# PROBABILITY
# AND
# MATHEMATICAL STATISTICS

**Prasanna Sahoo**
**Department of Mathematics**
**University of Louisville**
**Louisville, KY 40292 USA**

THIS BOOK IS DEDICATED TO
AMIT
SADHNA
MY PARENTS, TEACHERS
AND
STUDENTS

# PREFACE

This book is both a tutorial and a textbook. This book presents an introduction to probability and mathematical statistics and it is intended for students already having some elementary mathematical background. It is intended for a one-year junior or senior level undergraduate or beginning graduate level course in probability theory and mathematical statistics. The book contains more material than normally would be taught in a one-year course. This should give the teacher flexibility with respect to the selection of the content and level at which the book is to be used. This book is based on over 15 years of lectures in senior level calculus based courses in probability theory and mathematical statistics at the University of Louisville.

Probability theory and mathematical statistics are difficult subjects both for students to comprehend and teachers to explain. Despite the publication of a great many textbooks in this field, each one intended to provide an improvement over the previous textbooks, this subject is still difficult to comprehend. A good set of examples makes these subjects easy to understand. For this reason alone I have included more than 350 completely worked out examples and over 165 illustrations. I give a rigorous treatment of the fundamentals of probability and statistics using mostly calculus. I have given great attention to the clarity of the presentation of the materials. In the text, theoretical results are presented as theorems, propositions or lemmas, of which as a rule rigorous proofs are given. For the few exceptions to this rule references are given to indicate where details can be found. This book contains over 450 problems of varying degrees of difficulty to help students master their problem solving skill.

In many existing textbooks, the examples following the explanation of a topic are too few in number or too simple to obtain a through grasp of the principles involved. Often, in many books, examples are presented in abbreviated form that leaves out much material between steps, and requires that students derive the omitted materials themselves. As a result, students find examples difficult to understand. Moreover, in some textbooks, examples

are often worded in a confusing manner. They do not state the problem and then present the solution. Instead, they pass through a general discussion, never revealing what is to be solved for. In this book, I give many examples to illustrate each topic. Often we provide illustrations to promote a better understanding of the topic. All examples in this book are formulated as questions and clear and concise answers are provided in step-by-step detail.

There are several good books on these subjects and perhaps there is no need to bring a new one to the market. So for several years, this was circulated as a series of typeset lecture notes among my students who were preparing for the examination 110 of the Actuarial Society of America. Many of my students encouraged me to formally write it as a book. Actuarial students will benefit greatly from this book. The book is written in simple English; this might be an advantage to students whose native language is not English.

I cannot claim that all the materials I have written in this book are mine. I have learned the subject from many excellent books, such as *Introduction to Mathematical Statistics* by Hogg and Craig, and *An Introduction to Probability Theory and Its Applications* by Feller. In fact, these books have had a profound impact on me, and my explanations are influenced greatly by these textbooks. If there are some similarities, then it is due to the fact that I could not make improvements on the original explanations. I am very thankful to the authors of these great textbooks. I am also thankful to the Actuarial Society of America for letting me use their test problems. I thank all my students in my probability theory and mathematical statistics courses from 1988 to 2005 who helped me in many ways to make this book possible in the present form. Lastly, if it weren't for the infinite patience of my wife, Sadhna, this book would never get out of the hard drive of my computer.

The author on a Macintosh computer using TEX, the typesetting system designed by Donald Knuth, typeset the entire book. The figures were generated by the author using MATHEMATICA, a system for doing mathematics designed by Wolfram Research, and MAPLE, a system for doing mathematics designed by Maplesoft. The author is very thankful to the University of Louisville for providing many internal financial grants while this book was under preparation.

Prasanna Sahoo, *Louisville*

# TABLE OF CONTENTS

# Chapter 1
# PROBABILITY OF EVENTS

## 1.1. Introduction

During his lecture in 1929, Bertrand Russel said, "*Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means.*" Most people have some vague ideas about what *probability of an event* means. The interpretation of the word *probability* involves synonyms such as chance, odds, uncertainty, prevalence, risk, expectancy etc. "*We use probability when we want to make an affirmation, but are not quite sure,*" writes J.R. Lucas.

There are many distinct interpretations of the word *probability*. A complete discussion of these interpretations will take us to areas such as philosophy, theory of algorithm and randomness, religion, etc. Thus, we will only focus on two extreme interpretations. One interpretation is due to the so-called objective school and the other is due to the subjective school.

The subjective school defines probabilities as subjective assignments based on rational thought with available information. Some subjective probabilists interpret probabilities as the degree of belief. Thus, it is difficult to interpret the probability of an event.

The objective school defines probabilities to be "*long run*" relative frequencies. This means that one should compute a probability by taking the number of favorable outcomes of an experiment and dividing it by total numbers of the possible outcomes of the experiment, and then taking the limit as the number of trials becomes large. Some statisticians object to the word "long run". The philosopher and statistician John Keynes said "*in the long run we are all dead*". The objective school uses the theory developed by

Von Mises (1928) and Kolmogorov (1965). The Russian mathematician Kolmogorov gave the solid foundation of probability theory using measure theory. The advantage of Kolmogorov's theory is that one can construct probabilities according to the rules, compute other probabilities using axioms, and then interpret these probabilities.

In this book, we will study mathematically one interpretation of probability out of many. In fact, we will study probability theory based on the theory developed by the late Kolmogorov. There are many applications of probability theory. We are studying probability theory because we would like to study mathematical statistics. Statistics is concerned with the development of methods and their applications for collecting, analyzing and interpreting quantitative data in such a way that the reliability of a conclusion based on data may be evaluated objectively by means of probability statements. Probability theory is used to evaluate the reliability of conclusions and inferences based on data. Thus, probability theory is fundamental to mathematical statistics.

For an event $A$ of a discrete sample space $S$, the probability of $A$ can be computed by using the formula

$$P(A) = \frac{N(A)}{N(S)}$$

where $N(A)$ denotes the number of elements of $A$ and $N(S)$ denotes the number of elements in the sample space $S$. For a discrete case, the probability of an event $A$ can be computed by counting the number of elements in $A$ and dividing it by the number of elements in the sample space $S$.

In the next section, we develop various counting techniques. The branch of mathematics that deals with the various counting techniques is called combinatorics.

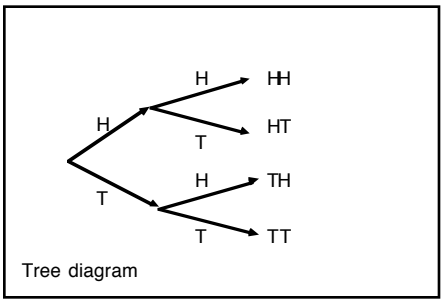
## 1.2. Counting Techniques

There are three basic counting techniques. They are multiplication rule, permutation and combination.

**1.2.1 Multiplication Rule.** If $E_1$ is an experiment with $n_1$ outcomes and $E_2$ is an experiment with $n_2$ possible outcomes, then the experiment which consists of performing $E_1$ first and then $E_2$ consists of $n_1 n_2$ possible outcomes.

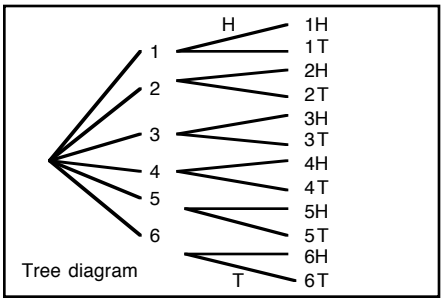
**Example 1.1.** Find the possible number of outcomes in a sequence of two tosses of a fair coin.

**Answer:** The number of possible outcomes is $2 \cdot 2 = 4$. This is evident from the following tree diagram.




**Example 1.2.** Find the number of possible outcomes of the rolling of a die and then tossing a coin.

**Answer:** Here $n_1 = 6$ and $n_2 = 2$. Thus by multiplication rule, the number of possible outcomes is 12.




**Example 1.3.** How many different license plates are possible if Kentucky uses three letters followed by three digits.

**Answer:**

$$\begin{aligned}(26)^3\,(10)^3 &= (17576)\,(1000)\\ &= 17,576,000.\end{aligned}$$

### 1.2.2. Permutation

Consider a set of 4 objects. Suppose we want to fill 3 positions with objects selected from the above 4. Then the number of possible ordered arrangements is 24 and they are

| | | | |
|---|---|---|---|
| a b c | b a c | c a b | d a b |
| a b d | b a d | c a d | d a c |
| a c b | b c a | c b a | d b c |
| a c d | b c d | c b d | d b a |
| a d c | b d a | c d b | d c a |
| a d b | b d c | c d a | d c b |

The number of possible ordered arrangements can be computed as follows: Since there are 3 positions and 4 objects, the first position can be filled in 4 different ways. Once the first position is filled the remaining 2 positions can be filled from the remaining 3 objects. Thus, the second position can be filled in 3 ways. The third position can be filled in 2 ways. Then the total number of ways 3 positions can be filled out of 4 objects is given by

$$(4)\,(3)\,(2) = 24.$$

In general, if $r$ positions are to be filled from $n$ objects, then the total number of possible ways they can be filled are given by

$$\begin{aligned} & n(n-1)(n-2)\cdots(n-r+1) \\ &= \frac{n!}{(n-r)!} \\ &= {}_nP_r. \end{aligned}$$

Thus, ${}_nP_r$ represents the number of ways $r$ positions can be filled from $n$ objects.

