1-p_1-p_2) - \cdots - np_{k-1}(1-p_{k-1}-p_k)$$
$$= -n(1-p_k) + n(p_1^2 + \cdots + p_{k-1}^2) + np_k(1-p_k).$$

Hence

$$\begin{aligned}\mathrm{Var}(x_k) &= n(p_1 + \cdots + p_{k-1}) - n(p_1^2 + \cdots + p_{k-1}^2) - n(1-p_k)\\ &\quad + n(p_1^2 + \cdots + p_{k-1}^2) + np_k(1-p_k)\\ &= np_k(1-p_k).\end{aligned}$$

Hence

$$\mathrm{Var}(x_i) = np_i(1-p_i), \quad i = 1,2,\dots,k$$

and

$$\mathrm{Cov}(x_i,x_j) = -np_ip_j, \quad i \neq j = 1,2,\dots,k. \tag{ii}$$

The $k \times k$ matrix of variances and covariances is given by

$$V_1 = \begin{bmatrix} np_1(1-p_1) & -np_1p_2 & \dots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & \dots & -np_2p_k \\ \vdots & \vdots & \dots & \vdots \\ -np_kp_1 & -np_kp_2 & \dots & np_k(1-p_k) \end{bmatrix}.$$

This $V_1$ is a singular matrix with determinant of $V_1$, denoted by $|V_1|$, is zero or $|V_1| = 0$. The non-singular covariance matrix in the multinomial case is given by

$$V = \begin{bmatrix} np_1(1-p_1) & -np_1p_2 & \dots & -np_1p_{k-1} \\ -np_2p_1 & np_2(1-p_2) & \dots & -np_2p_{k-1} \\ \vdots & \vdots & \dots & \vdots \\ -np_{k-1}p_1 & -np_{k-1}p_2 & \dots & np_{k-1}(1-p_{k-1}) \end{bmatrix}.$$

The most important quantity associated with $V$ and $V_1$ is Pearson's "goodness-of-fit" statistic $X^2$. The statistic, denoted by $X^2$, is given by

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}. \tag{iii}$$

This can be shown to be the square of a generalized distance between the observed vector $O$ and the expected vector $E$, where

$$O = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_{k-1} \end{bmatrix}, \quad E = \begin{bmatrix} np_1 \\ np_2 \\ \vdots \\ np_{k-1} \end{bmatrix}. \tag{iv}$$

The ordinary Euclidean distance between $O$ and $E$ is $[(O-E)'(O-E)]^{\frac{1}{2}}$. A generalized distance between $O$ and $E$ is available by scaling with the inverse of the square root of the covariance matrix $V$. That is the Euclidean distance between $V^{-\frac{1}{2}}O$ and $V^{-\frac{1}{2}}E$. That is,

$$[(V^{-\frac{1}{2}}O-V^{-\frac{1}{2}}E)'(V^{-\frac{1}{2}}O-V^{-\frac{1}{2}}E)]^{\frac{1}{2}}=[(O-E)'V^{-1}(O-E)]^{\frac{1}{2}}. \tag{v}$$

Hence the square of this generalized distance is $(O-E)'V^{-1}(O-E)$. This quantity should be approximately a chi-square with $k-1$ degrees of freedom according to the central limit theorem, see the note below which explains why this is chi-square distributed. Note that we can write $V$ as follows:

$$V=nD[I-JJ'D]=nD[D^{-1}-JJ']D,\quad D=\begin{bmatrix} p_1 & 0 & \dots & 0\\ 0 & p_2 & \dots & 0\\ \vdots & \vdots & \dots & \vdots\\ 0 & 0 & \dots & p_{k-1}\end{bmatrix},\quad J=\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}. \tag{vi}$$

It can be shown that

$$V^{-1}=n^{-1}\left[D^{-1}+\frac{JJ'}{p_k}\right] \tag{vii}$$

by inverting $D^{-1}-JJ'$ with the help of elementary transformations. Let $(O-E)'=(n_1-np_1,\dots,n_{k-1}-np_{k-1})$. Then

$$(O-E)'V^{-1}(O-E)=(O-E)'\left(\frac{D^{-1}}{n}\right)(O-E)+(O-E)'\left(\frac{JJ'}{np_k}\right)(O-E)$$

where

$$(O-E)'\left(\frac{D^{-1}}{n}\right)(O-E)=\sum_{j=1}^{k-1}\frac{(n_j-np_j)^2}{np_j}$$

and

$$\frac{1}{np_k}(O-E)'JJ'(O-E)=\frac{(n_k-np_k)^2}{np_k}.$$

Therefore, adding the above two terms we have

$$(O-E)'V^{-1}(O-E)=\sum_{j=1}^{k}\frac{(n_j-np_j)^2}{np_j}=X^2\quad(\text{Pearson's }X^2)\to\chi^2_{k-1}. \tag{viii}$$

## 16.2 Questions and answers on model building

**Question.** Is regression the same as least square analysis?

**Answer.** Definitely not. In a large number of books, regression is interpreted as least square estimation and often start to define regression as least square estimation of model building. Least square estimation is a procedure of model building, introduced by Gauss. The basic principle there is to minimize the square of the Euclidean distance between the observed value and the value estimated by the model and then fit the model in hand to the data at hand. If someone wishes to fit a second degree polynomial $y = a + bx + cx^2$, $c \neq 0$ to paired observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, then corresponding to $x_j$, an observation on $x$, there is an observed value of $y$, denoted by $y_j$. Naturally, $y_j$ need not be equal to $a + bx_j + cx_j^2$ unless $y = a + bx + cx^2$ is a mathematical relationship or the model selected is a perfect fit for all possible pairs of observations $(x_j, y_j)$. In reality, the selected model $y = a + bx + cx^2$ is taken as a possible behavior of the data in hand. Hence, naturally, there is an error which may be taken as $e_j = y_j - (a + bx_j + cx_j^2)$, the difference between the observed and modeled value of $y$.

One question the students ask is: can we take error as $(a + bx_j + cx_j^2) - y_j = e_j$ instead of taking the other way around? The answer is "yes" because we are going to consider only the distance and hence both will lead to the same answers at the end.

Here, the unknown quantities are $a, b, c$ because an arbitrary second degree polynomial is selected to fit the data in hand. The error sum of squares is then

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} [y_j - (a + bx_j + cx_j^2)]^2.$$

This is the squared Euclidean distance between the observed $y_j$ and the modeled $a + bx_j + cx_j^2$. If $a, b, c$ are estimated by minimizing the error sum of squares then the method is the method of least squares or the sum of squares of the error is made a minimum. Minimization can be done in many ways. If calculus is used, then consider the equations:

$$\frac{\partial}{\partial a} \sum_{j=1}^{n} e_j^2 = 0 \quad \Rightarrow \quad \sum_{j=1}^{n} (y_j - \hat{a} - \hat{b}x_j - \hat{c}x_j^2) = 0 \tag{i}$$

$$\frac{\partial}{\partial b} \sum_{j=1}^{m} e_j^2 = 0 \quad \Rightarrow \quad \sum_{j=1}^{n} (x_j y_j - \hat{a}x_j - \hat{b}x_j^2 - \hat{c}x_j^3) = 0 \tag{ii}$$

$$\frac{\partial}{\partial c} \sum_{j=1}^{n} e_j^2 = 0 \quad \Rightarrow \quad \sum_{j=1}^{m} (x_j^2 y_j - \hat{a}x_j^2 - \hat{b}x_j^3 - \hat{c}x_j^4) = 0. \tag{iii}$$

Note that (i), (ii), (iii) do not hold for all possible values of $a, b, c$ but they hold only at a point or points where $(\hat{a}, \hat{b}, \hat{c})$ satisfies all the equations (i), (ii), (iii). A solution of (i), (ii), (iii) for $(a, b, c)$ is denoted as $(\hat{a}, \hat{b}, \hat{c})$. Note that in (i), (ii), (iii) all $x_j$'s and $y_j$'s are data or known numbers and the only unknown quantities are $\hat{a}, \hat{b}, \hat{c}$.

**Question.** Should we have used $a, b, c$ in equations (i), (ii), (iii), usually that is done in many books, rather than putting hat for $a, b, c$ in these equations?

**Answer.** If we write the equations (i), (ii), (iii) without hat, then it means that the equations hold for all possible values of $a, b, c$, which is incorrect, and hence a proper way of writing the equations is with the hat for the unknown quantities $a, b, c$ to indicate specific values for which the equations hold.

A solution of (i), (ii), (iii) is called a *critical point* and the equations (i), (ii), (iii) are called *normal equations*.

**Question.** Why are they called normal equations, is there any normal distribution involved?

**Answer.** As far as this author knows, there is no normality involved and someone called the minimizing equations in least square procedure as normal equations and from then on they are called normal equations. The points corresponding to a local maximum or local minimum or saddle point is called a *critical point* when a calculus procedure is used.

In a general situation where one has a model $y = g(x_1, \dots, x_r, a_1, \dots, a_k)$ to be fitted to a given data where $y$ and $(x_1, \dots, x_r)$ are going to be observed, and hence they will be numbers eventually and the only parameters or the unknown quantities will be $a_1, \dots, a_k$. In our $g$, there is $g(x, a, b, c)$ is $a + bx + cx^2$ or $y = a + bx + cx^2$. In the general case, the normal equations are the following:

$$\frac{\partial}{\partial a_j}\sum_{j=1}^{n} e_j^2 = \frac{\partial}{\partial a_j}\sum_{j=1}^{n}[y_j - g(x_1, \dots, x_r, a_1, \dots, a_k)]\frac{\partial g}{\partial a_j} = 0, \quad j = 1, \dots, k.$$

If the unknowns $a_1, \dots, a_k$ are such that $g$ is a linear function in the unknowns $a_1, \dots, a_k$ then the procedure is called *linear least square procedure*. Note that $y$, $x_1, \dots, x_r$ are going to be observed and hence they will be numbers eventually, and hence if $g$ is a linear or non-linear function of $x_1, \dots, x_r$ it is still a linear least square problem as long as $g$ is linear in $a_1, \dots, a_k$. In our model, the function $g = a + bx + cx^2$ which is linear in $a, b, c$ but non-linear in $x$ but the model is a linear model because it is linear in the unknowns $a, b, c$. Students ask several questions in this connection.

**Question.** In least square analysis, usually we compute the critical points from the normal equation and construct the model. Why are we not computing the second-order derivatives and the matrix of second-order derivatives evaluated at a critical point to check for maxima/minima and instead of doing this why do we stop at critical points and claim that we have minimized the sum of squares?

**Answer.** Since the parameters $a_1, \dots, a_k$ are arbitrary, we can set them arbitrarily large. For example, in our example we can set $a, b, c$ arbitrarily large and, therefore, naturally, the maximum of $\sum_{j=1}^{n} e_j^2$ is at $+\infty$, and hence the critical point, usually unique, will correspond to a minimum.

**Question.** In least square and other model building situations, why are $x_1, \dots, x_r$ called *independent variables* and $y$ the *dependent variable*? What are they independent of?

**Answer.** *Independent* is a little unfortunate term. What it means is that we assign values to $x_1, \dots x_r$ or at given values of $x_1, \dots, x_r$ we wish to estimate $y$. In this sense, $x_1, \dots, x_r$ are called independent variables and $y$ the dependent variable. In $y - 2x - 3 = 0$, which one is independent variable and which is dependent variable because we can write this equation as $y = 2x + 3$ or as $x = \frac{1}{2}(y - 3)$. Independent and dependent variables do not depend on how we write the equation. It depends upon how are we going to use the equation. If we are going to evaluate $y$ at given value of $x$, then in this case $x$ is an independent variable and $y$ is the dependent variable or if we are going to evaluate $x$ at given $y$ then in this case $y$, is the independent variable and $x$ is the dependent variable. In the general model, $y = g(x_1, \dots, x_r, a_1, \dots, a_k)$ we are going to preassign values to $x_1, \dots, x_r$ and observe the corresponding value of $y$. Hence $x_1, \dots, x_r$ will be the independent variables here and $y$ is the only dependent variable.

**Question.** In almost all books on model building, a simple example is given claiming how to convert a non-linear model into a linear model by taking the model example as $y = ab^x$, then take logarithms and write as $Y = A + Bx$ where $Y = \ln y$, $A = \ln A$, $B = \ln b$. Can we convert non-linear models to linear models by such a procedure?

**Answer.** Unfortunately, the above procedure is incorrect. For taking logarithms, $y_j$'s must be positive since we are dealing with real numbers. Suppose that $y_j$'s are positive, $a, b$ are assumed to be positive. Still the procedure is incorrect. What happens to the error? In this case, $y_j = ab^{x_j} + e_j$, $j = 1, \dots, n$. We cannot write the logarithm of the sum on the right side and write as sum of the logarithms. Suppose that the error is entering into the model as a product such as $y_j = ab^{x_j} e_j$ then can we take logarithms and proceed? The meaning of error is that it can be a positive or negative quantity. If one can guarantee that in certain problems the error is always a positive number and the error enters into the model as a product and the model is of the form $ab^x$, then one can take logarithms. This is not a usual practical situation. The natural thing to do in a model such as $y = ab^x$ is to apply non-linear least square analysis.

**Question.** Then what is regression, if not least square analysis?

**Answer.** Regression is a prediction problem and it is not an estimation problem. We are trying to predict one variable, say $y$, by using other variables such as $x$ or $x_1, \dots, x_r$, using in the sense that what is the predicted value of $y$ at preassigned values of $x_1, \dots, x_r$? What is the "best" predictor function of $x$, "best" in some sense? We can use any arbitrary function $\phi(x)$, if there is only one variable $x$, to be used to predict, as a predictor function. But the predicted value $\phi(x)$ for $y$ and the true value of $y$ may be far apart. We would like to have $\phi(x)$ agreeing with $y$ whatever be the value of $y$.

This is not possible in a real-life situation unless there is a physical law behind it so that there is a mathematical relationship between $x$ and $y$. In the absence of a mathematical relationship, the best thing to do is to minimize a distance between $y$ and $\phi(x)$ and select a $\phi$. Then this $\phi$, which minimizes a "distance" between $y$ and $\phi(x)$, can be called the "best predictor". We can construct various measures of distance between $y$ and an arbitrary predictor function $\phi(x)$. A convenient squared distance is $E|y-\phi(x)|^2$ where $E$ denotes the expected value. $\phi(x)$ at given $x$ will be a constant, say $a$. Then the question is what is $a$ such that $E|y-a|^2$ is a minimum? We already know the answer to this. It is the expected value of $y$ at that preassigned value of $x$ or it is $E(y|x)$ = the conditional expectation of $y$ at preassigned value of $x$. Hence this conditional expectation is defined as the "best" predictor, "best" in the *minimum mean square sense* or the mean value or expected value of the squared Euclidean distance is minimized.

**Definition** (Regression of $y$ on $x$). It is defined as $E(y|x)$ or the conditional expectation of $y$ at preassigned value of $x$, if only one variable $x$ is used, and it is $E(y|x_1,\dots,x_r)$ if many variables $x_1,\dots,x_r$ are used for predicting $y$.

The reason for defining it like this is explained above, that it is the "best" predictor of $y$ "best" in the minimum mean square sense. Students usually ask the following question:

**Question.** Regress means to go back. Are we going back to something or why the word regression is used for the best predictor?

**Answer.** By this process, we are not going back to anything. It is simply the best predictor, best in the minimum mean square sense. But originally the problem was to study characteristics of offsprings and to say something about the parents or previous generation. Thus the original problem was a problem of going back. Nowadays, we are not using it for going back but using it to come with the best predictor whatever be the situation.

**Question.** Do we need to know the joint distribution of $x_1,\dots,x_k$ in order to construct the best predictor $E(x_1|x_2,\dots,x_k)$?

**Answer.** If the joint distribution is known and if the conditional expectation exists, then we can compute $E(x_1|x_2,\dots,x_k)$ or if the conditional distribution of $x_1$, given $x_2,\dots,x_k$, is known then also we can construct $E(x_1|x_2,\dots,x_k)$ provided it exists. Hence, we should know at least the conditional distribution, if not the joint distribution.

**Example 16.2.** From the joint density,

$$f(x,y)=\frac{1}{2\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-1-2x-3x^2)^2},\quad -\infty<y<\infty,\ 0\le x\le 2$$

and zero elsewhere, compute the best predictor of $y$ at given values of $x$.

**Solution 16.2.** In order to get the marginal density of $x$, integrate out $y$. But $\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-\mu)^2}\mathrm{d}y = 1$, $\mu = 1 + 2x + 3x^2$. Hence the marginal density of $x$ is $f_1(x) = \frac{1}{2}$, $0 \leq x \leq 2$ and zero elsewhere. Dividing $f(x,y)$ by this $f_1(x)$ we get the conditional density of $y$, given $x$, as

$$g_2(y|x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-\mu)^2}, \quad \mu = 1 + 2x + 3x^2.$$

Hence from this conditional normal density, $E(y|x) = 1 + 2x + 3x^2$ which is the best predictor of $y$ at given values of $x$ in this case. Note that here it is a non-linear function of $x$.

**Question.** It is said that linear regression uniquely determines a normal or Gaussian density. In the above example, we get a non-linear function of $x$, namely $1 + 2x + 3x^2$ as the regression of $y$ given $x$. Is there a contradiction here?

**Answer.** It is known that if $y$ and $x$ are jointly normally distributed then the regression of $y$ on $x$, and $x$ on $y$, are linear and the linear functions are the following:

$$E(y|x) = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1) \tag{16.1}$$

where $E(x) = \mu_1$, $E(y) = \mu_2$, $\mathrm{Var}(x) = \sigma_1^2$, $\mathrm{Var}(y) = \sigma_2^2$ and the correlation between $x$ and $y$ is $\rho$, and

$$E(x|y) = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y - \mu_2). \tag{16.2}$$

Also, linear regression, under some additional conditions characterize joint normality for $x$ and $y$; see the books on characterizations or see, for example, [11]. In Example 16.2, $x$ and $y$ are not jointly normally distributed. There, only the conditional density of $y$, given $x$, is normal with conditional expectation of $y$, given $x$, non-linear. We can have the conditional density of $y$ normal with all sorts of conditional expected values, linear or non-linear. For computing $E(y|x)$, we need only the conditional density of $y$, given $x$, and the joint density is not necessary. Hence there is no contradiction here.

**Question.** There is a theorem which says that $E(y) = E[E(y|x)]$. Is it true for all types of variables $x$ and $y$ as long as there is a conditional distribution of $y$, given $x$, and a marginal distribution of $x$, in the light of the answer to the previous question saying that all sorts of $E(y|x)$ can be there?

**Answer.** The theorem $E(y) = E[E(y|x)]$ does not hold everywhere. In some examples, the theorem will hold and in others it need not hold. Consider the following example.

**Example 16.3.** Evaluate $E(y)$, if possible, from the following joint density:

$$f(x,y) = \frac{1}{x^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-1-x)^2}, \quad -\infty < y < \infty, 1 \le x < \infty$$

and zero elsewhere.

**Solution 16.3.** Integrating out $y$ from the joint density we get the marginal density of $x$ here as

$$f_1(x) = \begin{cases} \frac{1}{x^2}, & 1 \le x < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

Here, as well as in the previous examples in this section, we cannot integrate out $x$ to get an analytic expression for the density of $y$. But we can try to evaluate $E(y)$ by using the theorem

$$E(y) = E[E(y|x)]. \tag{16.3}$$

Here, $E(y|x) = 1 + x$, and hence $E[E(y|x)] = E(1 + x) = 1 + E(x)$. Note that

$$E(x) = \int_1^\infty x\frac{1}{x^2}\mathrm{d}x = \int_1^\infty \frac{1}{x}\mathrm{d}x = [\ln x]_1^\infty = \infty$$

or does not exist. Hence $E(y)$ cannot be computed by using the above theorem. That theorem was valid only when all the expected values existed.

If the joint distribution or the conditional distribution is unknown, then the regression of $x_1$ on $x_2, \ldots, x_k$ cannot be evaluated. In this case, if we have some idea about the conditional expectation, such as it is a linear function of the conditioned variables $x_2, \ldots, x_k$ or some other specific form then we go for estimation of the regression function, where we use the method of least squares. Thus least square analysis comes in for estimating the regression function when the conditional distribution is not known. If the conditional expectation is suspected to be linear, that is,

$$E(x_1|x_2, \ldots, x_k) = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{i}$$

and if the conditional distribution of $x_1$, given $x_1, \ldots, x_k$ is unknown then we set up a model of the type

$$x_{1j} = a_1 + a_2 x_{2j} + \cdots + a_k x_{kj} + e_j, \quad j = 1, \ldots, n \tag{ii}$$

and try to estimate $a_1, a_2, \ldots, a_k$ so that the regression function can be estimated by using the estimated values $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k$ and substituting $\hat{a}_j$ for $\beta_j$ in (i). The observation matrix $x_1, x_2, \ldots, x_k$ is of the following form:

$$\begin{bmatrix} x_{11} & x_{21} & \ldots & x_{k1} \\ x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \ldots & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix}$$

Here, $X = (x_{ij})$. If the column averages are denoted by $\bar{x}_i = \frac{\sum_{j=1}^n x_{ij}}{n}$, $i = 1, \dots, k$. If $a_1$ is eliminated from the model by subtracting the averages, then we end up with the matrix corresponding to $X$ as

$$\tilde{X} = X - \bar{X}, \quad \bar{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_k \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_k \\ \vdots & \vdots & \dots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_k \end{bmatrix} \quad \Rightarrow \quad S = \tilde{X}'\tilde{X} = (X - \bar{X})'(X - \bar{X})$$

where $S = (s_{ij})$, $s_{ij} = \sum_{r=1}^n (x_{ir} - \bar{x}_i)(x_{jr} - \bar{x}_j)$. Then we may partition $S$ as

$$S = \begin{bmatrix} s_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \quad s_{11} \quad \text{is } 1 \times 1 \quad \text{and} \quad S_{22} \quad \text{is } (k-1) \times (k-1).$$

**Question.** Why are $s_{ij}$'s called "corrected" sum of products in some books? Was there a mistake somewhere in the procedure?

**Answer.** The phrase "corrected" is used to indicate that deviations from the respective averages $\bar{x}_i$, $i = 1, \dots, k$ are taken. That is all. There is no mistake anywhere.

**Question.** Can we take $S_{22}$ to be non-singular?

**Answer.** In a regression type model building situation, the $x_{ij}$'s are preassigned quantities for $i = 2, \dots, k$ and $j = 1, \dots, n$. When we preassign $v_j = (x_{2j}, \dots, x_{kj})$ for a specific $j$, we are not going to preassign again a multiple of this vector or a linear function of the points already selected because no additional information will be forthcoming. Hence we may assume, without loss of generality, that $S_{22}$ to be non-singular. If $x_{ij}$'s are coming from a design type model- then the $x_{ij}$'s are determined by the design of the experiment where usually $X$ will be a less than full rank matrix making $S$ singular.

Then the sample multiple correlation, denoted by $R_{1.2\dots k}$, is defined as

$$R^2_{1.2\dots k} = \frac{S_{12}S_{22}^{-1}S_{21}}{s_{11}}, \quad 1 - R^2_{1.2\dots k} = \frac{|S|}{|S_{22}|s_{11}} \tag{a}$$

parallel to the corresponding population quantities.

**Question.** Why do we take the form in (a)? Is it because the regression may be linear all the time?

**Answer.** If the regression is known to be linear, then we have proved in Chapter 14 that the population multiple correlation coefficient $\rho_{1.2\dots k} = \frac{\sqrt{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}}{\sqrt{\sigma_{11}}}$ is the maximum correlation in the class of linear predictors for $x_1$, based on $x_2, \dots, x_k$. If the regression is known to be non-linear then we have a corresponding measure of multiple correlation ratio and then that can be used. If nothing is known, whether the regression is linear or non-linear, then $\rho_{1.2...k}$ is a convenient quantity to calculate, and hence $\rho_{1.2...k}$ and the corresponding sample value $R_{1.2...k}$, given in (a), are used in practical situations.

**Question.** Is the sample multiple correlation coefficient a good measure to use to check the "goodness" of a model, the procedure that is usually done in practice?

**Answer.** No. It is not an indicator of the 'goodness" of the model in most practical situations. In Section 14.6.2, it is seen that the multiple correlation coefficient keeps on increasing if more and more variables are included, the variables themselves may not have any relevance in estimating $x_1$. If variables $x_1, x_2, \dots, x_k$ are jointly normally distributed, then one can check whether the sample multiple correlation $R^2_{1.2...k}$ is "significantly large" or not. In a practical situation, joint normality may not be there and it is difficult to check for joint normality also. This author suggests to use $s^2$ = the least square minimum, to check for the "goodness" of the model. $s^2$ is a squared distance between the observed $x_{1j}$ and the estimated $x_{1j}$, estimated by the model. Hence the criterion "smaller the distance better the model" can be used. At each stage compute $s^2$ and if $s^2$ increases or remains steady, then remove the new variable introduced. If $s^2$ decreases, then proceed and stop the process when the decrease in $s^2$ is insignificantly small.

**Question.** Usually, the hypothesis $H_0 : c_j = 0$ is tested by using a Student-t test to delete or retain $c_j$ in the model. Is it a proper procedure?

**Answer.** A Student-t statistic arises from the normality assumption, that is, the assumption that $x_1$, at given $x_2, \dots, x_k$, is conditionally normal or the conditional density is a normal density. There are many characterization theorems available to characterize or uniquely determine a normal distribution. In the light of these, one can analyze the practical situation at hand and decide whether a normality assumption is reasonable. If reasonable, then a Student-t test can be used. But $s^2$, the lest square minimum, avoids all such distributional problems.

If we are estimating $x_1$ by using $x_2$ assuming a simple linear model, then the model will be of the form

$$x_{1j} = c_0 + c_2 x_{2j} + e_j. \quad j = 1, \dots, n$$

and the estimated linear function is of the following form:

$$x_1 = \bar{x}_1 + \frac{s_{12}}{s_{22}}(x_2 - \bar{x}_2) \quad \Rightarrow \quad x_1 = \bar{x}_1 + \frac{r\sqrt{s_{11}}}{\sqrt{s_{22}}}(x_2 - \bar{x}_2$$

where $r$ represents the sample correlation coefficient.

**Question.** Suppose that we wish to test a hypothesis such as $c_2 = 0$ or construct a confidence interval for $c_2$. Is it equivalent to testing the hypothesis $\rho = 0$ the corresponding population correlation coefficient, or constructing confidence interval for $\rho$ and then converting it for $c_2$?

**Answer.** Unfortunately, the procedures are not equivalent. Decision making on the coefficient $c_2$ can be done in the conditional space at given values of $x_2$ whereas inference on $\rho$ will require the joint distribution of $x_1$ and $x_2$, conditional distribution is not sufficient. Hence the two procedures have two different premises and they are different. One can be done in the conditional space but the other needs the entire space or joint space.

**Question.** In the model building situations usually we assume the errors $e_j$'s to be normally distributed. Why assume normality? Are not the errors normally distributed according to Gauss?

**Answer.** If errors satisfy some basic conditions such as (i) $e_j$ is contributed by infinitely many unknown factors, all independently contributing and such contributions are infinitesimally small; (ii) contribution of each such factor can be positive or negative with probability $\frac{1}{2}$ each; (iii) the total variance, $\mathrm{Var}(e_j)$ is a finite quantity $\sigma^2$. Under these conditions, it can be mathematically derived that the error will be normally distributed with expected value zero and variance $\sigma^2$ or in such a situation $e_j \sim N(0,\sigma^2)$, $j=1,\dots,n$ and independently distributed. Hence normality is a reasonable assumption. But there are situations where the errors may not follow normal distributions.

**Question.** If we know beforehand that $y$ has a symmetric distribution so that $E(y)=0$, then should we take the model as $y_j = c_1x_{1j} + \cdots + c_kx_{kj} + e_j$, $j=1,\dots,n$ or including $c_0$ also?

**Answer.** When we preassign $x_1 = 0,\dots,x_k = 0$ if all observations on $y$ are zeros, then we can take the model with $c_0 = 0$. If that does not happen, usually in a practical situation this does not happen, then take the model with $c_0$ in and eliminate $c_0$ by the procedure described below and continue with the analysis. The error sum of squares is $\sum_{j=1}^n e_j^2$. Then the equation $\frac{\partial}{\partial c_0}\sum_{j=1}^n e_j^2 = 0$ gives

$$c_0 = \bar{y} - c_1\bar{x}_1 - \cdots - c_k\bar{x}_k, \quad \bar{x}_i = \frac{\sum_{j=1}^n x_{ij}}{n}, \quad i=1,\dots,k.$$

If we substitute for $c_0$, then we have the model

$$y_j - \bar{y} = c_1(x_{1j} - \bar{x}_1) + \cdots + c_k(x_{kj} - \bar{x}_k) + e_j, \quad j=1,\dots,n.$$

Then the normal equations become the following:

$$(X-\bar{X})'(X-\bar{X})\hat{\beta} = (X-\bar{X})'(Y-\bar{Y}) \tag{16.4}$$

where

$$\bar{Y} = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}, \quad \bar{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_k \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_k \\ \vdots & \vdots & \dots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_k \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \vdots \\ \hat{c}_k \end{bmatrix}, \quad X - \bar{X} \quad \text{is } n \times k.$$

Then, under non-singularity for $(X - \bar{X})'(X - \bar{X})$ we have the solution:

$$\hat{\beta} = [(X - \bar{X})'(X - \bar{X})]^{-1}(X - \bar{X})'(Y - \bar{Y}) \tag{16.5}$$

and the least square minimum

$$s^2 = (Y - \bar{Y})'\{I - (X - \bar{X})[(X - \bar{X})'(X - \bar{X})]^{-1}(X - \bar{X})'\}(Y - \bar{Y})$$

with $E(s^2) = (n - k)\sigma^2$ and under normality $\frac{s^2}{\sigma^2} \sim \chi^2_{n-k}$ and an unbiased estimator for $\sigma^2$ is $\frac{s^2}{n-k}$ or $\hat{\sigma}^2 = \frac{s^2}{n-k}$ = unbiased estimator for $\sigma^2$. Use this $\hat{\sigma}^2$ for constructing Student-t for testing hypotheses on individual parameters $c_1, \dots, c_k$. In this case, the degrees of freedom is $n - k$, not $n - (k + 1)$.

## 16.3 Questions and answers on tests of hypotheses

**Question.** Why $H_0$ is called the "null" hypothesis? Is there anything made empty or zero by this hypothesis?

**Answer.** Nothing is made zero or empty. It is simply the hypothesis being tested. The word "null" came due to historical reasons. Testing was originally developed for agricultural experiments, where the experimenter liked to make a claim that the expected yield under a particular fertilizer was different from that of another fertilizer, that is, a hypothesis of the type $\mu_1 \neq \mu_2$ where $\mu_1$ and $\mu_2$ are the expected values. But open statements cannot be tested, which will be seen later, and hence the hypothesis is made as $H_0 : \mu_1 = \mu_2$ and tested against $H_1 : \mu_1 \neq \mu_2$. Here, the negation is tested to say something about the hypothesis of interest. Hence "no difference" brought in the term "null". Nowadays there is no such meaning. $H_0$ simply means the hypothesis that is being tested.

**Question.** In many books for testing a hypothesis such as $\mu_1 > \mu_2$ it is suggested to take $H_0 : \mu_1 = \mu_2$ and test it against $\mu_1 > \mu_2$. Can we take alternate as we please or according to convenience?

**Answer.** No. The procedure is incorrect unless it is guaranteed that either $\mu_1 = \mu_2$ or $\mu_1 > \mu_2$ and nothing else is possible, which is not a practical situation. When

$H_0 : \mu_1 = \mu_2$, then the natural alternate is $H_1 : \mu_1 \neq \mu_2$. But if someone wishes to make a claim $\mu_1 > \mu_2$, then take $H_0 : \mu_1 \leq \mu_2$ and test it against the natural alternate $H_1 : \mu_1 > \mu_2$. This is a valid procedure. The null and alternate hypotheses must cover the entire parameter space. Hence one cannot select alternate according to convenience. Some times, a wrong procedure can lead to the same test criterion obtained through the correct procedure. This will be explained when we talk about test criteria.

**Question.** Is there a practical situation where both $H_0$ and $H_1$ are simple?

**Answer.** Usually not, but there can be situations where there are only two points in the parameter space. Consider a machine automatically filling 10 kg sugar bags. The expected weight of each bag is 10 but there can be slight variations from bag to bag because the machine is not counting sugar crystals, probably the machine is timing it or automatic weighing process is there. Suppose that a machine is automatically filling 20 kg sacks with coconuts, where the machine is not allowed to cut or chop any coconut. Then the variation from bag to bag will be substantial up to the weight of one coconut. In both of these examples, at one case $E(x) = 10$ = expected weight and in the other case $E(x) = 20$. Suppose that the sugar filling machine went out of control for a few minutes so that the setting went to 9.5 instead of the 10 before it was recognized and corrected, or a dishonest wholesaler set the machine purposely at 9.5 for some time. All bags are sent to the market. There are only two possibilities here, either the expected value is 10 or the expected value is 9.5 and nothing else or the parameter space has only two points 10 and 9.5, and hence a simple $H_0$ versus a simple $H_1$ is also possible or makes sense here.

**Question.** In many books, it is written "reject $H_0$, accept $H_0$". Is it non-rejection equivalent to accepting the hypothesis?

**Answer.** You will see later that the whole testing procedure is geared to rejecting $H_0$. If $H_0$ is not rejected, then nothing can be stated logically. Testing is carried out by using one data set. If $H_0$ is not rejected in one data set, there is no guarantee that $H_0$ is not rejected in all data sets. We have several real-life examples. The drug thalidomide was tested on mice, rabbits and several types of animals and the hypothesis that the drug was good or effective and safe was not rejected. The drug was introduced into the market and resulted in a large number of deformed babies and the drug was banned eventually. If $H_0$ is rejected in one data set, we are safe and justified in rejecting $H_0$ because it is rejected at least in one data set. Also, these are not mathematical statements. When we reject $H_0$ we are not saying that $H_0$ is not really true. It says only that as per the testing procedure used and as per the data in hand, $H_0$ is to be rejected. Logical way of formulating the decision is "reject $H_0$" and "not reject $H_0$". The question of acceptance does not arise anywhere in the testing procedure.