**Definition 1.1.** Each of the ${}_nP_r$ arrangements is called a permutation of $n$ objects taken $r$ at a time.

**Example 1.4.** How many permutations are there of all three of letters a, b, and c?

**Answer:**

$$\begin{aligned} {}_3P_3 &= \frac{n!}{(n-r)!} \\ &= \frac{3!}{0!} = 6 \end{aligned}.$$

**Example 1.5.** Find the number of permutations of $n$ distinct objects.

**Answer:**

$$ {}_nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!. $$

**Example 1.6.** Four names are drawn from the 24 members of a club for the offices of President, Vice-President, Treasurer, and Secretary. In how many different ways can this be done?

**Answer:**

$$ \begin{aligned} {}_{24}P_4 &= \frac{(24)!}{(20)!} \\ &= (24)\,(23)\,(22)\,(21) \\ &= 255,024. \end{aligned} $$

### 1.2.3. Combination

In permutation, order is important. But in many problems the order of selection is not important and interest centers only on the set of $r$ objects.

Let $c$ denote the number of subsets of size $r$ that can be selected from $n$ different objects. The $r$ objects in each set can be ordered in ${}_rP_r$ ways. Thus we have

$$ {}_nP_r = c\,({}_rP_r)\,. $$

From this, we get

$$ c = \frac{{}_nP_r}{{}_rP_r} = \frac{n!}{(n-r)!\,r!} $$

The number $c$ is denoted by $\binom{n}{r}$. Thus, the above can be written as

$$ \binom{n}{r} = \frac{n!}{(n-r)!\,r!}. $$

**Definition 1.2.** Each of the $\binom{n}{r}$ unordered subsets is called a combination of $n$ objects taken $r$ at a time.

**Example 1.7.** How many committees of two chemists and one physicist can be formed from 4 chemists and 3 physicists?

**Answer:**

$$\begin{aligned}&\binom{4}{2}\binom{3}{1}\\&= (6)\,(3)\\&= 18.\end{aligned}$$

Thus 18 different committees can be formed.

### 1.2.4. Binomial Theorem

We know from lower level mathematics courses that

$$\begin{aligned}(x+y)^2 &= x^2 + 2\,xy + y^2\\&= \binom{2}{0}x^2 + \binom{2}{1}xy + \binom{2}{2}y^2\\&= \sum_{k=0}^{2}\binom{2}{k}x^{2-k}\,y^k.\end{aligned}$$

Similarly

$$\begin{aligned}(x+y)^3 &= x^3 + 3\,x^2y + 3xy^2 + y^3\\&= \binom{3}{0}x^3 + \binom{3}{1}x^2y + \binom{3}{2}xy^2 + \binom{3}{3}y^3\\&= \sum_{k=0}^{3}\binom{3}{k}x^{3-k}\,y^k.\end{aligned}$$

In general, using induction arguments, we can show that

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}\,y^k.$$

This result is called the Binomial Theorem. The coefficient $\binom{n}{k}$ is called the binomial coefficient. A combinatorial proof of the Binomial Theorem follows. If we write $(x+y)^n$ as the $n$ times the product of the factor $(x+y)$, that is

$$(x+y)^n = (x+y)\,(x+y)\,(x+y)\cdots(x+y),$$

then the coefficient of $x^{n-k}y^k$ is $\binom{n}{k}$, that is the number of ways in which we can choose the $k$ factors providing the $y$'s.

**Remark 1.1.** In 1665, Newton discovered the Binomial Series. The Binomial Series is given by

$$\begin{aligned}(1+y)^{\alpha} &= 1+\binom{\alpha}{1}y+\binom{\alpha}{2}y^2+\cdots+\binom{\alpha}{n}y^n+\cdots \\ &= 1+\sum_{k=1}^{\infty}\binom{\alpha}{k}y^k,\end{aligned}$$

where $\alpha$ is a real number and

$$\binom{\alpha}{k}=\frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!}.$$

This $\binom{\alpha}{k}$ is called the generalized binomial coefficient.

Now, we investigate some properties of the binomial coefficients.

**Theorem 1.1.** Let $n \in \mathbf{N}$ (the set of natural numbers) and $r = 0, 1, 2, ..., n$. Then

$$\binom{n}{r}=\binom{n}{n-r}.$$

**Proof:** By direct verification, we get

$$\begin{aligned}\binom{n}{n-r} &= \frac{n!}{(n-n+r)!\,(n-r)!} \\ &= \frac{n!}{r!\,(n-r)!} \\ &= \binom{n}{r}.\end{aligned}$$

This theorem says that the binomial coefficients are symmetrical.

**Example 1.8.** Evaluate $\binom{3}{1}+\binom{3}{2}+\binom{3}{0}$.

**Answer:** Since the combinations of 3 things taken 1 at a time are 3, we get $\binom{3}{1}=3$. Similarly, $\binom{3}{0}$ is 1. By Theorem 1,

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

Hence

$$\binom{3}{1}+\binom{3}{2}+\binom{3}{0}=3+3+1=7.$$

**Theorem 1.2.** For any positive integer $n$ and $r = 1, 2, 3, ..., n$, we have

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}.$$

**Proof:**

$$\begin{aligned}
(1+y)^n &= (1+y)\,(1+y)^{n-1} \\
&= (1+y)^{n-1} + y\,(1+y)^{n-1} \\
\sum_{r=0}^{n} \binom{n}{r} y^r &= \sum_{r=0}^{n-1} \binom{n-1}{r} y^r + y \sum_{r=0}^{n-1} \binom{n-1}{r} y^r \\
&= \sum_{r=0}^{n-1} \binom{n-1}{r} y^r + \sum_{r=0}^{n-1} \binom{n-1}{r} y^{r+1}.
\end{aligned}$$

Equating the coefficients of $y^r$ from both sides of the above expression, we obtain

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

and the proof is now complete.

**Example 1.9.** Evaluate $\binom{23}{10} + \binom{23}{9} + \binom{24}{11}$.

**Answer:**

$$\begin{aligned}
&\binom{23}{10} + \binom{23}{9} + \binom{24}{11} \\
&\quad = \binom{24}{10} + \binom{24}{11} \\
&\quad = \binom{25}{11} \\
&\quad = \frac{25!}{(14)!\,(11)!} \\
&\quad = 4,457,400.
\end{aligned}$$

**Example 1.10.** Use the Binomial Theorem to show that $\sum_{r=0}^{n} (-1)^r \binom{n}{r} = 0$.

**Answer:** Using the Binomial Theorem, we get

$$(1+x)^n = \sum_{r=0}^{n} \binom{n}{r} x^r$$

for all real numbers $x$. Letting $x = -1$ in the above, we get

$$0 = \sum_{r=0}^{n} \binom{n}{r} (-1)^r.$$

**Theorem 1.3.** Let $m$ and $n$ be positive integers. Then

$$\sum_{r=0}^{k} \binom{m}{r} \binom{n}{k-r} = \binom{m+n}{k}.$$

**Proof:**

$$(1+y)^{m+n} = (1+y)^m (1+y)^n$$

$$\sum_{r=0}^{m+n} \binom{m+n}{r} y^r = \left\{ \sum_{r=0}^{m} \binom{m}{r} y^r \right\} \left\{ \sum_{r=0}^{n} \binom{n}{r} y^r \right\}.$$

Equating the coefficients of $y^k$ from the both sides of the above expression, we obtain

$$\binom{m+n}{k} = \binom{m}{0}\binom{n}{k} + \binom{m}{1}\binom{n}{k-1} + \cdots + \binom{m}{k}\binom{n}{k-k}$$

and the conclusion of the theorem follows.

**Example 1.11.** Show that

$$\sum_{r=0}^{n} \binom{n}{r}^2 = \binom{2n}{n}.$$

**Answer:** Let $k = n$ and $m = n$. Then from Theorem 3, we get

$$\sum_{r=0}^{k} \binom{m}{r} \binom{n}{k-r} = \binom{m+n}{k}$$

$$\sum_{r=0}^{n} \binom{n}{r} \binom{n}{n-r} = \binom{2n}{n}$$

$$\sum_{r=0}^{n} \binom{n}{r} \binom{n}{r} = \binom{2n}{n}$$

$$\sum_{r=0}^{n} \binom{n}{r}^2 = \binom{2n}{n}.$$

**Theorem 1.4.** Let $n$ be a positive integer and $k = 1, 2, 3, ..., n$. Then

$$\binom{n}{k} = \sum_{m=k-1}^{n-1} \binom{m}{k-1}.$$

**Proof:** In order to establish the above identity, we use the Binomial Theorem together with the following result of the elementary algebra

$$x^n - y^n = (x-y) \sum_{k=0}^{n-1} x^k y^{n-1-k}.$$

Note that

$$\begin{aligned}
\sum_{k=1}^{n} \binom{n}{k} x^k &= \sum_{k=0}^{n} \binom{n}{k} x^k - 1 \\
&= (x+1)^n - 1^n \qquad \text{by Binomial Theorem} \\
&= (x+1-1) \sum_{m=0}^{n-1} (x+1)^m \qquad \text{by above identity} \\
&= x \sum_{m=0}^{n-1} \sum_{j=0}^{m} \binom{m}{j} x^j \\
&= \sum_{m=0}^{n-1} \sum_{j=0}^{m} \binom{m}{j} x^{j+1} \\
&= \sum_{k=1}^{n} \sum_{m=k-1}^{n-1} \binom{m}{k-1} x^k.
\end{aligned}$$

Hence equating the coefficient of $x^k$, we obtain

$$\binom{n}{k} = \sum_{m=k-1}^{n-1} \binom{m}{k-1}.$$

This completes the proof of the theorem.