**Question.** In many books, it is stated “independent observations are taken”. Observations are numbers and how can we associate statistical independence to numbers?

**Answer.** A simple random sample or often called a sample, is a set of independently and identically distributed random variables or (iid) variables, not numbers. Suppose that $\{x_1, \dots, x_n\}$ be the sample of size $n$, where $x_1, \dots, x_n$ are iid variables and not numbers. If we take one observation on $x_1$, one observation on $x_2$, ..., one observation on $x_n$ then we say we have $n$ independent observations. “‘Independent observations” is used in this sense.

**Question.** Is the likelihood ratio test uniformly most powerful test (UMPT) in the light of the Neyman–Pearson lemma?

**Answer.** The Neyman–Pearson lemma is applicable in the case of simple $H_0$ versus simple $H_1$, which can be stretched to simple $H_0$ versus composite $H_1$, as illustrated in the worked example. But for composite $H_0$ versus composite $H_1$ there is no guarantee that we get the UMPT. In most of the problems that we consider in this book, we have UMPT.

**Question.** Is maximizing $L$, the likelihood function, the same as maximizing $\ln L$ and vice versa?

**Answer.** Yes. As long as $\phi(L)$ is a one to one function of $L$, then $\frac{\partial}{\partial\theta}L = 0 \Rightarrow \frac{\partial}{\partial\theta}\phi(L) = 0$.

**Question.** Instead of differentiation with respect to $\sigma^2$, as was done in the Gaussian case, if we had differentiated with respect to $\sigma$ do we get the same estimate for $\sigma^2$ as $s^2$?

**Answer.** Yes. We would have got the same answer. The reason being

$$\frac{\partial}{\partial\sigma}[\cdot] = 0 \quad \Rightarrow \quad \frac{\partial}{\partial\sigma^2}[\cdot]\frac{\partial}{\partial\sigma}\sigma^2 = 2\sigma \times \frac{\partial}{\partial\sigma^2}[\cdot] = 0$$

or both will lead to the same solution. In fact, for any non-trivial function $\psi(\sigma)$ for which $\frac{\partial\psi}{\partial\sigma} \neq 0$ the maximum likelihood estimate (MLE) of $\psi(\sigma)$ is $\psi(\hat{\sigma})$ where $\hat{\sigma}$ is the MLE of $\sigma$.

**Question.** In the maximum likelihood procedure, should we not have taken the second-order derivatives and checked for the definiteness of the matrix of second-order derivatives evaluated at the critical points to check for maxima/minima? Why stop at finding the critical points and claiming that the maximum occurs at the critical point?

**Answer.** Yes. We should have taken the second-order derivatives and checked for the definiteness of the matrix of second-order derivatives. Here (Gaussian case), it is easily seen that the matrix of second-order derivatives is negative definite:

$$
\begin{aligned}
&\frac{\partial}{\partial\mu}\ln L = \frac{n}{\sigma^2}(\bar{x}-\mu)\\
&\frac{\partial^2}{\partial\mu^2}\ln L = -\frac{n}{\sigma^2} = -\frac{n}{s^2} \quad \text{at the critical point}\\
&\frac{\partial}{\partial\theta}\frac{\partial}{\partial\mu}\ln L = -\frac{n}{\theta^2}(\bar{x}-\mu)\Big|_{\hat{\theta},\hat{\mu}} = 0, \quad \theta = \sigma^2\\
&\frac{\partial^2}{\partial\theta^2}\ln L = \left[\frac{n}{\theta^2} - \frac{1}{\theta^2}(ns^2 + n(\bar{x}-\mu)^2)\right]\Big|_{\hat{\theta},\hat{\mu}} = -\frac{n}{2\hat{\theta}^2}.
\end{aligned}
$$

Therefore, the matrix of second-order derivatives, evaluated at $\hat{\theta}, \hat{\mu}$ is given by

$$
\begin{bmatrix} -\frac{n}{\hat{\theta}} & 0 \\ 0 & -\frac{n}{2\hat{\theta}^2} \end{bmatrix}
$$

which is negative definite. Hence the critical point corresponds to a maximum. We can also note this by observing the behavior $\ln L$ for all possible values of $\mu$ and $\sigma^2$. You will see that $\ln L$ goes from $-\infty$ back to $-\infty$ through finite values, and hence the only critical point must correspond to a maximum.

**Question.** If the null hypothesis is $H_0 : \mu < \mu_0$, is there a MLE under the null hypothesis?

**Answer.** No. There does not exist a MLE if the hypothesis is in an open interval. Nothing can be maximized in an open interval. $H_0$ must have a boundary point, otherwise the MLE does not exist. This is a very important point. For a general $\theta$, if the claim is $\theta < \theta_0$ then formulate the null hypothesis as $\theta \ge \theta_0$ and test it against $\theta < \theta_0$. Such a test is possible under the likelihood ratio test because the MLE under $\Omega$ (parameter space) and under $H_0$ are both available.

**Question.** If we had taken $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ would it not give the same $\lambda$-criterion?

**Answer.** Yes. But the procedure is logically incorrect because the alternate must be the natural alternate. If $H_0$ is $\mu = \mu_0$, then the natural alternate is $H_1 : \mu \neq \mu_0$. If the possible parameter values are known to be $\mu_0 \ge \mu < \infty$, then the alternate for $H_0 : \mu = \mu_0$ is $H_1 : \mu > \mu_0$. Otherwise, the procedure is logically incorrect.

**Hint.** Note that the rejection region is in the direction of the alternate. Here, the alternate is $H_1 : \mu > \mu_0$ and we reject for large values of $z$ or for $z \ge z_\alpha$. This, in fact, is a general observation.

**Question.** Is it true that if $x_1, \dots, x_k$ are jointly normally distributed then any linear function of $x_1, \dots, x_k$ is univariate normal?

**Answer.** If the multivariate normal is taken as the usual multivariate function denoted by $N_p(\mu, \Sigma)$, $\Sigma > O$ or $\Sigma$ is at least positive semi-definite, then it can be proved that any linear function $u$ of $x_1, \dots, x_k$, containing at least one variable, is univariate normal with the parameters $E(u)$ and $\mathrm{Var}(u)$.

**Question.** Does non-rejection of $H_0$ mean that the hypothesis $H_0$ is accepted?

**Answer.** Non-rejection does not mean acceptance. Non-rejection may be due to many reasons. Our assumption of joint normality may be faulty and in another data set the decision may be to reject. The procedure of constructing the test is geared to rejecting $H_0$. We fixed the probability of rejection when $H_0$ is true as $\alpha$ (we give that much probability for our decision to be wrong) and we maximized the probability of rejection when $H_0$ is not true. The maximum that we can say is that the data seems to be consistent with the hypothesis.

**Question.** If we have the data in hand, then by looking at the data we can create a hypothesis that can never be rejected or that can always be rejected. Then what is the meaning of testing of hypotheses?

**Answer.** If we have the data in hand, then by looking at the data we can create a hypothesis of our interest to reject or not to reject. This is not the idea of testing a statistical hypothesis. We create a hypothesis first. Then we collect data in the form of a random sample on the relevant variables. Then carry out the test by using one of the testing procedures. This is the idea.

**Question.** Consider, for example, the hypothesis $H_0 : \mu \le \mu_0$ in a $N(\mu, \sigma^2)$ with $\sigma^2$ known. Then we reject for large values of $z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}$. Since an $n$ is present in the numerator, can we not take $n$ large enough so that we will not reject $H_0$ at all. Is it not true?

**Answer.** This is a question the critics of testing of statistical hypotheses raise. It is not quite correct. By the weak law of large numbers, $\bar{x}$ is going to the true value of $\mu$. If the null hypothesis is correct, then $\bar{x}$ is going to $\mu_0$ when $n \to \infty$. Hence when $n \to \infty$, $z$ is not going to $\infty$ freely. When $n$ is changing the density becomes more and more peaked and still the probability for large $z$ is kept as $\alpha$. Hence, even though it appears that $z$ is going to infinity when $n \to \infty$ it is not so.

**Question.** In the likelihood function relating to Bernoulli population, why are we not taking the binomial coefficient $\binom{n}{x}$?

**Answer.** We are taking a simple random sample from a Bernoulli population. Then the likelihood function is $\prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j}$ and there is no binomial coefficient $\binom{n}{x}$. If we had taken one observation from a binomial distribution then there would have been this binomial coefficient because we have derived the probability function of $x$, which has the binomial coefficient coming from the number of combinations of $n$ taking $x$ at a time, where $x$ = the binomial random variable is really the Bernoulli sum.

**Question.** In all the problems before, we had taken the probability statement as exactly equal to $\alpha$. Why do we take it as $\leq\alpha$, in the discrete case, why not $\geq\alpha$?

**Answer.** In all the previous problems, our population was continuous, and hence we were able to solve for the critical point with the right side exactly equal to $\alpha$. When the population is discrete, individual probability masses are at distinct points. When we start adding up from one end the sum need not hit exactly $\alpha = 0.05$, $\alpha = 0.01$, etc. Up to a certain point, the sum may be less than $\alpha$ and when you add the mass at the next point the total may exceed $\alpha$. Then we stop at the point where the sum is less than $\alpha$ if it did not hit $\alpha$. Why not take bigger than $\alpha$ or the point when it just exceeds $\alpha$? We are prefixing $\alpha$ or allowing a certain tolerance and smaller the tolerance level to go wrong is better. Hence we take less than or equal to $\alpha$.

**Question.** In using lack-of-fit or goodness-of-fit tests, if we had created a claim by looking at the data then we could have not rejected the hypothesis. Why not modify the claim?

**Answer.** In any testing procedure, the hypothesis or the claim has to come first, then we take the data and check the claim against the data and not the other way around. If the claim is made by looking at the data, then we can always make the claim either consistent with the data (not to reject $H_0$) or reject $H_0$. Then the purpose of testing hypothesis will be defeated.

**Question.** Does it mean that no other Poisson model is a better fit to the data [testing goodness-of-fit of a Poisson model]?

**Answer.** No. We only fitted one Poisson model with $\lambda = 2.4$. We did not exhaust all possible parameter values. We got 2.4 as the MLE, which as an estimator has many interesting properties. That is all.

**Question.** Even though MLE has many interesting properties, can we find a better fitting Poisson model to this data?

**Answer.** Usually, we will be able to find a better fitting model from the same family. In the present case, the answer is "yes". Take $\lambda = 2.5$. Then the $X^2$ value can be seen to be $2.36 < 4.04 < 9.49$ which shows that the Poisson models with $\lambda = 2.4$ and $\lambda = 2.5$ are good fits to the data. Further, since $X^2$ is a measure of generalized distance between the observed and expected frequencies, smaller $X^2$ value is better the model. Hence the Poisson model with $\lambda = 2.5$ is a better Poisson model to the data than the one with $\lambda = 2.4$ given by the MLE. A Poisson model with $\lambda = 2.3$ can also be seen to be a good fit ($H_0$ is not rejected), not better than the cases for $\lambda = 2.4$ and 2.5.

**Question.** From the above conclusions, how good is this "goodness-of-fit" test?

**Answer.** In "goodness-of-fit tests", the testing procedure defeats the purpose. As per the testing procedure in hypotheses testing, if $H_0$ is rejected then it is a valid conclusion and if $H_0$ is not rejected the procedure does not help to say anything further or the procedure does not allow you to "accept" $H_0$. The whole purpose of going for "goodness-of-fit" test is to claim that the model is a good fit to the data. Hence the statistical aspect is controversial and better to forget about the statistical part. Instead of relying on the chi-square critical point, one can treat Pearson's $X^2$ as the square of a distance between the observed and expected frequencies. Hence use the criterion: "smaller the distance better the model". By this process, if we want to fit a Poisson model to the above data and if the three models with $\lambda = 2.3$, $\lambda = 2.4$, $\lambda = 2.5$ are compared then we will select the model with $\lambda = 2.5$ because the distance there is the smallest among the three. One should call this class of tests "lack-of-fit" tests, that is what is exactly measured by the statistical procedure.

**Question.** Can we come up with any other discrete distribution, other than a Poisson model to fit this data?

**Answer.** Yes. We can come up with many other discrete models which will fit this data. For example, consider a multinomial model where the hypothesized probabilities are

$$p_1 = 0.1, \quad p_2 = 0.2, \quad p_3 = 0.25, \quad p_4 = 0.2, \quad p_5 = 0.15, \quad p_6 = 0.1$$

[the observed proportions]. The distance between the observed and expected frequencies is exactly zero. Hence there cannot be a better model than this to fit this data. Consider new multinomial models with slight changes in the above probabilities so that the $X^2$ value in each case will not reject $H_0$. There can be infinitely many such models which will all fit the data, even with $X^2$ value smaller than 2.36 the best among the three Poisson models that we have considered. Thus, many better fitting models can be constructed belonging to other families of distributions or may be to the same family.

When Pearson's $X^2$ test or any other so-called "goodness-of-fit" test is used, the decision of non-rejection cannot be given too much importance. If one model is found to be a good fit, under some criterion, we may be able to find several other models which are better fits or as good as the selected model, under the same criterion.

**Question.** Do we have to compute the full $X^2$ value [Pearson's $X^2$] to make a decision?

**Answer.** As illustrated in the example in this book, it is not necessary to compute the full $X^2$ value. We need to check only whether the observed $X^2$ exceeds the critical point or not. This may be possible by computing from a few cells.

**Question.** Can we use this procedure whatever be the number of cells in our classification?

**Answer.** No. In order to have a good chi-square approximation for Pearson's $X^2$ statistic, the following is a rule of thumb. If there is no estimation involved, then the number of cells $k \geq 5$ and the expected frequencies in each cell, under $H_0$, must be ≥5. If estimation of parameters is involved in obtaining the estimated expected frequencies and if the resulting degrees of freedom is $\nu$, then $\nu - 1 \geq 5$ and each of the estimated expected frequency is ≥5. This is a rule of thumb. If the number of cells is only 2, then binomial situation arises. Here, the total frequency $n \geq 20$ for a reasonable approximation if the true probability is not close to zero or 1. Similarly, for $k = 3, 4$ one can find separate conditions for a good chi-square approximation.

**Question.** Rejection of $H_0$ of no association means what [two-way contingency table]?

**Answer.** In a testing procedure, rejection is a logical conclusion. If $H_0$ is not rejected, then no valid conclusion can be made. Our decision is based on one data set. In another data set, perhaps the decision may be different unless the original data set is a representative of the whole population in every respect. This type of representation is not possible in practice. Hence the maximum that we can say is that the data seem to be consistent with the hypothesis when $H_0$ is not rejected. Here, our situation is different. The hypothesis of no association is rejected. Does it not mean that there is possibility of association between the characteristics of classification? Since rejection is a valid conclusion, we must admit that this data set suggests that there is possibility of some sort of association between the characteristics of classification or there is possibility of association between weights and intelligence according to this data set and according to this testing procedure.

**Question.** If the observations are $2, 5, 6$, how many populations are possible from where these observations came?

**Answer.** There can be infinitely many populations or all populations where the range, with non-zero probabilities, cover these observations.

**Question.** Suppose that we computed $D_n$ [Kolmogorov–Smirnov statistic for goodness-of-fit] for one sample with $n = 10$ and got the number 2.3. How many different populations are possible where an observed sample of size 10 gave the $D_n$ measure as 2.3? How logical is the statistical procedure in this situation?

**Answer.** Infinitely many populations are possible. The logical basis is very shaky. If $H_0$ is rejected, it is well and good. If $H_0$ is not rejected and if we wish to say anything about the selected model, then the statistical procedure cannot logically support the move. What one can say is to compute $D_n$ or any such distance measure and use the criterion "smaller the distance better the fit" and then say that the selected model seems to be a good fit if the distance is smaller than a preassigned number, remembering that there could be several other models which may also be good fits to the same data.

**Question.** Are not the procedure of constructing confidence intervals the same as testing of hypotheses; one seems to be a complement of the other?

**Answer.** Some people mix up the two procedures and give credence to the statement of "accepting $H_0$" saying that confidence intervals and "acceptance regions" coincide in many cases. The two procedures are different and the premises are different also. In some populations, such as the normal population the test statistics and pivotal quantities are similar and this aspect may give rise to this doubt. Take the binomial and Poisson parameters. Constructing confidence intervals is fully different from testing a hypothesis there. Pivotal quantities may not be there for constructing confidence intervals but under a null hypothesis the populations may be fully known. Testing is based on some motivating principle such as the likelihood ratio test, maximizing power, etc. whereas for constructing confidence intervals one may select a pivotal quantity arbitrarily, the same pivotal quantity as well as different pivotal quantities giving different confidence intervals for the same parameter with the same confidence coefficient. The procedure of constructing confidence intervals and testing of hypotheses should not be mixed up; these two have different premises.

# Tables of percentage points

**Table 1:** Binomial coefficients

**Entry:** $\binom{n}{x} = \binom{n}{n-x} = \frac{n!}{x!(n-x)!}$

| *n* | *x* = 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 5 | 10 | | | | | | |
| 6 | 1 | 6 | 15 | 20 | | | | | |
| 7 | 1 | 7 | 21 | 35 | | | | | |
| 8 | 1 | 8 | 28 | 56 | 70 | | | | |
| 9 | 1 | 9 | 36 | 84 | 126 | | | | |
| 10 | 1 | 10 | 45 | 120 | 210 | 252 | | | |
| 11 | 1 | 11 | 55 | 165 | 330 | 462 | | | |
| 12 | 1 | 12 | 66 | 220 | 495 | 792 | 924 | | |
| 13 | 1 | 13 | 78 | 286 | 715 | 1287 | 1716 | | |
| 14 | 1 | 14 | 91 | 364 | 1001 | 2002 | 3003 | 3432 | |
| 15 | 1 | 15 | 105 | 455 | 1365 | 3003 | 5005 | 6435 | |
| 16 | 1 | 16 | 120 | 560 | 1820 | 4368 | 8008 | 11440 | 12870 |
| 17 | 1 | 17 | 136 | 680 | 2380 | 6188 | 12376 | 19448 | 24310 |
| 18 | 1 | 18 | 153 | 816 | 3060 | 8568 | 18564 | 31824 | 43758 |
| 19 | 1 | 19 | 171 | 969 | 3876 | 11628 | 27132 | 50388 | 75582 |
| 20 | 1 | 20 | 190 | 1140 | 4845 | 15504 | 38760 | 77520 | 125970 |
| 21 | 1 | 21 | 210 | 1330 | 5985 | 20349 | 54264 | 116280 | 203499 |
| 22 | 1 | 22 | 231 | 1540 | 7315 | 26334 | 74613 | 170544 | 319770 |
| 23 | 1 | 23 | 253 | 1771 | 8855 | 33649 | 100947 | 245157 | 490314 |
| 24 | 1 | 24 | 276 | 2024 | 10626 | 42504 | 134596 | 346104 | 735471 |
| 25 | 1 | 25 | 300 | 2300 | 12650 | 53130 | 177100 | 480700 | 1081575 |
| 26 | 1 | 26 | 325 | 2600 | 14950 | 65780 | 230230 | 657800 | 1562275 |
| 27 | 1 | 27 | 351 | 2925 | 17550 | 80730 | 296010 | 888030 | 2220075 |
| 28 | 1 | 28 | 378 | 3276 | 20475 | 98280 | 376740 | 1184040 | 3108105 |
| 29 | 1 | 29 | 406 | 3654 | 23751 | 118755 | 475020 | 1560780 | 4292145 |
| 30 | 1 | 30 | 435 | 4060 | 27405 | 142506 | 593775 | 2035800 | 5852925 |

| *n* | *x* = 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|
| 18 | 48620 | | | | | | |
| 19 | 92378 | | | | | | |
| 20 | 167960 | 184756 | | | | | |
| 21 | 293930 | 352716 | | | | | |
| 22 | 497420 | 646646 | 705432 | | | | |
| 23 | 817190 | 1144066 | 1352078 | | | | |
| 24 | 1307504 | 1961256 | 2496144 | 2704156 | | | |
| 25 | 2042975 | 3268760 | 4457400 | 5200300 | | | |
| 26 | 3124550 | 5311735 | 7726160 | 9657700 | 10400600 | | |
| 27 | 4686825 | 8436285 | 13037895 | 17383860 | 20058300 | | |
| 28 | 6906900 | 3123110 | 21474180 | 30421755 | 37442160 | 40116600 | |
| 29 | 10015005 | 20030010 | 34597290 | 51895935 | 67863915 | 77558760 | |
| 30 | 14307150 | 30045015 | 54627300 | 86493225 | 119759850 | 145422675 | 155117520 |

https://doi.org/10.1515/9783110562545-017

**Table 2:** Cumulative binomial probabilities

**Entry:** $\sum_{r=0}^{x}\binom{n}{r}p^{r}(1-p)^{n-r} = \sum_{r=n-x}^{n}\binom{n}{r}(1-p)^{r}p^{n-r}$

| *n* | *x* | *p* = 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.9500 | 0.9000 | 0.8500 | 0.8000 | 0.7500 | 0.7000 | 0.6500 | 0.6000 | 0.5500 | 0.5000 |
|  | 1 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 2 | 0 | 0.9025 | 0.8100 | 0.7225 | 0.6400 | 0.5625 | 0.4900 | 0.4225 | 0.3600 | 0.3025 | 0.2500 |
|  | 1 | 0.9975 | 0.9900 | 0.9775 | 0.9600 | 0.9375 | 0.9100 | 0.8775 | 0.8400 | 0.7975 | 0.7500 |
|  | 2 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 3 | 0 | 0.8574 | 0.7290 | 0.6141 | 0.5120 | 0.4219 | 0.3430 | 0.2746 | 0.2160 | 0.1664 | 0.1250 |
|  | 1 | 0.9927 | 0.9720 | 0.9392 | 0.8960 | 0.8437 | 0.7840 | 0.7182 | 0.6480 | 0.5747 | 0.5000 |
|  | 2 | 0.9999 | 0.9990 | 0.9966 | 0.9920 | 0.9844 | 0.9730 | 0.9571 | 0.9390 | 0.9089 | 0.8750 |
|  | 3 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 4 | 0 | 0.8145 | 0.6561 | 0.5220 | 0.4096 | 0.3164 | 0.2401 | 0.1785 | 0.1296 | 0.0915 | 0.0625 |
|  | 1 | 0.9860 | 0.9477 | 0.8905 | 0.8912 | 0.7383 | 0.6517 | 0.5630 | 0.4752 | 0.3910 | 0.3125 |
|  | 2 | 0.9995 | 0.9963 | 0.9880 | 0.9728 | 0.9492 | 0.9163 | 0.8735 | 0.8208 | 0.7585 | 0.6875 |
|  | 3 | 1.0000 | 0.9999 | 0.9995 | 0.9984 | 0.9961 | 0.9919 | 0.9850 | 0.9744 | 0.9590 | 0.9375 |
|  | 4 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 5 | 0 | 0.7738 | 0.5905 | 0.4437 | 0.3277 | 0.2373 | 0.1681 | 0.1160 | 0.0778 | 0.0503 | 0.0313 |
|  | 1 | 0.9774 | 0.9185 | 0.8352 | 0.7373 | 0.6328 | 0.5282 | 0.4284 | 0.3370 | 0.2562 | 0.1875 |
|  | 2 | 0.9988 | 0.9914 | 0.9734 | 0.9421 | 0.8905 | 0.8369 | 0.7648 | 0.6826 | 0.5931 | 0.5000 |
|  | 3 | 1.0000 | 0.9995 | 0.9978 | 0.9933 | 0.9844 | 0.9692 | 0.9460 | 0.9130 | 0.8688 | 0.8125 |
|  | 4 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9990 | 0.9976 | 0.9947 | 0.9898 | 0.9815 | 0.9688 |
|  | 5 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 6 | 0 | 0.7351 | 0.5314 | 0.3771 | 0.2621 | 0.1780 | 0.1176 | 0.0754 | 0.0467 | 0.0277 | 0.0156 |
|  | 1 | 0.9672 | 0.8857 | 0.7765 | 0.6554 | 0.5339 | 0.4202 | 0.3191 | 0.2333 | 0.1636 | 0.1094 |
|  | 2 | 0.9978 | 0.9841 | 0.9527 | 0.9011 | 0.8306 | 0.7443 | 0.6471 | 0.5443 | 0.4415 | 0.3438 |
|  | 3 | 0.9999 | 0.9987 | 0.9941 | 0.9830 | 0.9624 | 0.9295 | 0.8826 | 0.8208 | 0.7447 | 0.6562 |
|  | 4 | 1.0000 | 0.9999 | 0.9996 | 0.9984 | 0.9954 | 0.9891 | 0.9777 | 0.9590 | 0.9308 | 0.8906 |
|  | 5 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9993 | 0.9982 | 0.9959 | 0.9917 | 0.9844 |
|  | 6 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 7 | 0 | 0.6983 | 0.4783 | 0.3206 | 0.2097 | 0.1335 | 0.0824 | 0.0490 | 0.0280 | 0.0152 | 0.0078 |
|  | 1 | 0.9556 | 0.8503 | 0.7166 | 0.5767 | 0.4449 | 0.3294 | 0.2338 | 0.1586 | 0.1024 | 0.0625 |
|  | 2 | 0.9962 | 0.9743 | 0.9262 | 0.8520 | 0.7564 | 0.6471 | 0.5323 | 0.4199 | 0.3164 | 0.2266 |
|  | 3 | 0.9998 | 0.9973 | 0.9879 | 0.9667 | 0.9294 | 0.8740 | 0.8002 | 0.7102 | 0.6083 | 0.5000 |
|  | 4 | 1.0000 | 0.9998 | 0.9988 | 0.9953 | 0.9871 | 0.9712 | 0.9444 | 0.9037 | 0.8471 | 0.7734 |
|  | 5 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9987 | 0.9962 | 0.9910 | 0.9812 | 0.9643 | 0.9375 |
|  | 6 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9994 | 0.9984 | 0.9963 | 0.9922 |
|  | 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 8 | 0 | 0.6634 | 0.4305 | 0.2725 | 0.1678 | 0.1001 | 0.0576 | 0.0319 | 0.0168 | 0.0084 | 0.0039 |
|  | 1 | 0.9428 | 0.8131 | 0.6572 | 0.5033 | 0.3671 | 0.2553 | 0.1691 | 0.1064 | 0.0632 | 0.0352 |
|  | 2 | 0.9942 | 0.9619 | 0.8948 | 0.7969 | 0.6786 | 0.5518 | 0.4278 | 0.3154 | 0.2201 | 0.1445 |
|  | 3 | 0.9996 | 0.9950 | 0.9786 | 0.9437 | 0.8862 | 0.8059 | 0.7064 | 0.5941 | 0.4770 | 0.3633 |

**Table 2:** (continued)

| *n* | *x* | *p* = 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 1.0000 | 0.9996 | 0.9971 | 0.9896 | 0.9727 | 0.9420 | 0.8939 | 0.8263 | 0.7396 | 0.6367 |
| | 5 | 1.0000 | 1.0000 | 0.9998 | 0.9988 | 0.9958 | 0.9887 | 0.9747 | 0.9502 | 0.9115 | 0.8555 |
| | 6 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9987 | 0.9964 | 0.9915 | 0.9819 | 0.9648 |
| | 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9993 | 0.9983 | 0.9961 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 9 | 0 | 0.6302 | 0.3874 | 0.2316 | 0.1342 | 0.0751 | 0.0404 | 0.0207 | 0.0101 | 0.0046 | 0.0020 |
| | 1 | 0.9288 | 0.7748 | 0.5995 | 0.4362 | 0.3003 | 0.1960 | 0.1211 | 0.0705 | 0.0385 | 0.0195 |
| | 2 | 0.9916 | 0.9470 | 0.8591 | 0.7382 | 0.6007 | 0.4628 | 0.3373 | 0.2318 | 0.1495 | 0.0898 |
| | 3 | 0.9994 | 0.9917 | 0.9661 | 0.9144 | 0.8343 | 0.7297 | 0.6089 | 0.4826 | 0.3614 | 0.2539 |
| | 4 | 1.0000 | 0.9991 | 0.9944 | 0.9804 | 0.9511 | 0.9012 | 0.8283 | 0.7334 | 0.6214 | 0.5000 |
| | 5 | 1.0000 | 0.9999 | 0.9994 | 0.9969 | 0.9900 | 0.9747 | 0.9464 | 0.9006 | 0.8342 | 0.7461 |
| | 6 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9987 | 0.9957 | 0.9888 | 0.9750 | 0.9502 | 0.9102 |
| | 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9986 | 0.9962 | 0.9909 | 0.9805 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9992 | 0.9980 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 10 | 0 | 0.5987 | 0.3487 | 0.1969 | 0.1074 | 0.0563 | 0.0282 | 0.0135 | 0.0060 | 0.0025 | 0.0010 |
| | 1 | 0.9130 | 0.7361 | 0.5443 | 0.3758 | 0.2440 | 0.1493 | 0.0860 | 0.0464 | 0.0233 | 0.0107 |
| | 2 | 0.9885 | 0.9298 | 0.8202 | 0.6778 | 0.5256 | 0.3828 | 0.2616 | 0.1673 | 0.0996 | 0.0547 |
| | 3 | 0.9990 | 0.9872 | 0.9500 | 0.8791 | 0.7759 | 0.6496 | 0.5138 | 0.3823 | 0.2660 | 0.1710 |
| | 4 | 0.9999 | 0.9984 | 0.9901 | 0.9672 | 0.9219 | 0.8497 | 0.7515 | 0.6331 | 0.5044 | 0.3770 |
| | 5 | 1.0000 | 0.9999 | 0.9986 | 0.9936 | 0.9803 | 0.9527 | 0.9051 | 0.8338 | 0.7384 | 0.6230 |
| | 6 | 1.0000 | 1.0000 | 0.9999 | 0.9991 | 0.9965 | 0.9894 | 0.9740 | 0.9452 | 0.8980 | 0.8281 |
| | 7 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9984 | 0.9952 | 0.9877 | 0.9726 | 0.9453 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9995 | 0.9983 | 0.9955 | 0.9803 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9990 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 11 | 0 | 0.5688 | 0.3138 | 0.1673 | 0.0859 | 0.0859 | 0.0422 | 0.0198 | 0.0088 | 0.0036 | 0.0005 |
| | 1 | 0.8981 | 0.6974 | 0.4922 | 0.3221 | 0.1971 | 0.1130 | 0.0606 | 0.0302 | 0.0139 | 0.0059 |
| | 2 | 0.9848 | 0.9104 | 0.7788 | 0.6174 | 0.4552 | 0.3127 | 0.2001 | 0.1189 | 0.0652 | 0.0327 |
| | 3 | 0.9984 | 0.9815 | 0.9306 | 0.8389 | 0.7133 | 0.5696 | 0.4256 | 0.2963 | 0.1911 | 0.1133 |
| | 4 | 0.9999 | 0.9972 | 0.9841 | 0.9496 | 0.8854 | 0.7897 | 0.6683 | 0.5328 | 0.3971 | 0.2744 |
| | 5 | 1.0000 | 0.9997 | 0.9973 | 0.9883 | 0.9657 | 0.9218 | 0.8513 | 0.7535 | 0.6331 | 0.5000 |
| | 6 | 1.0000 | 1.0000 | 0.9997 | 0.9980 | 0.9924 | 0.9784 | 0.9499 | 0.9006 | 0.8262 | 0.7256 |
| | 7 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9988 | 0.9957 | 0.9878 | 0.9707 | 0.9390 | 0.8867 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9980 | 0.9941 | 0.9852 | 0.9673 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9993 | 0.9978 | 0.9941 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9995 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 12 | 0 | 0.5404 | 0.2824 | 0.1422 | 0.0687 | 0.0317 | 0.0138 | 0.0057 | 0.0022 | 0.0008 | 0.0002 |
| | 1 | 0.8816 | 0.6590 | 0.4435 | 0.2749 | 0.1584 | 0.0850 | 0.0424 | 0.0196 | 0.0083 | 0.0032 |
| | 2 | 0.9804 | 0.8891 | 0.7358 | 0.5583 | 0.3907 | 0.2528 | 0.1513 | 0.0834 | 0.0421 | 0.0193 |
| | 3 | 0.9978 | 0.9744 | 0.9078 | 0.7946 | 0.6488 | 0.4925 | 0.3467 | 0.2253 | 0.1345 | 0.0730 |
| | 4 | 0.9998 | 0.9957 | 0.9761 | 0.9274 | 0.8424 | 0.7237 | 0.5833 | 0.4382 | 0.3044 | 0.1938 |