The following result

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{n_1+n_2+\cdots+n_m=n} \binom{n}{n_1, n_2, ..., n_m} x_1^{n_1} x_2^{n_2} \cdots x_m^{n_m}$$

is known as the multinomial theorem and it generalizes the binomial theorem. The sum is taken over all positive integers $n_1, n_2, ..., n_m$ such that $n_1 + n_2 + \cdots + n_m = n$, and

$$\binom{n}{n_1, n_2, ..., n_m} = \frac{n!}{n_1!\, n_2!, ..., n_m!}.$$

This coefficient is known as the multinomial coefficient.

### 1.3. Probability Measure

A random experiment is an experiment whose outcomes cannot be predicted with certainty. However, in most cases the collection of every possible outcome of a random experiment can be listed.

**Definition 1.3.** A sample space of a random experiment is the collection of all possible outcomes.

**Example 1.12.** What is the sample space for an experiment in which we select a rat at random from a cage and determine its sex?

**Answer:** The sample space of this experiment is

$$S = \{M,\ F\}$$

where $M$ denotes the male rat and $F$ denotes the female rat.

**Example 1.13.** What is the sample space for an experiment in which the state of Kentucky picks a three digit integer at random for its daily lottery?

**Answer:** The sample space of this experiment is

$$S = \{000, 001, 002, \cdots\cdots, 998, 999\}.$$

**Example 1.14.** What is the sample space for an experiment in which we roll a pair of dice, one red and one green?

**Answer:** The sample space $S$ for this experiment is given by

$$S = \begin{matrix} \{(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\} \end{matrix}$$

This set $S$ can be written as

$$S = \{(x, y) \,|\, 1 \le x \le 6,\ 1 \le y \le 6\}$$

where $x$ represents the number rolled on red die and $y$ denotes the number rolled on green die.

**Definition 1.4.** Each element of the sample space is called a sample point.

**Definition 1.5.** If the sample space consists of a countable number of sample points, then the sample space is said to be a countable sample space.

**Definition 1.6.** If a sample space contains an uncountable number of sample points, then it is called a continuous sample space.

An event $A$ is a subset of the sample space S. It seems obvious that if $A$ and $B$ are events in sample space $S$, then $A \cup B$, $A^c$, $A \cap B$ are also entitled to be events. Thus precisely we define an event as follows:

**Definition 1.7.** A subset $A$ of the sample space $S$ is said to be an event if it belongs to a collection $\mathcal{F}$ of subsets of $S$ satisfying the following three rules: (a) $S \in \mathcal{F}$; (b) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$; and (c) if $A_j \in \mathcal{F}$ for $j \geq 1$, then $\bigcup_{j=1}^{\infty} \in \mathcal{F}$. The collection $\mathcal{F}$ is called an event space or a $\sigma$-field. If $A$ is the outcome of an experiment, then we say that the event $A$ has occurred.

**Example 1.15.** Describe the sample space of rolling a die and interpret the event $\{1, 2\}$.

**Answer:** The sample space of this experiment is

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

The event $\{1, 2\}$ means getting either a 1 or a 2.

**Example 1.16.** First describe the sample space of rolling a pair of dice, then describe the event $A$ that the sum of numbers rolled is 7.

**Answer:** The sample space of this experiment is

$$S = \{(x, y) \,|\, x, y = 1, 2, 3, 4, 5, 6\}$$

and

$$A = \{(1, 6), (6, 1), (2, 5), (5, 2), (4, 3), (3, 4)\}.$$

**Definition 1.8.** Let $S$ be the sample space of a random experiment. A probability measure $P : \mathcal{F} \to [0, 1]$ is a set function which assigns real numbers to the various events of $S$ satisfying

(P1) $P(A) \geq 0$ for all event $A \in \mathcal{F}$,
(P2) $P(S) = 1$,

(P3) $P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$
if $A_1$, $A_2$, $A_3$, ..., $A_k$, ..... are mutually disjoint events of $S$.

Any set function with the above three properties is a probability measure for $S$. For a given sample space $S$, there may be more than one probability measure. The probability of an event $A$ is the value of the probability measure at $A$, that is

$$Prob(A) = P(A).$$

**Theorem 1.5.** If $\emptyset$ is a empty set (that is an impossible event), then

$$P(\emptyset) = 0.$$

**Proof:** Let $A_1 = S$ and $A_i = \emptyset$ for $i = 2, 3, ..., \infty$. Then

$$S = \bigcup_{i=1}^{\infty} A_i$$

where $A_i \cap A_j = \emptyset$ for $i \neq j$. By axiom 2 and axiom 3, we get

$$\begin{aligned} 1 &= P(S) \qquad \text{(by axiom 2)} \\ &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=1}^{\infty} P(A_i) \qquad \text{(by axiom 3)} \\ &= P(A_1) + \sum_{i=2}^{\infty} P(A_i) \\ &= P(S) + \sum_{i=2}^{\infty} P(\emptyset) \\ &= 1 + \sum_{i=2}^{\infty} P(\emptyset). \end{aligned}$$

Therefore

$$\sum_{i=2}^{\infty} P(\emptyset) = 0.$$

Since $P(\emptyset) \geq 0$ by axiom 1, we have

$$P(\emptyset) = 0$$

and the proof of the theorem is complete.

This theorem says that the probability of an impossible event is zero. Note that if the probability of an event is zero, that does not mean the event is empty (or impossible). There are random experiments in which there are infinitely many events each with probability 0. Similarly, if $A$ is an event with probability 1, then it does not mean $A$ is the sample space $S$. In fact there are random experiments in which one can find infinitely many events each with probability 1.

**Theorem 1.6.** Let $\{A_1, A_2, ..., A_n\}$ be a finite collection of $n$ events such that $A_i \cap E_j = \emptyset$ for $i \neq j$. Then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

**Proof:** Consider the collection $\{A_i'\}_{i=1}^{\infty}$ of the subsets of the sample space $S$ such that

$$A_1' = A_1, \ A_2' = A_2, ..., A_n' = A_n$$

and

$$A_{n+1}' = A_{n+2}' = A_{n+3}' = \cdots = \emptyset.$$

Hence

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= P\left(\bigcup_{i=1}^{\infty} A_i'\right) \\ &= \sum_{i=1}^{\infty} P\left(A_i'\right) \\ &= \sum_{i=1}^{n} P\left(A_i'\right) + \sum_{i=n+1}^{\infty} P\left(A_i'\right) \\ &= \sum_{i=1}^{n} P\left(A_i\right) + \sum_{i=n+1}^{\infty} P\left(\emptyset\right) \\ &= \sum_{i=1}^{n} P\left(A_i\right) + 0 \\ &= \sum_{i=1}^{n} P\left(A_i\right) \end{aligned}$$

and the proof of the theorem is now complete.

When $n = 2$, the above theorem yields $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ where $A_1$ and $A_2$ are disjoint (or mutually exclusive) events.

In the following theorem, we give a method for computing probability of an event $A$ by knowing the probabilities of the elementary events of the sample space $S$.

**Theorem 1.7.** If $A$ is an event of a discrete sample space $S$, then the probability of $A$ is equal to the sum of the probabilities of its elementary events.

**Proof:** Any set $A$ in $S$ can be written as the union of its singleton sets. Let $\{O_i\}_{i=1}^{\infty}$ be the collection of all the singleton sets (or the elementary events) of $A$. Then

$$A = \bigcup_{i=1}^{\infty} O_i.$$

By axiom (P3), we get

$$\begin{aligned} P(A) &= P\left(\bigcup_{i=1}^{\infty} O_i\right) \\ &= \sum_{i=1}^{\infty} P(O_i). \end{aligned}$$

**Example 1.17.** If a fair coin is tossed twice, what is the probability of getting at least one head?

**Answer:** The sample space of this experiment is

$$S = \{HH,\, HT,\, TH,\, TT\}.$$

The event $A$ is given by

$$\begin{aligned} A &= \{\,\text{at least one head}\,\} \\ &= \{HH,\, HT,\, TH\}. \end{aligned}$$

By Theorem 1.7, the probability of $A$ is the sum of the probabilities of its elementary events. Thus, we get

$$\begin{aligned} P(A) &= P(HH) + P(HT) + P(TH) \\ &= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} \\ &= \frac{3}{4}. \end{aligned}$$

**Remark 1.2.** Notice that here we are not computing the probability of the elementary events by taking the number of points in the elementary event and dividing by the total number of points in the sample space. We are using the randomness to obtain the probability of the elementary events. That is, we are assuming that each outcome is equally likely. This is why the randomness is an integral part of probability theory.

**Corollary 1.1.** If $S$ is a finite sample space with $n$ sample elements and $A$ is an event in $S$ with $m$ elements, then the probability of $A$ is given by

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

**Proof:** By the previous theorem, we get

$$\begin{aligned} P(A) &= P\left(\bigcup_{i=1}^{m} O_i\right) \\ &= \sum_{i=1}^{m} P(O_i) \\ &= \sum_{i=1}^{m} \frac{1}{n} \\ &= \frac{m}{n}. \end{aligned}$$

The proof is now complete.

**Example 1.18.** A die is loaded in such a way that the probability of the face with $j$ dots turning up is proportional to $j$ for $j = 1, 2, ..., 6$. What is the probability, in one roll of the die, that an odd number of dots will turn up?