**Table 2:** (continued)

| *n* | *x* | *p* = 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 1.0000 | 0.9995 | 0.9954 | 0.9806 | 0.9456 | 0.8822 | 0.7873 | 0.6652 | 0.5269 | 0.3872 |
| | 6 | 1.0000 | 0.9999 | 0.9993 | 0.9961 | 0.9857 | 0.9614 | 0.9154 | 0.8418 | 0.7393 | 0.6128 |
| | 7 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9972 | 0.9905 | 0.9745 | 0.9427 | 0.8883 | 0.8062 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9983 | 0.9944 | 0.9847 | 0.9644 | 0.9270 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9992 | 0.9972 | 0.9921 | 0.9807 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9989 | 0.9968 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 13 | 0 | 0.5133 | 0.2543 | 0.1209 | 0.0550 | 0.0238 | 0.0097 | 0.0037 | 0.0013 | 0.0004 | 0.0001 |
| | 1 | 0.8646 | 0.6213 | 0.3983 | 0.2336 | 0.1267 | 0.0637 | 0.0296 | 0.0126 | 0.0049 | 0.0017 |
| | 2 | 0.9755 | 0.8661 | 0.6920 | 0.5017 | 0.3326 | 0.2025 | 0.1132 | 0.0572 | 0.0269 | 0.0112 |
| | 3 | 0.9969 | 0.9658 | 0.8820 | 0.7473 | 0.5843 | 0.4206 | 0.2783 | 0.1686 | 0.0929 | 0.0461 |
| | 4 | 0.9997 | 0.9935 | 0.9658 | 0.9009 | 0.7940 | 0.6543 | 0.5005 | 0.3530 | 0.2279 | 0.1334 |
| | 5 | 1.0000 | 0.9991 | 0.9925 | 0.9700 | 0.9198 | 0.8346 | 0.7159 | 0.5744 | 0.4268 | 0.2905 |
| | 6 | 1.0000 | 0.9999 | 0.9987 | 0.9930 | 0.9757 | 0.9376 | 0.8705 | 0.7712 | 0.6437 | 0.5000 |
| | 7 | 1.0000 | 1.0000 | 0.9998 | 0.9988 | 0.9944 | 0.9818 | 0.9538 | 0.9023 | 0.8212 | 0.7095 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9990 | 0.9960 | 0.9874 | 0.9679 | 0.9302 | 0.8666 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9993 | 0.9975 | 0.9922 | 0.9797 | 0.9539 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9987 | 0.9959 | 0.9888 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9995 | 0.9983 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 14 | 0 | 0.4877 | 0.2288 | 0.1028 | 0.0440 | 0.0178 | 0.0068 | 0.0024 | 0.0008 | 0.0002 | 0.0001 |
| | 1 | 0.8470 | 0.5846 | 0.3567 | 0.1979 | 0.1010 | 0.0475 | 0.0205 | 0.0081 | 0.0029 | 0.0009 |
| | 2 | 0.9699 | 0.8416 | 0.6479 | 0.4481 | 0.2811 | 0.1608 | 0.0839 | 0.0398 | 0.0170 | 0.0065 |
| | 3 | 0.9958 | 0.9559 | 0.8535 | 0.6982 | 0.5213 | 0.3552 | 0.2205 | 0.1243 | 0.0632 | 0.0287 |
| | 4 | 0.9996 | 0.9908 | 0.9533 | 0.8702 | 0.7415 | 0.5842 | 0.4227 | 0.2793 | 0.1672 | 0.0898 |
| | 5 | 1.0000 | 0.9985 | 0.9885 | 0.9561 | 0.8883 | 0.7805 | 0.6405 | 0.4859 | 0.3373 | 0.2120 |
| | 6 | 1.0000 | 0.9998 | 0.9978 | 0.9884 | 0.9617 | 0.9064 | 0.8164 | 0.6925 | 0.5461 | 0.3953 |
| | 7 | 1.0000 | 1.0000 | 0.9997 | 0.9976 | 0.9897 | 0.9685 | 0.9247 | 0.8499 | 0.7414 | 0.6047 |
| | 8 | 1.0000 | 1.0000 | 1.0000 | 0.9996 | 0.9978 | 0.9917 | 0.9757 | 0.9417 | 0.8811 | 0.7880 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9983 | 0.9940 | 0.9825 | 0.9574 | 0.9102 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9989 | 0.9961 | 0.9886 | 0.9713 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9978 | 0.9935 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9991 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 15 | 0 | 0.4633 | 0.2059 | 0.0874 | 0.0352 | 0.0134 | 0.0047 | 0.0016 | 0.0005 | 0.0001 | 0.0000 |
| | 1 | 0.8290 | 0.5490 | 0.3186 | 0.1671 | 0.0802 | 0.0353 | 0.0142 | 0.0052 | 0.0017 | 0.0005 |
| | 2 | 0.9638 | 0.8159 | 0.6042 | 0.3980 | 0.2361 | 0.1268 | 0.0617 | 0.0271 | 0.0107 | 0.0037 |
| | 3 | 0.9945 | 0.9444 | 0.8227 | 0.6482 | 0.4613 | 0.2969 | 0.1727 | 0.0905 | 0.0424 | 0.0176 |
| | 4 | 0.9994 | 0.9873 | 0.8358 | 0.6865 | 0.5155 | 0.3519 | 0.3519 | 0.2173 | 0.1204 | 0.0592 |
| | 5 | 0.9999 | 0.9978 | 0.9832 | 0.9389 | 0.8516 | 0.7216 | 0.5643 | 0.4032 | 0.2608 | 0.1509 |

**Table 2:** (continued)

| *n* | *x* | *p* = 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 1.0000 | 0.9997 | 0.9964 | 0.9819 | 0.9434 | 0.8689 | 0.7548 | 0.6098 | 0.4522 | 0.3036 |
| | 7 | 1.0000 | 1.0000 | 0.9994 | 0.9958 | 0.9827 | 0.9500 | 0.8868 | 0.7869 | 0.6535 | 0.5000 |
| | 8 | 1.0000 | 1.0000 | 0.9999 | 0.9992 | 0.9958 | 0.9948 | 0.9578 | 0.9050 | 0.8182 | 0.6964 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9992 | 0.9963 | 0.9876 | 0.9662 | 0.9231 | 0.8491 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9993 | 0.9972 | 0.9907 | 0.9745 | 0.9408 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9995 | 0.9981 | 0.9937 | 0.9824 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9989 | 0.9963 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9995 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 16 | 0 | 0.4401 | 0.1853 | 0.0743 | 0.0281 | 0.0100 | 0.0033 | 0.0010 | 0.0003 | 0.0001 | 0.0000 |
| | 1 | 0.8108 | 0.5147 | 0.2839 | 0.1407 | 0.0635 | 0.0261 | 0.0098 | 0.0033 | 0.0010 | 0.0003 |
| | 2 | 0.9571 | 0.7892 | 0.5614 | 0.3518 | 0.1971 | 0.0994 | 0.0451 | 0.0183 | 0.0066 | 0.0021 |
| | 3 | 0.9930 | 0.9316 | 0.7899 | 0.5981 | 0.4050 | 0.2459 | 0.1339 | 0.0651 | 0.0281 | 0.0106 |
| | 4 | 0.9991 | 0.9830 | 0.9209 | 0.7982 | 0.6302 | 0.4499 | 0.2892 | 0.1666 | 0.0853 | 0.0384 |
| | 5 | 0.9999 | 0.9967 | 0.9765 | 0.9183 | 0.8103 | 0.6598 | 0.4900 | 0.3288 | 0.1976 | 0.1051 |
| | 6 | 1.0000 | 0.9995 | 0.9944 | 0.9733 | 0.9204 | 0.8247 | 0.6881 | 0.5272 | 0.3660 | 0.2272 |
| | 7 | 1.0000 | 0.9999 | 0.9989 | 0.9930 | 0.9729 | 0.9256 | 0.8406 | 0.7161 | 0.5629 | 0.4018 |
| | 8 | 1.0000 | 1.0000 | 0.9998 | 0.9985 | 0.9925 | 0.9743 | 0.9329 | 0.8577 | 0.7441 | 0.5982 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9984 | 0.9929 | 0.9771 | 0.9417 | 0.8750 | 0.7728 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9984 | 0.9938 | 0.9809 | 0.9514 | 0.8949 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9987 | 0.9851 | 0.9851 | 0.9616 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9991 | 0.9965 | 0.9894 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9979 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 |
| | 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 17 | 0 | 0.4181 | 0.1668 | 0.0631 | 0.0225 | 0.0075 | 0.0023 | 0.0007 | 0.0002 | 0.0000 | 0.0000 |
| | 1 | 0.7922 | 0.4818 | 0.2525 | 0.1182 | 0.0501 | 0.0193 | 0.0067 | 0.0021 | 0.0006 | 0.0001 |
| | 2 | 0.9497 | 0.7618 | 0.5198 | 0.3096 | 0.1637 | 0.0774 | 0.0327 | 0.0123 | 0.0041 | 0.0012 |
| | 3 | 0.9912 | 0.9174 | 0.7556 | 0.5489 | 0.3530 | 0.2019 | 0.1028 | 0.0464 | 0.0184 | 0.0064 |
| | 4 | 0.9988 | 0.9779 | 0.9013 | 0.7582 | 0.5739 | 0.3887 | 0.2348 | 0.1260 | 0.0596 | 0.0245 |
| | 5 | 0.9999 | 0.9953 | 0.9681 | 0.8943 | 0.7653 | 0.5968 | 0.4197 | 0.2639 | 0.1471 | 0.0717 |
| | 6 | 1.0000 | 0.9992 | 0.9917 | 0.9623 | 0.8929 | 0.7752 | 0.6188 | 0.4478 | 0.2902 | 0.1662 |
| | 7 | 1.0000 | 0.9999 | 0.9983 | 0.9891 | 0.9598 | 0.8954 | 0.7872 | 0.4405 | 0.4743 | 0.3145 |
| | 8 | 1.0000 | 1.0000 | 0.9997 | 0.9974 | 0.9876 | 0.9597 | 0.9006 | 0.8011 | 0.6626 | 0.5000 |
| | 9 | 1.0000 | 1.0000 | 1.0000 | 0.9995 | 0.9969 | 0.9873 | 0.9617 | 0.9081 | 0.8166 | 0.6855 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9968 | 0.9880 | 0.9652 | 0.9174 | 0.8338 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9993 | 0.9970 | 0.9894 | 0.9699 | 0.9283 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9975 | 0.9914 | 0.9755 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9995 | 0.9981 | 0.9936 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9988 |
| | 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |
| | 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**Table 2:** (continued)

| *n* | *x* | *p* = 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | 0 | 0.3972 | 0.1501 | 0.0536 | 0.0180 | 0.0056 | 0.0016 | 0.0004 | 0.0001 | 0.0000 | 0.0000 |
| | 1 | 0.7735 | 0.4503 | 0.2241 | 0.0991 | 0.0395 | 0.0142 | 0.0046 | 0.0013 | 0.0003 | 0.0001 |
| | 2 | 0.9419 | 0.7338 | 0.4797 | 0.2713 | 0.1353 | 0.0600 | 0.0236 | 0.0082 | 0.0025 | 0.0007 |
| | 3 | 0.9891 | 0.9018 | 0.7202 | 0.5010 | 0.3057 | 0.1646 | 0.0783 | 0.0328 | 0.0120 | 0.0038 |
| | 4 | 0.9985 | 0.9718 | 0.8794 | 0.7164 | 0.5187 | 0.3327 | 0.1886 | 0.0942 | 0.0411 | 0.0154 |
| | 5 | 0.9998 | 0.9936 | 0.9581 | 0.8671 | 0.7175 | 0.5344 | 0.3550 | 0.2088 | 0.1077 | 0.0481 |
| | 6 | 1.0000 | 0.9988 | 0.9882 | 0.9487 | 0.8610 | 0.7217 | 0.5491 | 0.3743 | 0.2258 | 0.1189 |
| | 7 | 1.0000 | 0.9998 | 0.9973 | 0.9837 | 0.9431 | 0.8593 | 0.7283 | 0.5634 | 0.3915 | 0.2403 |
| | 8 | 1.0000 | 1.0000 | 0.9995 | 0.9957 | 0.9807 | 0.9404 | 0.8609 | 0.7368 | 0.5778 | 0.4073 |
| | 9 | 1.0000 | 1.0000 | 0.9999 | 0.9991 | 0.9946 | 0.9790 | 0.9403 | 0.8653 | 0.7473 | 0.5927 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9988 | 0.9939 | 0.9788 | 0.9424 | 0.8720 | 0.7597 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9986 | 0.9938 | 0.9797 | 0.9463 | 0.8811 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9956 | 0.9942 | 0.9817 | 0.9519 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9987 | 0.9951 | 0.9846 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9990 | 0.9962 |
| | 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9993 |
| | 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |
| | 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 19 | 0 | 0.3774 | 0.1351 | 0.0456 | 0.0144 | 0.0042 | 0.0011 | 0.0003 | 0.0001 | 0.0000 | 0.0000 |
| | 1 | 0.7547 | 0.4203 | 0.1985 | 0.0829 | 0.0310 | 0.0104 | 0.0031 | 0.0008 | 0.0002 | 0.0000 |
| | 2 | 0.9335 | 0.7054 | 0.4413 | 0.2369 | 0.1113 | 0.0462 | 0.0170 | 0.0055 | 0.0015 | 0.0004 |
| | 3 | 0.9868 | 0.8850 | 0.6841 | 0.4551 | 0.2631 | 0.1332 | 0.0591 | 0.0230 | 0.0077 | 0.0022 |
| | 4 | 0.9980 | 0.9643 | 0.8556 | 0.6733 | 0.4654 | 0.2822 | 0.1500 | 0.0696 | 0.0280 | 0.0096 |
| | 5 | 0.9998 | 0.9914 | 0.9463 | 0.8369 | 0.6678 | 0.4739 | 0.2968 | 0.1629 | 0.0777 | 0.0318 |
| | 6 | 1.0000 | 0.9983 | 0.9837 | 0.9324 | 0.8251 | 0.6655 | 0.4812 | 0.3081 | 0.1727 | 0.0835 |
| | 7 | 1.0000 | 0.9997 | 0.9959 | 0.9767 | 0.9225 | 0.8180 | 0.6656 | 0.4878 | 0.3169 | 0.1796 |
| | 8 | 1.0000 | 1.0000 | 0.9992 | 0.9933 | 0.9713 | 0.9161 | 0.8145 | 0.6675 | 0.4940 | 0.3238 |
| | 9 | 1.0000 | 1.0000 | 0.9999 | 0.9984 | 0.9911 | 0.9674 | 0.9125 | 0.8139 | 0.6710 | 0.5000 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9977 | 0.9895 | 0.9653 | 0.9115 | 0.8159 | 0.6762 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9995 | 0.9972 | 0.9886 | 0.9648 | 0.9129 | 0.8204 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9969 | 0.9884 | 0.9658 | 0.9165 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9993 | 0.9969 | 0.9891 | 0.9682 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9994 | 0.9972 | 0.9904 |
| | 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9995 | 0.9978 |
| | 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 |
| | 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 20 | 0 | 0.3585 | 0.1216 | 0.0388 | 0.0115 | 0.0032 | 0.0008 | 0.0002 | 0.0000 | 0.0000 | 0.0000 |
| | 1 | 0.7358 | 0.3917 | 0.1756 | 0.0692 | 0.0243 | 0.0076 | 0.0021 | 0.0005 | 0.0001 | 0.0000 |
| | 2 | 0.9245 | 0.6769 | 0.4049 | 0.2061 | 0.0913 | 0.0355 | 0.0121 | 0.0036 | 0.0009 | 0.0002 |
| | 3 | 0.9841 | 0.8670 | 0.6477 | 0.4114 | 0.2252 | 0.1071 | 0.0444 | 0.0160 | 0.0049 | 0.0013 |
| | 4 | 0.9974 | 0.9568 | 0.8298 | 0.6296 | 0.4148 | 0.2375 | 0.1182 | 0.0510 | 0.0189 | 0.0059 |
| | 5 | 0.9997 | 0.9887 | 0.9327 | 0.8042 | 0.6172 | 0.4164 | 0.2454 | 0.1256 | 0.0553 | 0.0207 |
| | 6 | 1.0000 | 0.9976 | 0.9781 | 0.9133 | 0.7858 | 0.6080 | 0.4166 | 0.2500 | 0.1299 | 0.0577 |

**Table 2:** (continued)

| $n$ | $x$ | $p = 0.05$ | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 7 | 1.0000 | 0.9996 | 0.9941 | 0.9679 | 0.8982 | 0.7723 | 0.6010 | 0.4159 | 0.2520 | 0.1316 |
| | 8 | 1.0000 | 0.9999 | 0.9987 | 0.9900 | 0.9591 | 0.8867 | 0.7624 | 0.5956 | 0.4143 | 0.2517 |
| | 9 | 1.0000 | 1.0000 | 0.9998 | 0.9974 | 0.9861 | 0.9520 | 0.8782 | 0.7553 | 0.5914 | 0.4119 |
| | 10 | 1.0000 | 1.0000 | 1.0000 | 0.9994 | 0.9961 | 0.9829 | 0.9468 | 0.8725 | 0.7507 | 0.5881 |
| | 11 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9991 | 0.9949 | 0.9804 | 0.9435 | 0.8692 | 0.7483 |
| | 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 | 0.9987 | 0.9940 | 0.9790 | 0.9420 | 0.8684 |
| | 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9985 | 0.9935 | 0.9786 | 0.9423 |
| | 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9984 | 0.9936 | 0.9793 |
| | 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9985 | 0.9941 |
| | 16 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9997 | 0.9987 |
| | 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9998 |
| | 18 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**Table 3:** Cumulative poisson probabilities

**Entry:** $\sum_{r=0}^{x} \frac{\lambda^r e^{-\lambda}}{r!}$

| $x$ | $\lambda = 0.1$ | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.9048 | 0.8187 | 0.7408 | 0.6703 | 0.6065 | 0.5488 | 0.4966 | 0.4493 | 0.4066 | 0.3679 |
| 1 | 0.9953 | 0.9825 | 0.9631 | 0.9384 | 0.9098 | 0.8781 | 0.8442 | 0.8088 | 0.7725 | 0.7358 |
| 2 | 0.9998 | 0.9989 | 0.9964 | 0.9921 | 0.9856 | 0.9769 | 0.9659 | 0.9526 | 0.9371 | 0.9197 |
| 3 | 1.0000 | 0.9999 | 0.9997 | 0.9992 | 0.9982 | 0.9966 | 0.9942 | 0.9909 | 0.9865 | 0.9810 |
| 4 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9996 | 0.9992 | 0.9986 | 0.9977 | 0.9963 |
| 5 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9997 | 0.9994 |
| 6 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |
| 7 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

| $x$ | $\lambda = 1.1$ | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.3329 | 0.3012 | 0.2725 | 0.2466 | 0.2231 | 0.2019 | 0.1827 | 0.1653 | 0.1496 | 0.1353 |
| 1 | 0.6990 | 0.6626 | 0.6268 | 0.5918 | 0.5578 | 0.5249 | 0.4932 | 0.4628 | 0.4337 | 0.4060 |
| 2 | 0.9004 | 0.8795 | 0.8571 | 0.8335 | 0.8088 | 0.7834 | 0.7572 | 0.7306 | 0.7037 | 0.6767 |
| 3 | 0.9743 | 0.9662 | 0.9569 | 0.9463 | 0.9344 | 0.9212 | 0.9068 | 0.8913 | 0.8747 | 0.8571 |
| 4 | 0.9946 | 0.9923 | 0.9893 | 0.9857 | 0.9814 | 0.9763 | 0.9704 | 0.9636 | 0.9559 | 0.9473 |
| 5 | 0.9990 | 0.9985 | 0.9978 | 0.9968 | 0.9955 | 0.9940 | 0.9920 | 0.9868 | 0.9868 | 0.9834 |
| 6 | 0.9999 | 0.9997 | 0.9996 | 0.9994 | 0.9991 | 0.9987 | 0.9981 | 0.9974 | 0.9966 | 0.9955 |
| 7 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9998 | 0.9997 | 0.9996 | 0.9994 | 0.9992 | 0.9989 |
| 8 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9998 | 0.9998 |
| 9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**Table 3:** (continued)

| *x* | $\lambda = 2.1$ | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.1225 | 0.1108 | 0.1003 | 0.0907 | 0.0821 | 0.0743 | 0.0672 | 0.0608 | 0.0550 | 0.0498 |
| 1 | 0.3796 | 0.3546 | 0.3309 | 0.3084 | 0.2873 | 0.2674 | 0.2487 | 0.2311 | 0.2146 | 0.1991 |
| 2 | 0.6496 | 0.6227 | 0.5960 | 0.5967 | 0.5438 | 0.5184 | 0.4396 | 0.4695 | 0.4460 | 0.4232 |
| 3 | 0.8386 | 0.8194 | 0.7993 | 0.7787 | 0.7576 | 0.7360 | 0.7141 | 0.6919 | 0.6696 | 0.6472 |
| 4 | 0.9379 | 0.9275 | 0.9162 | 0.9041 | 0.8912 | 0.8774 | 0.8629 | 0.8477 | 0.8318 | 0.8153 |
| 5 | 0.9796 | 0.9751 | 0.9700 | 0.9643 | 0.9580 | 0.9510 | 0.9433 | 0.9349 | 0.9258 | 0.9161 |
| 6 | 0.9941 | 0.9925 | 0.9906 | 0.9884 | 0.9858 | 0.9828 | 0.9794 | 0.9756 | 0.9713 | 0.9665 |
| 7 | 0.9985 | 0.9980 | 0.9974 | 0.9967 | 0.9958 | 0.9947 | 0.9934 | 0.9919 | 0.9901 | 0.9881 |
| 8 | 0.9997 | 0.9995 | 0.9994 | 0.9991 | 0.9989 | 0.9985 | 0.9981 | 0.9976 | 0.9969 | 0.9962 |
| 9 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9997 | 0.9996 | 0.9995 | 0.9993 | 0.9991 | 0.9989 |
| 10 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9997 |
| 11 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 |
| 12 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

| *x* | $\lambda = 3.1$ | 3.2 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0450 | 0.0408 | 0.0369 | 0.0334 | 0.0302 | 0.0273 | 0.0247 | 0.0224 | 0.0202 | 0.0183 |
| 1 | 0.1857 | 0.1712 | 0.1586 | 0.1468 | 0.1359 | 0.1257 | 0.1162 | 0.1074 | 0.0992 | 0.0916 |
| 2 | 0.4012 | 0.3799 | 0.3594 | 0.3397 | 0.3208 | 0.3027 | 0.2854 | 0.2689 | 0.2531 | 0.2381 |
| 3 | 0.6248 | 0.6025 | 0.5803 | 0.5584 | 0.5366 | 0.5152 | 0.4942 | 0.4735 | 0.4532 | 0.4335 |
| 4 | 0.7982 | 0.7806 | 0.7626 | 0.7442 | 0.7254 | 0.7064 | 0.6872 | 0.6678 | 0.6484 | 0.6288 |
| 5 | 0.9057 | 0.8946 | 0.8829 | 0.8705 | 0.8576 | 0.8441 | 0.8301 | 0.8156 | 0.8006 | 0.7851 |
| 6 | 0.9612 | 0.9554 | 0.9490 | 0.9421 | 0.9247 | 0.9267 | 0.9182 | 0.9091 | 0.8995 | 0.8893 |
| 7 | 0.9858 | 0.9832 | 0.9802 | 0.9769 | 0.9733 | 0.9692 | 0.9648 | 0.9599 | 0.9546 | 0.9489 |
| 8 | 0.9953 | 0.9943 | 0.9931 | 0.9917 | 0.9901 | 0.9883 | 0.9863 | 0.9840 | 0.9815 | 0.9786 |
| 9 | 0.9986 | 0.9982 | 0.9978 | 0.9973 | 0.9967 | 0.9960 | 0.9952 | 0.9942 | 0.9931 | 0.9919 |
| 10 | 0.9996 | 0.9995 | 0.9994 | 0.9992 | 0.9990 | 0.9987 | 0.9984 | 0.9981 | 0.9977 | 0.9972 |
| 11 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9997 | 0.9996 | 0.9995 | 0.9994 | 0.9993 | 0.9991 |
| 12 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9997 |
| 13 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

| *x* | $\lambda = 4.1$ | 4.2 | 4.3 | 4.4 | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 5.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0166 | 0.0150 | 0.0136 | 0.0123 | 0.0111 | 0.0101 | 0.0091 | 0.0082 | 0.0074 | 0.0067 |
| 1 | 0.0845 | 0.0780 | 0.0719 | 0.0663 | 0.0611 | 0.0563 | 0.0518 | 0.0477 | 0.0439 | 0.0404 |
| 2 | 0.2238 | 0.2102 | 0.1974 | 0.1851 | 0.1736 | 0.1626 | 0.1523 | 0.1425 | 0.1333 | 0.1247 |
| 3 | 0.4142 | 0.3954 | 0.3772 | 0.3594 | 0.3423 | 0.3257 | 0.3097 | 0.2942 | 0.2793 | 0.2650 |
| 4 | 0.6093 | 0.5898 | 0.5704 | 0.5512 | 0.5321 | 0.5132 | 0.4946 | 0.4763 | 0.4582 | 0.4405 |
| 5 | 0.7693 | 0.7531 | 0.7367 | 0.7190 | 0.7029 | 0.6858 | 0.6684 | 0.6510 | 0.6335 | 0.6160 |
| 6 | 0.8786 | 0.8675 | 0.8558 | 0.8436 | 0.8311 | 0.8180 | 0.8046 | 0.7908 | 0.7767 | 0.7622 |
| 7 | 0.9427 | 0.9361 | 0.9290 | 0.9214 | 0.9134 | 0.9049 | 0.8960 | 0.8867 | 0.8769 | 0.8066 |
| 8 | 0.9755 | 0.9721 | 0.9683 | 0.9642 | 0.9597 | 0.9549 | 0.9497 | 0.9442 | 0.9382 | 0.9319 |

**Table 3:** (continued)

| $x$ | $\lambda = 4.1$ | 4.2 | 4.3 | 4.4 | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 5.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 0.9905 | 0.9889 | 0.9871 | 0.9851 | 0.9829 | 0.9805 | 0.9778 | 0.9749 | 0.9717 | 0.9682 |
| 10 | 0.9966 | 0.9959 | 0.9952 | 0.9943 | 0.9933 | 0.9922 | 0.9910 | 0.9896 | 0.9880 | 0.9863 |
| 11 | 0.9989 | 0.9986 | 0.9983 | 0.9980 | 0.9976 | 0.9971 | 0.9966 | 0.9960 | 0.9953 | 0.9945 |
| 12 | 0.9997 | 0.9996 | 0.9995 | 0.9993 | 0.9992 | 0.9990 | 0.9988 | 0.9986 | 0.9983 | 0.9980 |
| 13 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9997 | 0.9997 | 0.9996 | 0.9995 | 0.9994 | 0.9993 |
| 14 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 |
| 15 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 |

| $x$ | $\lambda = 5.1$ | 5.2 | 5.3 | 5.4 | 5.5 | 5.6 | 5.7 | 5.8 | 5.9 | 6.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0061 | 0.0055 | 0.0050 | 0.0045 | 0.0041 | 0.0037 | 0.0033 | 0.0030 | 0.0027 | 0.0025 |
| 1 | 0.0372 | 0.0342 | 0.0314 | 0.0266 | 0.0244 | 0.0244 | 0.0224 | 0.0206 | 0.0189 | 0.0174 |
| 2 | 0.1165 | 0.1088 | 0.1016 | 0.0948 | 0.0884 | 0.0824 | 0.0768 | 0.0715 | 0.0666 | 0.0620 |
| 3 | 0.2513 | 0.2381 | 0.2254 | 0.2133 | 0.0217 | 0.1906 | 0.1800 | 0.1700 | 0.1604 | 0.1512 |
| 4 | 0.4231 | 0.4061 | 0.3895 | 0.3733 | 0.3575 | 0.3422 | 0.3272 | 0.3127 | 0.2987 | 0.2851 |
| 5 | 0.5984 | 0.5809 | 0.5635 | 0.5461 | 0.5289 | 0.5119 | 0.4950 | 0.4783 | 0.4619 | 0.4457 |
| 6 | 0.7474 | 0.7324 | 0.7171 | 0.7017 | 0.6860 | 0.6703 | 0.6544 | 0.6384 | 0.6224 | 0.6063 |
| 7 | 0.8560 | 0.8449 | 0.8335 | 0.8217 | 0.8095 | 0.7970 | 0.7841 | 0.7710 | 0.7576 | 0.7440 |
| 8 | 0.9252 | 0.9181 | 0.9106 | 0.9027 | 0.8944 | 0.8857 | 0.8766 | 0.8672 | 0.8574 | 0.8472 |
| 9 | 0.9644 | 0.9603 | 0.9559 | 0.9512 | 0.9462 | 0.9409 | 0.9352 | 0.9292 | 0.9228 | 0.9161 |
| 10 | 0.9844 | 0.9823 | 0.9800 | 0.9775 | 0.9747 | 0.9718 | 0.9686 | 0.9651 | 0.9614 | 0.9574 |
| 11 | 0.9937 | 0.9927 | 0.9916 | 0.9904 | 0.9890 | 0.9875 | 0.9859 | 0.9841 | 0.9821 | 0.9799 |
| 12 | 0.9976 | 0.9972 | 0.9967 | 0.9962 | 0.9955 | 0.9949 | 0.9941 | 0.9932 | 0.9922 | 0.9912 |
| 13 | 0.9992 | 0.9990 | 0.9988 | 0.9986 | 0.9983 | 0.9980 | 0.9977 | 0.9973 | 0.9969 | 0.9964 |
| 14 | 0.9997 | 0.9997 | 0.9996 | 0.9995 | 0.9994 | 0.9993 | 0.9991 | 0.9990 | 0.9988 | 0.9986 |
| 15 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9998 | 0.9997 | 0.9996 | 0.9996 | 0.9995 |
| 16 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 |
| 17 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |

| $x$ | $\lambda = 6.1$ | 6.2 | 6.3 | 6.4 | 6.5 | 6.6 | 6.7 | 6.8 | 6.9 | 7.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0022 | 0.0020 | 0.0018 | 0.0017 | 0.0015 | 0.0014 | 0.0012 | 0.0011 | 0.0010 | 0.0009 |
| 1 | 0.0159 | 0.0146 | 0.0134 | 0.0123 | 0.0113 | 0.0103 | 0.0095 | 0.0087 | 0.0080 | 0.0073 |
| 2 | 0.0577 | 0.0536 | 0.0498 | 0.0463 | 0.0430 | 0.0400 | 0.0371 | 0.0344 | 0.0320 | 0.0296 |
| 3 | 0.1425 | 0.1342 | 0.1264 | 0.1189 | 0.1118 | 0.1052 | 0.0988 | 0.0928 | 0.0871 | 0.0818 |
| 4 | 0.2719 | 0.2592 | 0.2469 | 0.2351 | 0.2237 | 0.2127 | 0.2022 | 0.1920 | 0.1823 | 0.1730 |
| 5 | 0.4298 | 0.4141 | 0.3988 | 0.3837 | 0.3690 | 0.3547 | 0.3406 | 0.3270 | 0.3137 | 0.3007 |
| 6 | 0.5902 | 0.5742 | 0.5582 | 0.5423 | 0.5265 | 0.5108 | 0.4953 | 0.4799 | 0.4647 | 0.4497 |
| 7 | 0.7301 | 0.7160 | 0.7017 | 0.6873 | 0.6728 | 0.6581 | 0.6433 | 0.6285 | 0.6136 | 0.5987 |
| 8 | 0.8367 | 0.8259 | 0.8148 | 0.8033 | 0.7916 | 0.7796 | 0.7673 | 0.7548 | 0.7420 | 0.7291 |
| 9 | 0.9090 | 0.9016 | 0.8939 | 0.8858 | 0.8774 | 0.8686 | 0.8596 | 0.8502 | 0.8405 | 0.8305 |
| 10 | 0.9531 | 0.9486 | 0.9437 | 0.9386 | 0.9332 | 0.9274 | 0.9214 | 0.9151 | 0.9084 | 0.9015 |
| 11 | 0.9776 | 0.9750 | 0.9723 | 0.9693 | 0.9661 | 0.9627 | 0.9591 | 0.9552 | 0.9510 | 0.9467 |

**Table 3:** (continued)

| *x* | *λ* = 6.1 | 6.2 | 6.3 | 6.4 | 6.5 | 6.6 | 6.7 | 6.8 | 6.9 | 7.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 0.9900 | 0.9887 | 0.9873 | 0.9857 | 0.9840 | 0.9821 | 0.9801 | 0.9779 | 0.9755 | 0.9730 |
| 13 | 0.9958 | 0.9952 | 0.9945 | 0.9937 | 0.9929 | 0.9920 | 0.9909 | 0.9898 | 0.9885 | 0.9872 |
| 14 | 0.9984 | 0.9981 | 0.9978 | 0.9974 | 0.9970 | 0.9966 | 0.9961 | 0.9956 | 0.9950 | 0.9943 |
| 15 | 0.9994 | 0.9993 | 0.9992 | 0.9990 | 0.9988 | 0.9986 | 0.9984 | 0.9982 | 0.9979 | 0.9976 |
| 16 | 0.9998 | 0.9997 | 0.9997 | 0.9996 | 0.9996 | 0.9995 | 0.9994 | 0.9993 | 0.9992 | 0.9990 |
| 17 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9998 | 0.9997 | 0.9997 | 0.9996 |
| 18 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 19 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

| *x* | *λ* = 7.1 | 7.2 | 7.3 | 7.4 | 7.5 | 7.6 | 7.7 | 7.8 | 7.9 | 8.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0008 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0003 |
| 1 | 0.0067 | 0.0061 | 0.0056 | 0.0051 | 0.0047 | 0.0043 | 0.0039 | 0.0036 | 0.0033 | 0.0030 |
| 2 | 0.0275 | 0.0255 | 0.0236 | 0.0219 | 0.0203 | 0.0188 | 0.0174 | 0.0161 | 0.0149 | 0.0138 |
| 3 | 0.0767 | 0.0719 | 0.0674 | 0.0632 | 0.0591 | 0.0554 | 0.0518 | 0.0485 | 0.0453 | 0.0424 |
| 4 | 0.1641 | 0.1555 | 0.1473 | 0.1395 | 0.1321 | 0.1249 | 0.1181 | 0.1117 | 0.1055 | 0.0996 |
| 5 | 0.2881 | 0.2759 | 0.2640 | 0.2526 | 0.2414 | 0.2307 | 0.2203 | 0.2103 | 0.2006 | 0.1912 |
| 6 | 0.4349 | 0.4204 | 0.4060 | 0.3920 | 0.3782 | 0.3646 | 0.3514 | 0.3384 | 0.3257 | 0.3134 |
| 7 | 0.5838 | 0.5689 | 0.5541 | 0.5393 | 0.5246 | 0.5100 | 0.4596 | 0.4812 | 0.4670 | 0.4530 |
| 8 | 0.7160 | 0.7027 | 0.6892 | 0.6757 | 0.6620 | 0.6482 | 0.6343 | 0.6204 | 0.6065 | 0.5925 |
| 9 | 0.8202 | 0.8096 | 0.7988 | 0.7877 | 0.7764 | 0.7649 | 0.7531 | 0.7411 | 0.7290 | 0.7166 |
| 10 | 0.8942 | 0.8867 | 0.8788 | 0.8707 | 0.8622 | 0.8535 | 0.8445 | 0.8352 | 0.8257 | 0.8159 |
| 11 | 0.9420 | 0.9371 | 0.9319 | 0.9265 | 0.9208 | 0.9148 | 0.9085 | 0.9020 | 0.8952 | 0.8881 |
| 12 | 0.9703 | 0.9673 | 0.9642 | 0.9609 | 0.9573 | 0.9536 | 0.9496 | 0.9454 | 0.9309 | 0.9362 |
| 13 | 0.9857 | 0.9841 | 0.9824 | 0.9805 | 0.9784 | 0.9762 | 0.9739 | 0.9714 | 0.9087 | 0.9658 |
| 14 | 0.9935 | 0.9927 | 0.9918 | 0.9908 | 0.9897 | 0.9886 | 0.9873 | 0.9859 | 0.9844 | 0.9827 |
| 15 | 0.9972 | 0.9969 | 0.9964 | 0.9959 | 0.9954 | 0.9948 | 0.9941 | 0.9934 | 0.9926 | 0.9918 |
| 16 | 0.9989 | 0.9987 | 0.9985 | 0.9983 | 0.9980 | 0.9978 | 0.9974 | 0.9971 | 0.9967 | 0.9963 |
| 17 | 0.9996 | 0.9995 | 0.9994 | 0.9993 | 0.9992 | 0.9991 | 0.9989 | 0.9988 | 0.9986 | 0.9984 |
| 18 | 0.9998 | 0.9998 | 0.9998 | 0.9997 | 0.9997 | 0.9996 | 0.9996 | 0.9995 | 0.9994 | 0.9993 |
| 19 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9998 | 0.9997 |
| 20 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 21 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