**Answer:**

$$\begin{aligned} P(\{j\}) &\propto j \\ &= k\, j \end{aligned}$$

where $k$ is a constant of proportionality. Next, we determine this constant $k$ by using the axiom (P2). Using Theorem 1.5, we get

$$\begin{aligned} P(S) &= P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\}) \\ &= k + 2k + 3k + 4k + 5k + 6k \\ &= (1 + 2 + 3 + 4 + 5 + 6)\, k \\ &= \frac{(6)(6+1)}{2}\, k \\ &= 21k. \end{aligned}$$

Using (P2), we get

$$21k = 1.$$

Thus $k = \frac{1}{21}$. Hence, we have

$$P(\{j\}) = \frac{j}{21}.$$

Now, we want to find the probability of the odd number of dots turning up.

$$\begin{aligned} P(\text{odd numbered dot will turn up}) &= P(\{1\}) + P(\{3\}) + P(\{5\}) \\ &= \frac{1}{21} + \frac{3}{21} + \frac{5}{21} \\ &= \frac{9}{21}. \end{aligned}$$

**Remark 1.3.** Recall that the sum of the first $n$ integers is equal to $\frac{n}{2}(n+1)$. That is,

$$1 + 2 + 3 + \cdots\cdots + (n-2) + (n-1) + n = \frac{n(n+1)}{2}.$$

This formula was first proven by Gauss (1777-1855) when he was a young school boy.

**Remark 1.4.** Gauss proved that the sum of the first $n$ positive integers is $n\frac{(n+1)}{2}$ when he was a school boy. Kolmogorov, the father of modern probability theory, proved that the sum of the first $n$ odd positive integers is $n^2$, when he was five years old.

### 1.4. Some Properties of the Probability Measure

Next, we present some theorems that will illustrate the various intuitive properties of a probability measure.

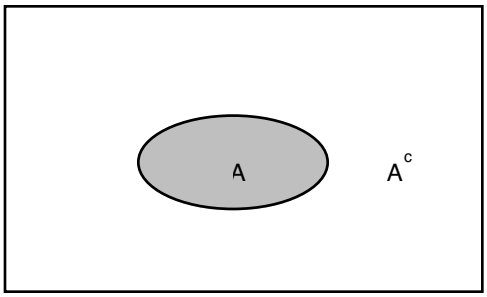
**Theorem 1.8.** If $A$ be any event of the sample space $S$, then

$$P(A^c) = 1 - P(A)$$

where $A^c$ denotes the complement of $A$ with respect to $S$.

**Proof:** Let $A$ be any subset of $S$. Then $S = A \cup A^c$. Further $A$ and $A^c$ are mutually disjoint. Thus, using (P3), we get

$$\begin{aligned} 1 = P(S) &= P(A \cup A^c) \\ &= P(A) + P(A^c). \end{aligned}$$



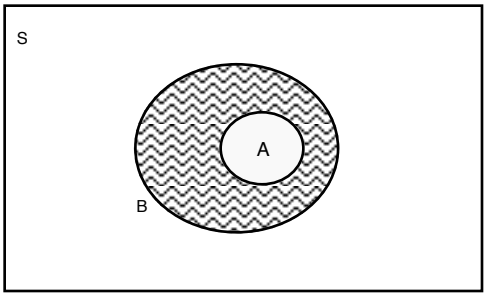

Hence, we see that

$$P(A^c) = 1 - P(A).$$

This completes the proof.

**Theorem 1.9.** If $A \subseteq B \subseteq S$, then

$$P(A) \leq P(B).$$




**Proof:** Note that $B = A \cup (B \setminus A)$ where $B \setminus A$ denotes all the elements $x$ that are in $B$ but not in $A$. Further, $A \cap (B \setminus A) = \emptyset$. Hence by (P3), we get

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

By axiom (P1), we know that $P(B \setminus A) \geq 0$. Thus, from the above, we get

$$P(B) \geq P(A)$$

and the proof is complete.

**Theorem 1.10.** If $A$ is any event in $S$, then

$$0 \leq P(A) \leq 1.$$

**Proof:** Follows from axioms (P1) and (P2) and Theorem 1.8.

**Theorem 1.10.** If $A$ and $B$ are any two events, then

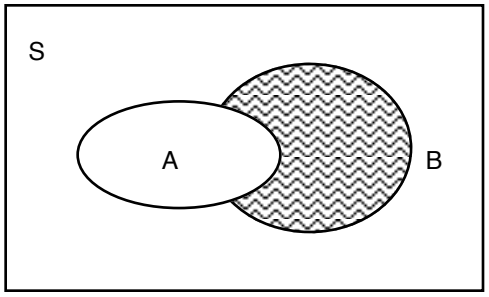
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

**Proof:** It is easy to see that

$$A \cup B = A \cup (A^c \cap B)$$

and

$$A \cap (A^c \cap B) = \emptyset.$$

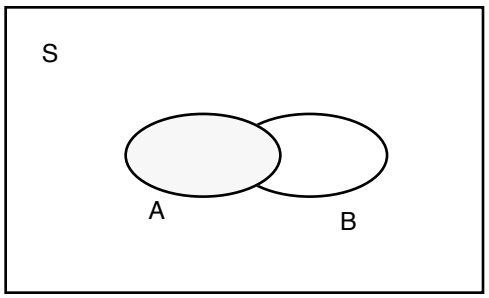



Hence by (P3), we get

$$P(A \cup B) = P(A) + P(A^c \cap B) \tag{1.1}$$

But the set $B$ can also be written as

$$B = (A \cap B) \cup (A^c \cap B)$$



Therefore, by (P3), we get

$$P(B) = P(A \cap B) + P(A^c \cap B). \tag{1.2}$$

Eliminating $P(A^c \cap B)$ from (1.1) and (1.2), we get

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

and the proof of the theorem is now complete.

This above theorem tells us how to calculate the probability that at least one of $A$ and $B$ occurs.

**Example 1.19.** If $P(A) = 0.25$ and $P(B) = 0.8$, then show that $0.05 \leq P(A \cap B) \leq 0.25$.

**Answer:** Since $A \cap B \subseteq A$ and $A \cap B \subseteq B$, by Theorem 1.8, we get

$$P(A \cap B) \leq P(A) \qquad \text{and also} \quad P(A \cap B) \leq P(B).$$

Hence

$$P(A \cap B) \leq \min\{P(A), P(B)\}.$$

This shows that

$$P(A \cap B) \leq 0.25. \tag{1.3}$$

Since $A \cup B \subseteq S$, by Theorem 1.8, we get

$$P(A \cup B) \leq P(S)$$

That is, by Theorem 1.10

$$P(A) + P(B) - P(A \cap B) \leq P(S).$$

Hence, we obtain

$$0.8 + 0.25 - P(A \cap B) \leq 1$$

and this yields

$$0.8 + 0.25 - 1 \leq P(A \cap B).$$

From this, we get

$$0.05 \leq P(A \cap B). \tag{1.4}$$

From (1.3) and (1.4), we get

$$0.05 \leq P(A \cap B) \leq 0.25.$$

**Example 1.20.** Let $A$ and $B$ be events in a sample space $S$ such that $P(A) = \frac{1}{2} = P(B)$ and $P(A^c \cap B^c) = \frac{1}{3}$. Find $P(A \cup B^c)$.

**Answer:** Notice that

$$A \cup B^c = A \cup (A^c \cap B^c).$$

Hence,

$$\begin{aligned} P(A \cup B^c) &= P(A) + P(A^c \cap B^c) \\ &= \frac{1}{2} + \frac{1}{3} \\ &= \frac{5}{6}. \end{aligned}$$

**Theorem 1.11.** If $A_1$ and $A_2$ are two events such that $A_1 \subseteq A_2$, then

$$P(A_2 \setminus A_1) = P(A_2) - P(A_1).$$

**Proof:** The event $A_2$ can be written as

$$A_2 = A_1 \bigcup (A_2 \setminus A_1)$$

where the sets $A_1$ and $A_2 \setminus A_1$ are disjoint. Hence

$$P(A_2) = P(A_1) + P(A_2 \setminus A_1)$$

which is

$$P(A_2 \setminus A_1) = P(A_2) - P(A_1)$$

and the proof of the theorem is now complete.

From calculus we know that a real function $f : \mathbb{R} \to \mathbb{R}$ (the set of real numbers) is continuous on $\mathbb{R}$ if and only if, for every convergent sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbb{R}$,

$$\lim_{n \to \infty} f(x_n) = f\left(\lim_{n \to \infty} x_n\right).$$

**Theorem 1.12.** If $A_1, A_2, ..., A_n, ...$ is a sequence of events in sample space $S$ such that $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n).$$

Similarly, if $B_1, B_2, ..., B_n, ...$ is a sequence of events in sample space $S$ such that $B_1 \supseteq B_2 \supseteq \cdots \supseteq B_n \supseteq \cdots$, then

$$P\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} P(B_n).$$

**Proof:** Given an increasing sequence of events

$$A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$$

we define a disjoint collection of events as follows:

$$\begin{aligned} E_1 &= A_1 \\ E_n &= A_n \setminus A_{n-1} \qquad \forall n \geq 2. \end{aligned}$$

Then $\{E_n\}_{n=1}^{\infty}$ is a disjoint collection of events such that

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} E_n.$$

Further

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} A_n\right) &= P\left(\bigcup_{n=1}^{\infty} E_n\right) \\ &= \sum_{n=1}^{\infty} P\left(E_n\right) \\ &= \lim_{m\to\infty} \sum_{n=1}^{m} P\left(E_n\right) \\ &= \lim_{m\to\infty} \left[P(A_1) + \sum_{n=2}^{m} \left[P\left(A_n\right) - P\left(A_{n-1}\right)\right]\right] \\ &= \lim_{m\to\infty} P(A_m) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

The second part of the theorem can be proved similarly.