| *x* | *λ* = 8.1 | 8.2 | 8.3 | 8.4 | 8.5 | 8.6 | 8.7 | 8.8 | 8.9 | 9.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0003 | 0.0003 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0001 | 0.0001 |
| 1 | 0.0028 | 0.0025 | 0.0023 | 0.0021 | 0.0019 | 0.0018 | 0.0016 | 0.0015 | 0.0014 | 0.0012 |
| 2 | 0.0127 | 0.0118 | 0.0109 | 0.0100 | 0.0093 | 0.0086 | 0.0079 | 0.0073 | 0.0068 | 0.0062 |
| 3 | 0.0396 | 0.0370 | 0.0346 | 0.0323 | 0.0301 | 0.0281 | 0.0262 | 0.0244 | 0.0228 | 0.0212 |
| 4 | 0.0940 | 0.0887 | 0.0837 | 0.0789 | 0.0744 | 0.0701 | 0.0660 | 0.0621 | 0.0584 | 0.0550 |
| 5 | 0.1822 | 0.1736 | 0.1653 | 0.1573 | 0.1496 | 0.1422 | 0.1352 | 0.1284 | 0.1219 | 0.1157 |
| 6 | 0.3013 | 0.2896 | 0.2781 | 0.2670 | 0.2562 | 0.2457 | 0.2355 | 0.2256 | 0.2160 | 0.2068 |
| 7 | 0.4391 | 0.4254 | 0.4119 | 0.3987 | 0.3856 | 0.3728 | 0.3602 | 0.3478 | 0.3357 | 0.3239 |

**Table 3:** (continued)

| *x* | *λ* = 8.1 | 8.2 | 8.3 | 8.4 | 8.5 | 8.6 | 8.7 | 8.8 | 8.9 | 9.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 0.5786 | 0.5647 | 0.5507 | 0.5369 | 0.5231 | 0.5094 | 0.4958 | 0.4823 | 0.4689 | 0.4557 |
| 9 | 0.7041 | 0.6915 | 0.6788 | 0.6659 | 0.6530 | 0.6400 | 0.6269 | 0.6137 | 0.6006 | 0.5874 |
| 10 | 0.8058 | 0.7955 | 0.7850 | 0.7743 | 0.7634 | 0.7522 | 0.7409 | 0.7294 | 0.7178 | 0.7060 |
| 11 | 0.8807 | 0.8731 | 0.8652 | 0.8571 | 0.8487 | 0.8400 | 0.8311 | 0.8220 | 0.8126 | 0.8030 |
| 12 | 0.9313 | 0.9261 | 0.9207 | 0.9150 | 0.9091 | 0.9029 | 0.8965 | 0.8898 | 0.8829 | 0.8758 |
| 13 | 0.9628 | 0.9595 | 0.9561 | 0.9524 | 0.9486 | 0.9445 | 0.9403 | 0.9358 | 0.9311 | 0.9261 |
| 14 | 0.9810 | 0.9791 | 0.9771 | 0.9749 | 0.9726 | 0.9701 | 0.9675 | 0.9647 | 0.9617 | 0.9585 |
| 15 | 0.9908 | 0.9898 | 0.9887 | 0.9875 | 0.9862 | 0.9848 | 0.9832 | 0.9816 | 0.9798 | 0.9780 |
| 16 | 0.9958 | 0.9953 | 0.9947 | 0.9941 | 0.9934 | 0.9926 | 0.9918 | 0.9909 | 0.9899 | 0.9889 |
| 17 | 0.9982 | 0.9979 | 0.9977 | 0.9973 | 0.9970 | 0.9966 | 0.9962 | 0.9957 | 0.9952 | 0.9947 |
| 18 | 0.9992 | 0.9991 | 0.9990 | 0.9989 | 0.9987 | 0.9985 | 0.9983 | 0.9981 | 0.9978 | 0.9976 |
| 19 | 0.9997 | 0.9997 | 0.9996 | 0.9995 | 0.9995 | 0.9994 | 0.9993 | 0.9992 | 0.9991 | 0.9989 |
| 20 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9997 | 0.9997 | 0.9996 | 0.9996 |
| 21 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 |
| 22 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 |

| *x* | *λ* = 9.1 | 9.2 | 9.3 | 9.4 | 9.5 | 9.6 | 9.7 | 9.8 | 9.9 | 10.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.00001 | 0.0001 | 0.0001 | 0.0000 |
| 1 | 0.0011 | 0.0010 | 0.0009 | 0.0009 | 0.0008 | 0.0007 | 0.0007 | 0.0006 | 0.0005 | 0.0005 |
| 2 | 0.0058 | 0.0053 | 0.0049 | 0.0045 | 0.0042 | 0.0038 | 0.0035 | 0.0033 | 0.0030 | 0.0028 |
| 3 | 0.0198 | 0.0184 | 0.0172 | 0.0160 | 0.0149 | 0.0138 | 0.0129 | 0.0120 | 0.0111 | 0.0103 |
| 4 | 0.0517 | 0.0486 | 0.0456 | 0.0429 | 0.0403 | 0.0378 | 0.0355 | 0.0333 | 0.0312 | 0.0293 |
| 5 | 0.1098 | 0.1041 | 0.0986 | 0.0935 | 0.0885 | 0.0838 | 0.0793 | 0.0750 | 0.0710 | 0.0671 |
| 6 | 0.1978 | 0.1892 | 0.1808 | 0.1727 | 0.1649 | 0.1574 | 0.1502 | 0.1433 | 0.1366 | 0.1301 |
| 7 | 0.3123 | 0.3010 | 0.2900 | 0.2792 | 0.2687 | 0.2584 | 0.2485 | 0.2388 | 0.2294 | 0.2202 |
| 8 | 0.4126 | 0.4296 | 0.4168 | 0.4042 | 0.3918 | 0.3798 | 0.3676 | 0.3558 | 0.3442 | 0.3328 |
| 9 | 0.5742 | 0.5611 | 0.5479 | 0.5349 | 0.5218 | 0.5089 | 0.4960 | 0.4832 | 0.4705 | 0.4579 |
| 10 | 0.6941 | 0.6820 | 0.6699 | 0.6576 | 0.6453 | 0.6329 | 0.6205 | 0.6080 | 0.5955 | 0.5830 |
| 11 | 0.7932 | 0.8732 | 0.7730 | 0.7626 | 0.7520 | 0.7412 | 0.7303 | 0.7193 | 0.7081 | 0.6968 |
| 12 | 0.8684 | 0.8607 | 0.8529 | 0.8448 | 0.8364 | 0.8279 | 0.8191 | 0.8101 | 0.8009 | 0.7916 |
| 13 | 0.9210 | 0.9156 | 0.9100 | 0.9042 | 0.8981 | 0.8919 | 0.8853 | 0.8786 | 0.8716 | 0.8615 |
| 14 | 0.9552 | 0.9517 | 0.9480 | 0.9441 | 0.9400 | 0.9357 | 0.9312 | 0.9265 | 0.9216 | 0.9165 |
| 15 | 0.9760 | 0.9738 | 0.9715 | 0.9691 | 0.9665 | 0.9638 | 0.9609 | 0.9579 | 0.9546 | 0.9513 |
| 16 | 0.9878 | 0.9865 | 0.9852 | 0.9838 | 0.9823 | 0.9806 | 0.9789 | 0.9770 | 0.9751 | 0.9730 |
| 17 | 0.9941 | 0.9934 | 0.9927 | 0.9919 | 0.9911 | 0.9902 | 0.9892 | 0.9881 | 0.9870 | 0.9857 |
| 18 | 0.9973 | 0.9969 | 0.9966 | 0.9962 | 0.9957 | 0.9952 | 0.9947 | 0.9941 | 0.9935 | 0.9928 |
| 19 | 0.9988 | 0.9986 | 0.9985 | 0.9983 | 0.9980 | 0.9978 | 0.9975 | 0.9972 | 0.9969 | 0.9965 |
| 20 | 0.9995 | 0.9994 | 0.9993 | 0.9992 | 0.9991 | 0.9990 | 0.9989 | 0.9987 | 0.9986 | 0.9984 |
| 21 | 0.9998 | 0.9998 | 0.9997 | 0.9997 | 0.9996 | 0.9996 | 0.9995 | 0.9995 | 0.9994 | 0.9994 |
| 22 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9998 | 0.9998 | 0.9998 | 0.9997 | 0.9997 |
| 23 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 24 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**Table 3:** (continued)

| $x$ | $\lambda = 11$ | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 0.0012 | 0.0005 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 3 | 0.0049 | 0.0023 | 0.0011 | 0.0005 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 4 | 0.0151 | 0.0076 | 0.0037 | 0.0018 | 0.0009 | 0.0004 | 0.0002 | 0.0001 | 0.0000 | 0.0000 |
| 5 | 0.0375 | 0.0203 | 0.0107 | 0.0055 | 0.0028 | 0.0014 | 0.0007 | 0.0003 | 0.0002 | 0.0001 |
| 6 | 0.0786 | 0.0458 | 0.0259 | 0.0142 | 0.0076 | 0.0040 | 0.0021 | 0.0010 | 0.0005 | 0.0003 |
| 7 | 0.1432 | 0.0895 | 0.0540 | 0.0316 | 0.0180 | 0.0100 | 0.0054 | 0.0029 | 0.0015 | 0.0008 |
| 8 | 0.2320 | 0.1550 | 0.0998 | 0.0621 | 0.0374 | 0.0220 | 0.0126 | 0.0071 | 0.0039 | 0.0021 |
| 9 | 0.3405 | 0.2424 | 0.1658 | 0.1094 | 0.0699 | 0.0433 | 0.0261 | 0.0154 | 0.0089 | 0.0050 |
| 10 | 0.4599 | 0.3472 | 0.2517 | 0.1757 | 0.1185 | 0.0774 | 0.0491 | 0.0304 | 0.0183 | 0.0108 |
| 11 | 0.5793 | 0.4616 | 0.3532 | 0.2600 | 0.1848 | 0.1270 | 0.0847 | 0.0549 | 0.0347 | 0.0214 |
| 12 | 0.6887 | 0.5760 | 0.4631 | 0.3585 | 0.2676 | 0.1931 | 0.1350 | 0.0917 | 0.0606 | 0.0390 |
| 13 | 0.7813 | 0.6815 | 0.5730 | 0.4644 | 0.3032 | 0.2745 | 0.2009 | 0.1426 | 0.0984 | 0.0661 |
| 14 | 0.8540 | 0.7720 | 0.6751 | 0.5704 | 0.4657 | 0.3675 | 0.2808 | 0.2081 | 0.1497 | 0.1049 |
| 15 | 0.9074 | 0.8444 | 0.7636 | 0.6694 | 0.5681 | 0.4667 | 0.3715 | 0.2867 | 0.2148 | 0.1565 |
| 16 | 0.9441 | 0.8987 | 0.8355 | 0.7559 | 0.6641 | 0.5660 | 0.4677 | 0.3751 | 0.2920 | 0.2211 |
| 17 | 0.9678 | 0.9370 | 0.8905 | 0.8272 | 0.7489 | 0.6593 | 0.5640 | 0.4686 | 0.3784 | 0.2970 |
| 18 | 0.9823 | 0.9626 | 0.9302 | 0.8826 | 0.8195 | 0.7423 | 0.6550 | 0.5622 | 0.4695 | 0.3814 |
| 19 | 0.9907 | 0.9787 | 0.9573 | 0.9325 | 0.8752 | 0.8122 | 0.7363 | 0.6509 | 0.5606 | 0.4703 |
| 20 | 0.9953 | 0.9884 | 0.9750 | 0.9521 | 0.9170 | 0.8682 | 0.8055 | 0.7307 | 0.6472 | 0.5591 |
| 21 | 0.9977 | 0.9939 | 0.9859 | 0.9712 | 0.9469 | 0.9108 | 0.8615 | 0.7991 | 0.7255 | 0.6437 |
| 22 | 0.9990 | 0.9970 | 0.9924 | 0.9833 | 0.9673 | 0.9418 | 0.9047 | 0.8551 | 0.7931 | 0.7206 |
| 23 | 0.9995 | 0.9985 | 0.9960 | 0.9907 | 0.9805 | 0.9633 | 0.9367 | 0.8989 | 0.8490 | 0.7875 |
| 24 | 0.9998 | 0.9993 | 0.9980 | 0.9950 | 0.9888 | 0.9777 | 0.9594 | 0.9317 | 0.8933 | 0.8432 |
| 25 | 0.9999 | 0.9997 | 0.9990 | 0.9974 | 0.9938 | 0.9869 | 0.9748 | 0.9554 | 0.9269 | 0.8878 |
| 26 | 1.0000 | 0.9999 | 0.9995 | 0.9987 | 0.9967 | 0.9925 | 0.9848 | 0.9718 | 0.9514 | 0.9221 |
| 27 | 1.0000 | 0.9999 | 0.9998 | 0.9994 | 0.9983 | 0.9959 | 0.9912 | 0.9827 | 0.9687 | 0.9475 |
| 28 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9991 | 0.9978 | 0.9950 | 0.9897 | 0.9805 | 0.9657 |
| 29 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9989 | 0.9973 | 0.9941 | 0.9881 | 0.9782 |
| 30 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9994 | 0.9986 | 0.9967 | 0.9930 | 0.9865 |
| 31 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9993 | 0.9982 | 0.9960 | 0.9919 |
| 32 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9996 | 0.9990 | 0.9978 | 0.9953 |
| 33 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9995 | 0.9988 | 0.9973 |
| 34 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9994 | 0.9985 |
| 35 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9997 | 0.9992 |
| 36 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 | 0.9996 |
| 37 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 | 0.9998 |
| 38 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9999 |
| 39 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**Table 4:** Normal probabilities

**Entry: Probability** = $\int_0^x \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt$

| $x$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1627 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2518 | 0.2549 |
| 0.7 | 0.2580 | 0.2612 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2882 | 0.2910 | 0.2939 | 0.2967 | 0.2996 | 0.3023 | 0.3051 | 0.3079 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3290 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3414 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3888 | 0.3888 | 0.3906 | 0.3925 | 0.3943 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4146 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4278 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4453 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4610 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4648 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4762 | 0.4767 |
| 2.0 | 0.4773 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.2 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4914 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4933 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4986 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |

**Table 5:** Student-*t* table, right tail

**Entry: $t_{v,\alpha}$ where $\int_{t_{v,\alpha}}^{\infty} f(t_v)\mathrm{d}t_v = \alpha$**
**and $f(t_v)$ is the density of a Student-*t* with $v$ degrees of freedom**

| $v$ | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.025$ | $\alpha = 0.01$ | $\alpha = 0.005$ |
|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| $\infty$ | 1.282 | 1.645 | 1.966 | 2.326 | 2.576 |