Note that

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$$

and

$$\lim_{n\to\infty} B_n = \bigcap_{n=1}^{\infty} B_n.$$

Hence the results above theorem can be written as

$$P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

and

$$P\left(\lim_{n\to\infty} B_n\right) = \lim_{n\to\infty} P(B_n)$$

and the Theorem 1.12 is called the continuity theorem for the probability measure.

### 1.5. Review Exercises

**1.** If we randomly pick two television sets in succession from a shipment of 240 television sets of which 15 are defective, what is the probability that they will both be defective?

**2.** A poll of 500 people determines that 382 like ice cream and 362 like cake. How many people like both if each of them likes at least one of the two? (Hint: Use $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ).

**3.** The Mathematics Department of the University of Louisville consists of 8 professors, 6 associate professors, 13 assistant professors. In how many of all possible samples of size 4, chosen without replacement, will every type of professor be represented?

**4.** A pair of dice consisting of a *six-sided* die and a *four-sided* die is rolled and the sum is determined. Let A be the event that a sum of 5 is rolled and let B be the event that a sum of 5 or a sum of 9 is rolled. Find (a) $P(A)$, (b) $P(B)$, and (c) $P(A \cap B)$.

**5.** A faculty leader was meeting two students in Paris, one arriving by train from Amsterdam and the other arriving from Brussels at approximately the same time. Let $A$ and $B$ be the events that the trains are on time, respectively. If $P(A) = 0.93$, $P(B) = 0.89$ and $P(A \cap B) = 0.87$, then find the probability that at least one train is on time.

**6.** Bill, George, and Ross, in order, roll a die. The first one to roll an even number wins and the game is ended. What is the probability that Bill will win the game?

**7.** Let $A$ and $B$ be events such that $P(A) = \frac{1}{2} = P(B)$ and $P(A^c \cap B^c) = \frac{1}{3}$. Find the probability of the event $A^c \cup B^c$.

**8.** Suppose a box contains 4 blue, 5 white, 6 red and 7 green balls. In how many of all possible samples of size 5, chosen without replacement, will every color be represented?

**9.** Using the Binomial Theorem, show that $\sum_{k=0}^{n} k\binom{n}{k} = n\,2^{n-1}$.

**10.** A function consists of a domain $A$, a co-domain $B$ and a rule $f$. The rule $f$ assigns to each number in the domain $A$ one and only one letter in the co-domain $B$. If $A = \{1, 2, 3\}$ and $B = \{x, y, z, w\}$, then find all the distinct functions that can be formed from the set $A$ into the set $B$.

**11.** Let $S$ be a countable sample space. Let $\{O_i\}_{i=1}^{\infty}$ be the collection of all the elementary events in $S$. What should be the value of the constant $c$ such that $P(O_i) = c\left(\frac{1}{3}\right)^i$ will be a probability measure in $S$?

**12.** A box contains five green balls, three black balls, and seven red balls. Two balls are selected at random without replacement from the box. What is the probability that both balls are the same color?

**13.** Find the sample space of the random experiment which consists of tossing a coin until the first head is obtained. Is this sample space discrete?

**14.** Find the sample space of the random experiment which consists of tossing a coin infinitely many times. Is this sample space discrete?

**15.** Five fair dice are thrown. What is the probability that a full house is thrown (that is, where two dice show one number and other three dice show a second number)?

**16.** If a fair coin is tossed repeatedly, what is the probability that the third head occurs on the $n^{\text{th}}$ toss?

**17.** In a particular softball league each team consists of 5 women and 5 men. In determining a batting order for 10 players, a woman must bat first, and successive batters must be of opposite sex. How many different batting orders are possible for a team?

**18.** An urn contains 3 red balls, 2 green balls and 1 yellow ball. Three balls are selected at random and without replacement from the urn. What is the probability that at least 1 color is not drawn?

**19.** A box contains four \$10 bills, six \$5 bills and two \$1 bills. Two bills are taken at random from the box without replacement. What is the probability that both bills will be of the same denomination?

**20.** An urn contains $n$ white counters numbered 1 through $n$, $n$ black counters numbered 1 through $n$, and $n$ red counter numbered 1 through $n$. If two counters are to be drawn at random without replacement, what is the probability that both counters will be of the same color or bear the same number?

**21.** Two people take turns rolling a fair die. Person $X$ rolls first, then person $Y$, then $X$, and so on. The winner is the first to roll a 6. What is the probability that person $X$ wins?

**22.** Mr. Flowers plants 10 rose bushes in a row. Eight of the bushes are white and two are red, and he plants them in random order. What is the probability that he will consecutively plant seven or more white bushes?

**23.** Using mathematical induction, show that

$$\frac{d^n}{dx^n}\left[f(x)\cdot g(x)\right]=\sum_{k=0}^{n}\binom{n}{k}\frac{d^k}{dx^k}\left[f(x)\right]\cdot\frac{d^{n-k}}{dx^{n-k}}\left[g(x)\right].$$

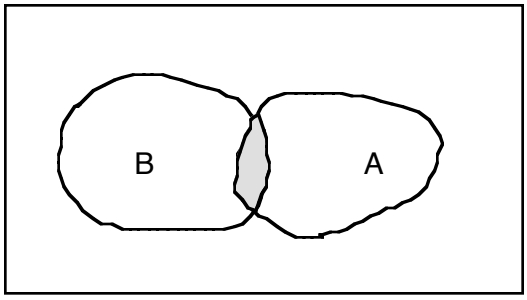
# Chapter 2

# CONDITIONAL PROBABILITIES AND BAYES' THEOREM

## 2.1. Conditional Probabilities

First, we give a heuristic argument for the definition of conditional probability, and then based on our heuristic argument, we define the conditional probability.

Consider a random experiment whose sample space is $S$. Let $B \subset S$. In many situations, we are only concerned with those outcomes that are elements of $B$. This means that we consider $B$ to be our new sample space.




For the time being, suppose $S$ is a nonempty finite sample space and $B$ is a nonempty subset of $S$. Given this new discrete sample space $B$, how do we define the probability of an event $A$? Intuitively, one should define the probability of $A$ with respect to the new sample space $B$ as (see the figure above)

$$P(A \text{ given } B) = \frac{\text{the number of elements in A} \cap \text{B}}{\text{the number of elements in B}}.$$

We denote the conditional probability of $A$ given the new sample space $B$ as $P(A/B)$. Hence with this notation, we say that

$$\begin{aligned} P(A/B) &= \frac{N(A \cap B)}{N(B)} \\ &= \frac{P(A \cap B)}{P(B)}, \end{aligned}$$

since $N(S) \neq 0$. Here $N(S)$ denotes the number of elements in $S$.

Thus, if the sample space is finite, then the above definition of the probability of an event $A$ given that the event $B$ has occurred makes sense intuitively. Now we define the conditional probability for any sample space (discrete or continuous) as follows.

**Definition 2.1.** Let $S$ be a sample space associated with a random experiment. The conditional probability of an event $A$, given that event $B$ has occurred, is defined by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

provided $P(B) > 0$.

This conditional probability measure $P(A/B)$ satisfies all three axioms of a probability measure. That is,

(CP1) $P(A/B) \geq 0$ for all event $A$

(CP2) $P(B/B) = 1$

(CP3) If $A_1, A_2, ..., A_k, ...$ are mutually exclusive events, then

$$P(\bigcup_{k=1}^{\infty} A_k/B) = \sum_{k=1}^{\infty} P(A_k/B).$$

Thus, it is a probability measure with respect to the new sample space $B$.

**Example 2.1.** A drawer contains 4 black, 6 brown, and 8 olive socks. Two socks are selected at random from the drawer. (a) What is the probability that both socks are of the same color? (b) What is the probability that both socks are olive if it is known that they are of the same color?

**Answer:** The sample space of this experiment consists of

$$S = \{(x, y) \,|\, x, y \in Bl,\ Ol,\ Br\}.$$

The cardinality of $S$ is

$$N(S) = \binom{18}{2} = 153.$$

Let $A$ be the event that two socks selected at random are of the same color. Then the cardinality of $A$ is given by

$$\begin{aligned} N(A) &= \binom{4}{2} + \binom{6}{2} + \binom{8}{2} \\ &= 6 + 15 + 28 \\ &= 49. \end{aligned}$$

Therefore, the probability of $A$ is given by

$$P(A) = \frac{49}{\binom{18}{2}} = \frac{49}{153}.$$

Let $B$ be the event that two socks selected at random are olive. Then the cardinality of $B$ is given by

$$N(B) = \binom{8}{2}$$

and hence

$$P(B) = \frac{\binom{8}{2}}{\binom{18}{2}} = \frac{28}{153}.$$

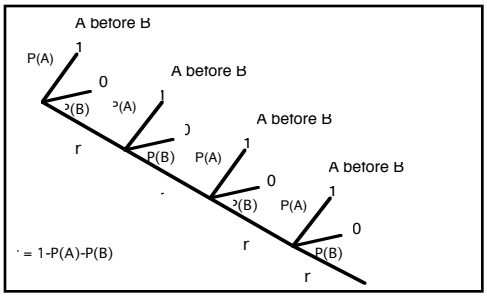
Notice that $B \subset A$. Hence,

$$\begin{aligned} P(B/A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(B)}{P(A)} \\ &= \left(\frac{28}{153}\right)\left(\frac{153}{49}\right) \\ &= \frac{28}{49} = \frac{4}{7}. \end{aligned}$$

Let $A$ and $B$ be two mutually disjoint events in a sample space $S$. We want to find a formula for computing the probability that the event $A$ occurs before the event $B$ in a sequence trials. Let $P(A)$ and $P(B)$ be the probabilities that $A$ and $B$ occur, respectively. Then the probability that neither $A$ nor $B$ occurs is $1 - P(A) - P(B)$. Let us denote this probability by $r$, that is $r = 1 - P(A) - P(B)$.

In the first trial, either $A$ occurs, or $B$ occurs, or neither $A$ nor $B$ occurs. In the first trial if $A$ occurs, then the probability of $A$ occurs before $B$ is 1.

If $B$ occurs in the first trial, then the probability of $A$ occurs before $B$ is 0. If neither $A$ nor $B$ occurs in the first trial, we look at the outcomes of the second trial. In the second trial if $A$ occurs, then the probability of $A$ occurs before $B$ is 1. If $B$ occurs in the second trial, then the probability of $A$ occurs before $B$ is 0. If neither $A$ nor $B$ occurs in the second trial, we look at the outcomes of the third trial, and so on. This argument can be summarized in the following diagram.




Hence the probability that the event $A$ comes before the event $B$ is given by

$$\begin{aligned} P(\text{A before B}) &= P(A) + r\,P(A) + r^2\,P(A) + r^3\,P(A) + \cdots + r^n\,P(A) + \cdots \\ &= P(A)\ [1 + r + r^2 + \cdots + r^n + \cdots] \\ &= P(A)\ \frac{1}{1-r} \\ &= P(A)\ \frac{1}{1-[1-P(A)-P(B)]} \\ &= \frac{P(A)}{P(A)+P(B)}. \end{aligned}$$

The event $A$ before $B$ can also be interpreted as a conditional event. In this interpretation the event $A$ before $B$ means the occurrence of the event $A$ given that $A \cup B$ has already occurred. Thus we again have

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

**Example 2.2.** A pair of four-sided dice is rolled and the sum is determined. What is the probability that a sum of 3 is rolled before a sum of 5 is rolled in a sequence of rolls of the dice?

**Answer:** The sample space of this random experiment is

$$S = \begin{matrix} \{(1,1) & (1,2) & (1,3) & (1,4) \\ (2,1) & (2,2) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,3) & (3,4) \\ (4,1) & (4,2) & (4,3) & (4,4)\}. \end{matrix}$$

Let $A$ denote the event of getting a sum of 3 and $B$ denote the event of getting a sum of 5. The probability that a sum of 3 is rolled before a sum of 5 is rolled can be thought of as the conditional probability of a sum of 3, given that a sum of 3 or 5 has occurred. That is, $P(A/A \cup B)$. Hence

$$\begin{aligned} P(A/A \cup B) &= \frac{P(A \cap (A \cup B))}{P(A \cup B)} \\ &= \frac{P(A)}{P(A) + P(B)} \\ &= \frac{N(A)}{N(A) + N(B)} \\ &= \frac{2}{2+4} \\ &= \frac{1}{3}. \end{aligned}$$

**Example 2.3.** If we randomly pick two television sets in succession from a shipment of 240 television sets of which 15 are defective, what is the probability that they will be both defective?

**Answer:** Let $A$ denote the event that the first television picked was defective. Let $B$ denote the event that the second television picked was defective. Then $A \cap B$ will denote the event that both televisions picked were defective. Using the conditional probability, we can calculate

$$\begin{aligned} P(A \cap B) &= P(A)\, P(B/A) \\ &= \left(\frac{15}{240}\right) \left(\frac{14}{239}\right) \\ &= \frac{7}{1912}. \end{aligned}$$

In Example 2.3, we assume that we are sampling without replacement.

**Definition 2.2.** If an object is selected and then replaced before the next object is selected, this is known as sampling with replacement. Otherwise, it is called sampling without replacement.

Rolling a die is equivalent to sampling with replacement, whereas dealing a deck of cards to players is sampling without replacement.

**Example 2.4.** A box of fuses contains 20 fuses, of which 5 are defective. If 3 of the fuses are selected at random and removed from the box in succession without replacement, what is the probability that all three fuses are defective?

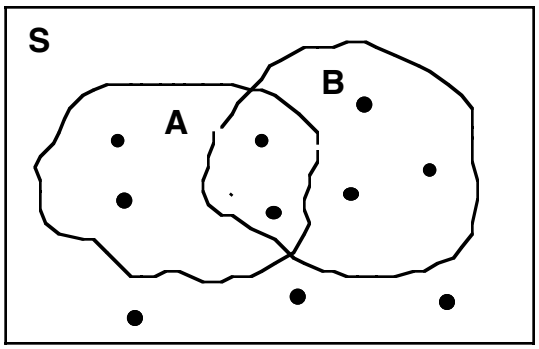
**Answer:** Let $A$ be the event that the first fuse selected is defective. Let $B$ be the event that the second fuse selected is defective. Let $C$ be the event that the third fuse selected is defective. The probability that all three fuses selected are defective is $P(A \cap B \cap C)$. Hence

$$\begin{aligned} P(A \cap B \cap C) &= P(A)\, P(B/A)\, P(C/A \cap B) \\ &= \left(\frac{5}{20}\right)\left(\frac{4}{19}\right)\left(\frac{3}{18}\right) \\ &= \frac{1}{114}. \end{aligned}$$

**Definition 2.3.** Two events $A$ and $B$ of a sample space $S$ are called independent if and only if

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

**Example 2.5.** The following diagram shows two events $A$ and $B$ in the sample space $S$. Are the events $A$ and $B$ independent?




**Answer:** There are 10 black dots in $S$ and event $A$ contains 4 of these dots. So the probability of $A$, is $P(A) = \frac{4}{10}$. Similarly, event $B$ contains 5 black dots. Hence $P(B) = \frac{5}{10}$. The conditional probability of $A$ given $B$ is

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{5}.$$

This shows that $P(A/B) = P(A)$. Hence $A$ and $B$ are independent.

**Theorem 2.1.** Let $A, B \subseteq S$. If $A$ and $B$ are independent and $P(B) > 0$, then

$$P(A/B) = P(A).$$

**Proof:**

$$\begin{aligned} P(A/B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)\,P(B)}{P(B)} \\ &= P(A). \end{aligned}$$

**Theorem 2.2.** If $A$ and $B$ are independent events. Then $A^c$ and $B$ are independent. Similarly $A$ and $B^c$ are independent.

**Proof:** We know that $A$ and $B$ are independent, that is

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

and we want to show that $A^c$ and $B$ are independent, that is

$$P(A^c \cap B) = P(A^c)\,P(B).$$

Since

$$\begin{aligned} P(A^c \cap B) &= P(A^c/B)\,P(B) \\ &= [1 - P(A/B)]\,P(B) \\ &= P(B) - P(A/B)P(B) \\ &= P(B) - P(A \cap B) \\ &= P(B) - P(A)\,P(B) \\ &= P(B)\,[1 - P(A)] \\ &= P(B)P(A^c), \end{aligned}$$

the events $A^c$ and $B$ are independent. Similarly, it can be shown that $A$ and $B^c$ are independent and the proof is now complete.

**Remark 2.1.** The concept of independence is fundamental. In fact, it is this concept that justifies the mathematical development of probability as a separate discipline from measure theory. Mark Kac said, "independence of events is not a purely mathematical concept." It can, however, be made plausible

that it should be interpreted by the rule of multiplication of probabilities and this leads to the mathematical definition of independence.

**Example 2.6.** Flip a coin and then independently cast a die. What is the probability of observing heads on the coin and a 2 or 3 on the die?

**Answer:** Let $A$ denote the event of observing a head on the coin and let $B$ be the event of observing a 2 or 3 on the die. Then

$$\begin{aligned} P(A \cap B) &= P(A)\, P(B) \\ &= \left(\frac{1}{2}\right)\, \left(\frac{2}{6}\right) \\ &= \frac{1}{6}. \end{aligned}$$

**Example 2.7.** An urn contains 3 red, 2 white and 4 yellow balls. An ordered sample of size 3 is drawn from the urn. If the balls are drawn with replacement so that one outcome does not change the probabilities of others, then what is the probability of drawing a sample that has balls of each color? Also, find the probability of drawing a sample that has two yellow balls and a red ball or a red ball and two white balls?

**Answer:**

$$P(RWY) = \left(\frac{3}{9}\right)\, \left(\frac{2}{9}\right)\, \left(\frac{4}{9}\right) = \frac{8}{243}$$

and

$$P(YYR\, or\, RWW) = \left(\frac{4}{9}\right)\, \left(\frac{4}{9}\right)\, \left(\frac{3}{9}\right) + \left(\frac{3}{9}\right)\, \left(\frac{2}{9}\right)\, \left(\frac{2}{9}\right) = \frac{20}{243}.$$

If the balls are drawn without replacement, then

$$P(RWY) = \left(\frac{3}{9}\right)\, \left(\frac{2}{8}\right)\, \left(\frac{4}{7}\right) = \frac{1}{21}.$$

$$P(YYR\, or\, RWW) = \left(\frac{4}{9}\right)\, \left(\frac{3}{8}\right)\, \left(\frac{3}{7}\right) + \left(\frac{3}{9}\right)\, \left(\frac{2}{8}\right)\, \left(\frac{1}{7}\right) = \frac{7}{84}.$$

There is a tendency to equate the concepts "mutually exclusive" and "independence". This is a fallacy. Two events $A$ and $B$ are *mutually exclusive* if $A \cap B = \emptyset$ and they are called *possible* if $P(A) \neq 0 \neq P(B)$.