**Table 6:** The chi-square table, right tail

**Entry:** $\chi^2_{v,\alpha}$ **where** $\int_{\chi_{v,\alpha}}^{\infty} f(\chi_v^2)\mathrm{d}\chi_v^2 = \alpha$

**with** $f(\chi_v^2)$ **being the density of a chi-square with** $v$ **degrees of freedom**

| $v$ | $\alpha$ = 0.995 | 0.99 | 0.975 | 0.95 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0000 | 0.0002 | 0.0010 | 0.0039 | 0.2.71 | 3.84 | 5.02 | 6.63 | 7.88 | 10.83 |
| 2 | 0.0100 | 00201 | 0.0506 | 0.1030 | 0.4.61 | 5.99 | 7.38 | 9.21 | 10.60 | 13.81 |
| 3 | 0.0717 | 0.1148 | 0.2160 | 0.3520 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 | 16.27 |
| 4 | 0.2070 | 0.2970 | 0.4844 | 0.7110 | 7.78 | 9.49 | 11.14 | 13.26 | 14.86 | 18.47 |
| 5 | 0.412 | 0.5543 | 0.831 | 1.15 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 | 20.52 |
| 6 | 0.676 | 0.872 | 1.24 | 1.64 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 | 22.46 |
| 7 | 0.989 | 1.24 | 1.69 | 2.17 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 | 24.32 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 | 26.12 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 | 27.88 |
| 10 | 2.16 | 2.56 | 3.25 | 3.94 | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 | 29.59 |
| 11 | 2.60 | 3.05 | 3.82 | 4.57 | 17.28 | 19.68 | 21.92 | 24.73 | 26.76 | 31.26 |
| 12 | 3.07 | 3.57 | 4.40 | 5.23 | 18.55 | 21.03 | 23.34 | 26.22 | 28.30 | 32.91 |
| 13 | 3.57 | 4.11 | 5.01 | 5.89 | 19.81 | 22.36 | 24.74 | 27.69 | 29.82 | 34.53 |
| 14 | 4.07 | 4.66 | 5.63 | 6.57 | 21.06 | 23.68 | 26.12 | 29.14 | 31.32 | 36.12 |
| 15 | 4.60 | 5.23 | 6.26 | 7.26 | 22.31 | 25.00 | 27.49 | 30.58 | 32.80 | 37.70 |
| 16 | 5.14 | 5.81 | 6.91 | 7.96 | 23.54 | 26.30 | 28.85 | 32.00 | 34.27 | 39.25 |
| 17 | 5.70 | 6.41 | 7.56 | 8.67 | 24.77 | 27.59 | 30.19 | 33.41 | 35.72 | 30.79 |
| 18 | 6.26 | 7.01 | 8.23 | 9.39 | 25.99 | 28.87 | 31.53 | 34.81 | 37.16 | 42.31 |
| 19 | 6.84 | 7.63 | 8.91 | 10.12 | 27.20 | 30.14 | 32.85 | 36.19 | 38.58 | 43.82 |
| 20 | 7.43 | 8.26 | 9.59 | 10.85 | 28.41 | 31.41 | 34.17 | 37.57 | 40.00 | 45.31 |
| 21 | 8.03 | 8.90 | 10.28 | 11.59 | 29.62 | 32.67 | 35.48 | 38.93 | 41.40 | 46.80 |
| 22 | 8.64 | 9.54 | 10.98 | 12.34 | 30.81 | 33.92 | 36.78 | 40.29 | 42.80 | 48.27 |
| 23 | 9.26 | 10.20 | 11.69 | 13.09 | 32.01 | 35.17 | 38.08 | 41.64 | 44.18 | 49.73 |
| 24 | 9.89 | 10.86 | 12.40 | 13.85 | 33.20 | 36.42 | 39.36 | 42.98 | 45.56 | 51.18 |
| 25 | 10.52 | 11.52 | 13.12 | 14.61 | 34.38 | 37.65 | 40.65 | 44.31 | 46.93 | 52.62 |
| 26 | 11.16 | 12.20 | 13.84 | 15.38 | 35.56 | 38.89 | 41.92 | 45.64 | 48.29 | 54.05 |
| 27 | 11.81 | 12.88 | 14.57 | 16.15 | 36.74 | 40.11 | 43.19 | 46.96 | 49.64 | 55.48 |
| 28 | 12.46 | 13.56 | 15.31 | 16.93 | 37.92 | 41.34 | 44.46 | 48.28 | 50.99 | 56.89 |
| 29 | 13.12 | 14.26 | 16.05 | 17.71 | 39.09 | 42.56 | 45.72 | 49.59 | 52.34 | 58.30 |
| 30 | 13.79 | 14.95 | 16.79 | 18.49 | 40.26 | 43.77 | 46.98 | 50.89 | 53.67 | 59.70 |
| 40 | 20.71 | 22.16 | 24.43 | 26.51 | 51.81 | 55.76 | 59.34 | 63.69 | 66.77 | 73.40 |
| 50 | 27.99 | 29.71 | 32.36 | 34.76 | 63.17 | 67.50 | 71.42 | 76.15 | 79.49 | 86.66 |
| 60 | 35.53 | 37.48 | 40.48 | 43.19 | 74.40 | 79.08 | 83.30 | 88.38 | 91.95 | 99.61 |
| 70 | 43.28 | 45.44 | 48.76 | 51.74 | 85.53 | 90.53 | 95.02 | 100.4 | 104.2 | 112.3 |
| 80 | 51.17 | 53.54 | 57.15 | 60.39 | 96.58 | 101.9 | 106.6 | 112.3 | 116.3 | 124.8 |
| 90 | 59.20 | 61.75 | 65.75 | 69.13 | 107.6 | 113.1 | 118.1 | 124.1 | 128.3 | 137.2 |
| 100 | 67.33 | 70.06 | 74.22 | 77.93 | 118.5 | 124.3 | 129.6 | 135.8 | 140.2 | 149.4 |

**Table 7:** $F$-distribution, right tail, 5% points

**Entry: $F_{v_1,v_2,0.05}$ where $\int_{F_{v_1,v_2,0.05}}^{\infty} f(F_{v_1,v_2})\mathrm{d}F_{v_1,v2} = 0.05$**
**with $f(F_{v_1,v_2})$ being the density of $F$-variable with $v_1$ and $v_2$ degrees of freedom**

| $v_2$ | $v_1 = 1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 10 | 12 | 24 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 241.9 | 243.9 | 249.0 | 254.3 |
| 2 | 18.5 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.5 | 19.5 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.79 | 8.74 | 8.64 | 8.53 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 5.96 | 5.91 | 5.77 | 5.63 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.74 | 4.68 | 4.53 | 4.36 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.06 | 4.00 | 3.84 | 3.67 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.37 | 3.64 | 3.57 | 3.41 | 3.23 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.35 | 3.28 | 3.12 | 2.93 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.38 | 3.37 | 3.29 | 3.23 | 3.14 | 3.07 | 2.90 | 2.71 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 2.98 | 2.91 | 2.74 | 2.54 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.85 | 2.79 | 2.61 | 2.40 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.75 | 2.69 | 2.51 | 2.30 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.67 | 2.60 | 2.42 | 2.21 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.60 | 2.53 | 2.35 | 2.13 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.54 | 2.48 | 2.29 | 2.07 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.49 | 2.42 | 2.24 | 2.01 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.45 | 2.38 | 2.19 | 1.96 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.41 | 2.34 | 2.15 | 1.92 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.38 | 2.31 | 2.11 | 1.88 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.35 | 2.28 | 2.08 | 1.84 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.32 | 2.25 | 2.05 | 1.81 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.30 | 2.23 | 2.03 | 1.78 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.27 | 2.20 | 2.00 | 1.76 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.25 | 2.18 | 1.98 | 1.73 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.24 | 2.16 | 1.96 | 1.71 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.22 | 2.15 | 1.95 | 1.69 |
| 27 | 4.21 | 3.25 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.20 | 2.13 | 1.93 | 1.67 |
| 28 | 4.20 | 3.34 | 2.96 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.19 | 2.12 | 1.91 | 1.65 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.18 | 2.10 | 1.90 | 1.64 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.16 | 2.09 | 1.89 | 1.62 |
| 32 | 4.15 | 3.29 | 2.90 | 1.67 | 2.51 | 2.40 | 2.31 | 2.24 | 2.14 | 2.07 | 1.86 | 1.59 |
| 34 | 4.13 | 3.28 | 2.88 | 2.65 | 2.49 | 2.38 | 2.29 | 2.23 | 2.12 | 2.05 | 1.84 | 1.57 |
| 36 | 4.11 | 3.26 | 2.87 | 2.63 | 2.48 | 2.36 | 2.28 | 2.21 | 2.11 | 2.03 | 1.82 | 1.55 |
| 38 | 4.10 | 3.24 | 2.85 | 2.62 | 2.46 | 2.35 | 2.26 | 2.19 | 2.09 | 2.02 | 1.81 | 1.53 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.08 | 2.00 | 1.79 | 1.51 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 1.99 | 1.92 | 1.70 | 1.89 |
| 100 | 3.92 | 3.07 | 2.63 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.91 | 1.83 | 1.61 | 1.25 |
| $\infty$ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.83 | 1.75 | 1.52 | 1.00 |

**Table 8:** $F$-distribution, right tail, 1% points

**Entry:** $F_{\nu_1,\nu_2,0.01}$ **where** $\int_{F_{\nu_1,\nu_2,0.01}}^{\infty} f(F_{\nu_1,\nu_2})\mathrm{d}F_{\nu_1,\nu 2} = 0.01$
**with** $f(F_{\nu_1,\nu_2})$ **being the density of** $F$**-variable with** $\nu_1$ **and** $\nu_2$ **degrees of freedom**

| $\nu_2$ | $\nu_1 = 1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 10 | 12 | 24 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4052 | 4999.5 | 5403 | 5625 | 5764 | 5859 | 5928 | 5981 | 6056 | 6106 | 6235 | 6366 |
| 2 | 98.5 | 99.0 | 99.2 | 99.2 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.5 | 99.5 |
| 3 | 34.1 | 30.8 | 29.5 | 28.7 | 28.2 | 27.9 | 27.7 | 27.5 | 27.2 | 27.1 | 26.6 | 26.1 |
| 4 | 21.2 | 18.0 | 16.7 | 16.0 | 15.5 | 15.2 | 15.0 | 14.8 | 14.5 | 14.4 | 13.9 | 13.5 |
| 5 | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.05 | 9.89 | 9.47 | 9.02 |
| 6 | 13.74 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.87 | 7.72 | 7.31 | 6.88 |
| 7 | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.62 | 6.47 | 6.07 | 5.65 |
| 8 | 11.26 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.81 | 5.67 | 5.28 | 4.86 |
| 9 | 10.56 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.26 | 5.11 | 4.73 | 4.31 |
| 10 | 10.04 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.85 | 4.71 | 4.33 | 3.91 |
| 11 | 0.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.54 | 4.40 | 4.02 | 3.60 |
| 12 | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.30 | 4.16 | 3.78 | 3.36 |
| 13 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.10 | 3.96 | 3.59 | 3.17 |
| 14 | 8.86 | 6.51 | 5.56 | 5.04 | 4.70 | 4.46 | 4.28 | 4.14 | 3.94 | 3.80 | 3.43 | 3.00 |
| 15 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.80 | 3.67 | 3.29 | 2.87 |
| 16 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.69 | 3.55 | 3.18 | 2.75 |
| 17 | 8.40 | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.59 | 3.46 | 3.08 | 2.65 |
| 18 | 8.20 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.51 | 3.37 | 3.00 | 2.57 |
| 19 | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.43 | 3.30 | 2.92 | 2.49 |
| 20 | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.37 | 3.23 | 2.86 | 2.42 |
| 21 | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.31 | 3.17 | 2.80 | 2.36 |
| 22 | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.26 | 3.12 | 2.75 | 2.31 |
| 23 | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.21 | 3.07 | 2.70 | 2.26 |
| 24 | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.17 | 3.03 | 2.66 | 2.21 |
| 25 | 7.77 | 5.57 | 4.68 | 4.18 | 3.86 | 3.63 | 3.46 | 3.32 | 3.13 | 2.99 | 2.62 | 2.17 |
| 26 | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.09 | 2.96 | 2.58 | 2.13 |
| 27 | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.06 | 2.93 | 2.55 | 2.10 |
| 28 | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.03 | 2.90 | 2.52 | 2.06 |
| 29 | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.00 | 2.87 | 2.49 | 2.03 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 2.98 | 2.84 | 2.47 | 2.01 |
| 32 | 7.50 | 5.34 | 4.46 | 3.97 | 3.65 | 3.43 | 3.26 | 3.13 | 2.93 | 2.80 | 2.42 | 1.96 |
| 34 | 7.45 | 5.29 | 4.42 | 3.93 | 3.61 | 3.39 | 3.22 | 3.09 | 2.90 | 2.76 | 2.38 | 1.91 |
| 36 | 7.40 | 5.25 | 4.38 | 3.89 | 3.58 | 3.35 | 3.18 | 3.05 | 2.86 | 2.72 | 2.35 | 1.87 |
| 38 | 7.35 | 5.21 | 4.34 | 3.86 | 3.54 | 3.32 | 3.15 | 3.02 | 2.83 | 2.69 | 2.32 | 1.84 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.80 | 2.66 | 2.29 | 1.80 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.63 | 2.50 | 2.12 | 1.60 |
| 120 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.47 | 2.334 | 1.95 | 1.38 |
| $\infty$ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.32 | 2.18 | 1.79 | 1.00 |

**Table 9:** Kolmogorov–Smirnov $D_n$

**Entry:** $D_{n,\alpha}$ **where** $\Pr\{D_n \geq D_{n,\alpha}\} = \alpha$

| $n \downarrow$ | $\alpha \rightarrow$ | **0.20** | **0.15** | **0.10** | **0.05** | **0.01** |
|---|---|---|---|---|---|---|
| 1 | | 0.900 | 0.925 | 0.950 | 0.975 | 0.995 |
| 2 | | 0.684 | 0.726 | 0.776 | 0.842 | 0.929 |
| 3 | | 0.565 | 0.597 | 0.642 | 0.708 | 0.828 |
| 4 | | 0.494 | 0.525 | 0.564 | 0.624 | 0.733 |
| 5 | | 0.446 | 0.474 | 0.510 | 0.565 | 0.669 |
| 6 | | 0.410 | 0.436 | 0.470 | 0.521 | 0.618 |
| 7 | | 0.381 | 0.405 | 0.438 | 0.486 | 0.577 |
| 8 | | 0.358 | 0.381 | 0.411 | 0.457 | 0.543 |
| 9 | | 0.339 | 0.360 | 0.388 | 0.432 | 0.514 |
| 10 | | 0.322 | 0.342 | 0.368 | 0.410 | 0.490 |
| 11 | | 0.307 | 0.326 | 0.352 | 0.391 | 0.468 |
| 12 | | 0.295 | 0.313 | 0.338 | 0.375 | 0.450 |
| 13 | | 0.284 | 0.302 | 0.325 | 0.361 | 0.433 |
| 14 | | 0.274 | 0.292 | 0.314 | 0.349 | 0.418 |
| 15 | | 0.266 | 0.283 | 0.304 | 0.338 | 0.404 |
| 16 | | 0.258 | 0.274 | 0.295 | 0.328 | 0.392 |
| 17 | | 0.250 | 0.266 | 0.280 | 0.318 | 0.381 |
| 18 | | 0.244 | 0.259 | 0.278 | 0.309 | 0.371 |
| 19 | | 0.237 | 0.252 | 0.272 | 0.301 | 0.363 |
| 20 | | 0.231 | 0.246 | 0.264 | 0.294 | 0.356 |
| 25 | | 0.210 | 0.220 | 0.240 | 0.270 | 0.320 |
| 30 | | 0.190 | 0.200 | 0.220 | 0.240 | 0.290 |
| 35 | | 0.180 | 0.190 | 0.210 | 0.230 | 0.270 |
| >35 | | $10.07/\sqrt{n}$ | $1.14/\sqrt{n}$ | $1.22/\sqrt{n}$ | $1.36/\sqrt{n}$ | $1.63/\sqrt{n}$ |

# References

[1] A. M. Mathai, An approximate analysis of the two-way layout, *Biometrics* **21** (1965), 376–385.

[2] A. M. Mathai, *A Handbook of Generalized Special Functions for Statistical and Physical Sciences*, Oxford University Press, Oxford, 1993.

[3] A. M. Mathai, *Jacobians of Matrix Transformations and Functions of Matrix Argument*, World Scientific Publishing, Amsterdam, 1997.

[4] A. M. Mathai, *An Introduction to Geometrical Probability: Distributional Aspects with Applications*, Gordon and Breach, Amsterdam, 1999.

[5] A. M. Mathai, A pathway to matrix variate gamma and normal densities, *Linear Algebra and Its Applications* **396** (2005), 317–328.

[6] A. M. Mathai, Some properties of Mittag–Leffler function and matrix-variate analogues: A statistical perspective, *Fractional Calculus & Applied Analysis* **13**(1) (2010), 113–132.

[7] A. M. Mathai, Stochastic models under power transformations and exponentiation, *Journal of the Indian Society for Probability and Statistics* **13** (2012), 1–19.

[8] A. M. Mathai and H. J. Haubold, Pathway model, superstatistics, Tsallis statistics and a generalized measure of entropy, *Physica A* **375** (2007), 110–122.

[9] A. M. Mathai and R. S. Katiyar, A new algorithm for non-linear least squares, *Researches in Mathematical Statistics* **207**(10) (1993), 143–147 (in Russian; English translation by the American Mathematical Society).

[10] A. M. Mathai and P. Moschopoulos, On a multivariate beta, *Statistica* **LIII**(2) (1993), 231–241.

[11] A. M. Mathai and G. Pederzoli, *Characterizations of the Normal Probability Law*, Wiley Halsted, New York and Wiley Eastern, New Delhi, 1977.

[12] A. M. Mathai and S. B. Provost, *Quadratic Forms in Random Variables: Theory and Applications*, Marcel Dekker, New York, 1992.

[13] A. M. Mathai, S. B. Provost and T. Hayakawa, *Bilinear Forms and Zonal Polynomials*, Lecture Notes in Statistics, Springer, New York, 1995.

[14] A. M. Mathai and P. N. Rathie, *Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications*, Wiley, New York and New Delhi, 1975.

[15] C. R. Rao, *Linear Statistical Inference and Its Applications*, second edition, Wiley, New York, 2001.

https://doi.org/10.1515/9783110562545-018

# Index