**Theorem 2.2.** Two possible mutually exclusive events are always dependent (that is not independent).

**Proof:** Suppose not. Then

$$\begin{aligned} P(A \cap B) &= P(A)\, P(B) \\ P(\emptyset) &= P(A)\, P(B) \\ 0 &= P(A)\, P(B). \end{aligned}$$

Hence, we get either $P(A) = 0$ or $P(B) = 0$. This is a contradiction to the fact that $A$ and $B$ are possible events. This completes the proof.

**Theorem 2.3.** Two possible independent events are not mutually exclusive.

**Proof:** Let $A$ and $B$ be two independent events and suppose $A$ and $B$ are mutually exclusive. Then

$$\begin{aligned} P(A)\, P(B) &= P(A \cap B) \\ &= P(\emptyset) \\ &= 0. \end{aligned}$$

Therefore, we get either $P(A) = 0$ or $P(B) = 0$. This is a contradiction to the fact that $A$ and $B$ are possible events.

The possible events $A$ and $B$ exclusive implies $A$ and $B$ are not independent; and $A$ and $B$ independent implies $A$ and $B$ are not exclusive.

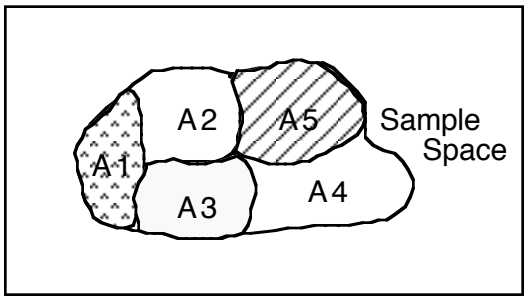
### 2.2. Bayes' Theorem

There are many situations where the ultimate outcome of an experiment depends on what happens in various intermediate stages. This issue is resolved by the Bayes' Theorem.

**Definition 2.4.** Let $S$ be a set and let $\mathcal{P} = \{A_i\}_{i=1}^{m}$ be a collection of subsets of $S$. The collection $\mathcal{P}$ is called a partition of $S$ if

$$(a) \quad S = \bigcup_{i=1}^{m} A_i$$

$$(b) \quad A_i \cap A_j = \emptyset \quad \text{for } i \neq j.$$



**Theorem 2.4.** If the events $\{B_i\}_{i=1}^m$ constitute a partition of the sample space $S$ and $P(B_i) \neq 0$ for $i = 1, 2, ..., m$, then for any event $A$ in $S$

$$P(A) = \sum_{i=1}^{m} P(B_i)\, P(A/B_i).$$

**Proof:** Let $S$ be a sample space and $A$ be an event in $S$. Let $\{B_i\}_{i=1}^m$ be any partition of $S$. Then

$$A = \bigcup_{i=1}^{m} (A \cap B_i)\,.$$

Thus

$$\begin{aligned} P(A) &= \sum_{i=1}^{m} P\,(A \cap B_i) \\ &= \sum_{i=1}^{m} P(B_i)\, P\,(A/B_i)\,. \end{aligned}$$

**Theorem 2.5.** If the events $\{B_i\}_{i=1}^m$ constitute a partition of the sample space $S$ and $P(B_i) \neq 0$ for $i = 1, 2, ..., m$, then for any event $A$ in $S$ such that $P(A) \neq 0$

$$P(B_k/A) = \frac{P(B_k)\, P(A/B_k)}{\sum_{i=1}^{m} P(B_i)\, P(A/B_i)} \quad k = 1, 2, ..., m.$$

**Proof:** Using the definition of conditional probability, we get

$$P(B_k/A) = \frac{P(A \cap B_k)}{P(A)}.$$

Using Theorem 1, we get

$$P(B_k/A) = \frac{P(A \cap B_k)}{\sum_{i=1}^{m} P(B_i)\, P(A/B_i)}.$$

This completes the proof.

This Theorem is called Bayes Theorem. The probability $P(B_k)$ is called prior probability. The probability $P(B_k/A)$ is called posterior probability.

**Example 2.8.** Two boxes containing marbles are placed on a table. The boxes are labeled $B_1$ and $B_2$. Box $B_1$ contains 7 green marbles and 4 white

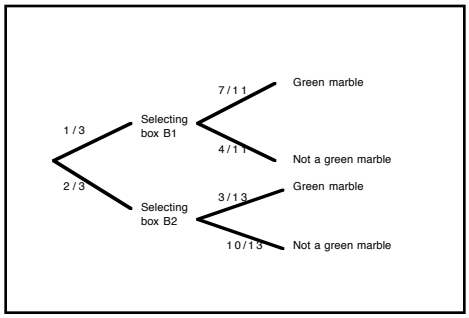
marbles. Box $B_2$ contains 3 green marbles and 10 yellow marbles. The boxes are arranged so that the probability of selecting box $B_1$ is $\frac{1}{3}$ and the probability of selecting box $B_2$ is $\frac{2}{3}$. Kathy is blindfolded and asked to select a marble. She will win a color TV if she selects a green marble. (a) What is the probability that Kathy will win the TV (that is, she will select a green marble)? (b) If Kathy wins the color TV, what is the probability that the green marble was selected from the first box?

**Answer:** Let $A$ be the event of drawing a green marble. The prior probabilities are $P(B_1) = \frac{1}{3}$ and $P(B_2) = \frac{2}{3}$.

(a) The probability that Kathy will win the TV is

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) \\ &= P(A/B_1)\, P(B_1) + P(A/B_2)\, P(B_2) \\ &= \left(\frac{7}{11}\right)\left(\frac{1}{3}\right) + \left(\frac{3}{13}\right)\left(\frac{2}{3}\right) \\ &= \frac{7}{33} + \frac{2}{13} \\ &= \frac{91}{429} + \frac{66}{429} \\ &= \frac{157}{429}. \end{aligned}$$

(b) Given that Kathy won the TV, the probability that the green marble was selected from $B_1$ is



$$P(B_1/A) = \frac{P(A/B_1)\,P(B_1)}{P(A/B_1)\,P(B_1) + P(A/B_2)\,P(B_2)}$$

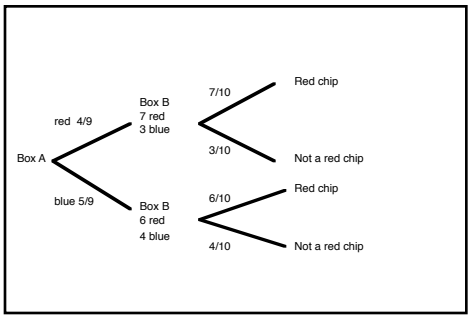
$$= \frac{\left(\frac{7}{11}\right)\,\left(\frac{1}{3}\right)}{\left(\frac{7}{11}\right)\,\left(\frac{1}{3}\right) + \left(\frac{3}{13}\right)\,\left(\frac{2}{3}\right)}$$

$$= \frac{91}{157}.$$

Note that $P(A/B_1)$ is the probability of selecting a green marble from $B_1$ whereas $P(B_1/A)$ is the probability that the green marble was selected from box $B_1$.

**Example 2.9.** Suppose box $A$ contains 4 red and 5 blue chips and box $B$ contains 6 red and 3 blue chips. A chip is chosen at random from the box $A$ and placed in box $B$. Finally, a chip is chosen at random from among those now in box $B$. What is the probability a blue chip was transferred from box $A$ to box $B$ given that the chip chosen from box $B$ is red?

**Answer:** Let $E$ represent the event of moving a blue chip from box $A$ to box $B$. We want to find the probability of a blue chip which was moved from box $A$ to box $B$ given that the chip chosen from $B$ was red. The probability of choosing a red chip from box $A$ is $P(R) = \frac{4}{9}$ and the probability of choosing a blue chip from box $A$ is $P(B) = \frac{5}{9}$. If a red chip was moved from box $A$ to box $B$, then box $B$ has 7 red chips and 3 blue chips. Thus the probability of choosing a red chip from box $B$ is $\frac{7}{10}$. Similarly, if a blue chip was moved from box $A$ to box $B$, then the probability of choosing a red chip from box $B$ is $\frac{6}{10}$.



Hence, the probability that a blue chip was transferred from box $A$ to box $B$ given that the chip chosen from box $B$ is red is given by

$$P(E/R) = \frac{P(R/E)\,P(E)}{P(R)}$$

$$= \frac{\left(\frac{6}{10}\right)\,\left(\frac{5}{9}\right)}{\left(\frac{7}{10}\right)\,\left(\frac{4}{9}\right) + \left(\frac{6}{10}\right)\,\left(\frac{5}{9}\right)}$$

$$= \frac{15}{29}.$$

**Example 2.10.** Sixty percent of new drivers have had driver education. During their first year, new drivers without driver education have probability 0.08 of having an accident, but new drivers with driver education have only a 0.05 probability of an accident. What is the probability a new driver has had driver education, given that the driver has had no accident the first year?

**Answer:** Let $A$ represent the new driver who has had driver education and $B$ represent the new driver who has had an accident in his first year. Let $A^c$ and $B^c$ be the complement of $A$ and $B$, respectively. We want to find the probability that a new driver has had driver education, given that the driver has had no accidents in the first year, that is $P(A/B^c)$.

$$P(A/B^c) = \frac{P(A \cap B^c)}{P(B^c)}$$

$$= \frac{P(B^c/A)\,P(A)}{P(B^c/A)\,P(A) + P(B^c/A^c)\,P(A^c)}$$

$$= \frac{[1 - P(B/A)]\,P(A)}{[1 - P(B/A)]\,P(A) + [1 - P(B/A^c)]\,[1 - P(A)]}$$

$$= \frac{\left(\frac{60}{100}\right)\,\left(\frac{95}{100}\right)}{\left(\frac{40}{100}\right)\,\left(\frac{92}{100}\right) + \left(\frac{60}{100}\right)\,\left(\frac{95}{100}\right)}$$

$$= 0.6077.$$

**Example 2.11.** One-half percent of the population has AIDS. There is a test to detect AIDS. A positive test result is supposed to mean that you

have AIDS but the test is not perfect. For people with AIDS, the test misses the diagnosis 2% of the times. And for the people without AIDS, the test incorrectly tells 3% of them that they have AIDS. (a) What is the probability that a person picked at random will test positive? (b) What is the probability that you have AIDS given that your test comes back positive?

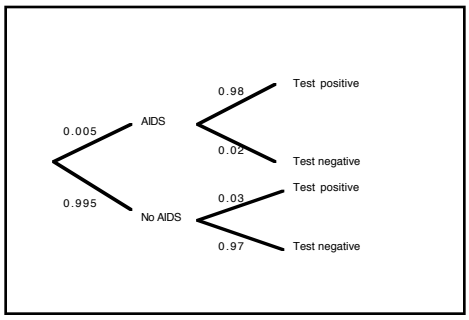
**Answer:** Let $A$ denote the event of one who has AIDS and let $B$ denote the event that the test comes out positive.

(a) The probability that a person picked at random will test positive is given by

$$\begin{aligned} P(\text{test positive}) &= (0.005)\,(0.98) + (0.995)\,(0.03) \\ &= 0.0049 + 0.0298 = 0.035. \end{aligned}$$

(b) The probability that you have AIDS given that your test comes back positive is given by

$$\begin{aligned} P(A/B) &= \frac{\text{favorable positive branches}}{\text{total positive branches}} \\ &= \frac{(0.005)\,(0.98)}{(0.005)\,(0.98) + (0.995)\,(0.03)} \\ &= \frac{0.0049}{0.035} = 0.14. \end{aligned}$$




**Remark 2.2.** This example illustrates why Bayes' theorem is so important. What we would really like to know in this situation is a first-stage result: Do you have AIDS? But we cannot get this information without an autopsy. The first stage is hidden. But the second stage is not hidden. The best we can do is make a prediction about the first stage. This illustrates why backward conditional probabilities are so useful.

### 2.3. Review Exercises

**1.** Let $P(A) = 0.4$ and $P(A \cup B) = 0.6$. For what value of $P(B)$ are $A$ and $B$ independent?

**2.** A die is loaded in such a way that the probability of the face with $j$ dots turning up is proportional to $j$ for $j = 1, 2, 3, 4, 5, 6$. In 6 independent throws of this die, what is the probability that each face turns up exactly once?

**3.** A system engineer is interested in assessing the reliability of a rocket composed of three stages. At take off, the engine of the first stage of the rocket must lift the rocket off the ground. If that engine accomplishes its task, the engine of the second stage must now lift the rocket into orbit. Once the engines in both stages 1 and 2 have performed successfully, the engine of the third stage is used to complete the rocket's mission. The reliability of the rocket is measured by the probability of the completion of the mission. If the probabilities of successful performance of the engines of stages 1, 2 and 3 are 0.99, 0.97 and 0.98, respectively, find the reliability of the rocket.

**4.** Identical twins come from the same egg and hence are of the same sex. Fraternal twins have a 50-50 chance of being the same sex. Among twins the probability of a fraternal set is $\frac{1}{3}$ and an identical set is $\frac{2}{3}$. If the next set of twins are of the same sex, what is the probability they are identical?

**5.** In rolling a pair of fair dice, what is the probability that a sum of 7 is rolled before a sum of 8 is rolled ?

**6.** A card is drawn at random from an ordinary deck of 52 cards and replaced. This is done a total of 5 independent times. What is the conditional probability of drawing the ace of spades exactly 4 times, given that this ace is drawn at least 4 times?

**7.** Let $A$ and $B$ be independent events with $P(A) = P(B)$ and $P(A \cup B) = 0.5$. What is the probability of the event $A$?

**8.** An urn contains 6 red balls and 3 blue balls. One ball is selected at random and is replaced by a ball of the other color. A second ball is then chosen. What is the conditional probability that the first ball selected is red, given that the second ball was red?

**9.** A family has five children. Assuming that the probability of a girl on each birth was 0.5 and that the five births were independent, what is the probability the family has at least one girl, given that they have at least one boy?

**10.** An urn contains 4 balls numbered 0 through 3. One ball is selected at random and removed from the urn and not replaced. All balls with nonzero numbers less than that of the selected ball are also removed from the urn. Then a second ball is selected at random from those remaining in the urn. What is the probability that the second ball selected is numbered 3?

**11.** English and American spelling are *rigour* and *rigor*, respectively. A man staying at Al Rashid hotel writes this word, and a letter taken at random from his spelling is found to be a vowel. If 40 percent of the English-speaking men at the hotel are English and 60 percent are American, what is the probability that the writer is an Englishman?

**12.** A diagnostic test for a certain disease is said to be 90% accurate in that, if a person has the disease, the test will detect with probability 0.9. Also, if a person does not have the disease, the test will report that he or she doesn't have it with probability 0.9. Only 1% of the population has the disease in question. If the diagnostic test reports that a person chosen at random from the population has the disease, what is the conditional probability that the person, in fact, has the disease?

**13.** A small grocery store had 10 cartons of milk, 2 of which were sour. If you are going to buy the $6^{\text{th}}$ carton of milk sold that day at random, find the probability of selecting a carton of sour milk.

**14.** Suppose $Q$ and $S$ are independent events such that the probability that at least one of them occurs is $\frac{1}{3}$ and the probability that $Q$ occurs but $S$ does not occur is $\frac{1}{9}$. What is the probability of $S$?

**15.** A box contains 2 green and 3 white balls. A ball is selected at random from the box. If the ball is green, a card is drawn from a deck of 52 cards. If the ball is white, a card is drawn from the deck consisting of just the 16 pictures. (a) What is the probability of drawing a king? (b) What is the probability of a white ball was selected given that a king was drawn?

**16.** Five urns are numbered 3,4,5,6 and 7, respectively. Inside each urn is $n^2$ dollars where $n$ is the number on the urn. The following experiment is performed: An urn is selected at random. If its number is a prime number the experimenter receives the amount in the urn and the experiment is over. If its number is not a prime number, a second urn is selected from the remaining four and the experimenter receives the total amount in the two urns selected. What is the probability that the experimenter ends up with exactly twenty-five dollars?

**17.** A cookie jar has 3 red marbles and 1 white marble. A shoebox has 1 red marble and 1 white marble. Three marbles are chosen at random without replacement from the cookie jar and placed in the shoebox. Then 2 marbles are chosen at random and without replacement from the shoebox. What is the probability that both marbles chosen from the shoebox are red?

**18.** A urn contains $n$ black balls and $n$ white balls. Three balls are chosen from the urn at random and without replacement. What is the value of $n$ if the probability is $\frac{1}{12}$ that all three balls are white?

**19.** An urn contains 10 balls numbered 1 through 10. Five balls are drawn at random and without replacement. Let $A$ be the event that "Exactly two odd-numbered balls are drawn and they occur on odd-numbered draws from the urn." What is the probability of event $A$?

**20.** I have five envelopes numbered 3, 4, 5, 6, 7 all hidden in a box. I pick an envelope – if it is prime then I get the square of that number in dollars. Otherwise (without replacement) I pick another envelope and then get the sum of squares of the two envelopes I picked (in dollars). What is the probability that I will get \$25?

# Chapter 3
# RANDOM VARIABLES AND DISTRIBUTION FUNCTIONS

## 3.1. Introduction

In many random experiments, the elements of sample space are not necessarily numbers. For example, in a coin tossing experiment the sample space consists of

$$S = \{\text{Head},\ \text{Tail}\}.$$

Statistical methods involve primarily numerical data. Hence, one has to 'mathematize' the outcomes of the sample space. This mathematization, or quantification, is achieved through the notion of random variables.

**Definition 3.1.** Consider a random experiment whose sample space is $S$. A *random variable* $X$ is a function from the sample space $S$ into the set of real numbers $\mathbb{R}$ such that for each interval $I$ in $\mathbb{R}$, the set $\{s \in S \mid X(s) \in I\}$ is an event in $S$.

In a particular experiment a random variable $X$ would be some function that assigns a real number $X(s)$ to each possible outcome $s$ in the sample space. Given a random experiment, there can be many random variables. This is due to the fact that given two (finite) sets $A$ and $B$, the number of distinct functions one can come up with is $|B|^{|A|}$. Here $|A|$ means the cardinality of the set $A$.

Random variable is not a variable. Also, it is not random. Thus someone named it inappropriately. The following analogy speaks the role of the random variable. Random variable is like the *Holy Roman Empire – it was*

*not holy, it was not Roman, and it was not an empire.* A random variable is neither random nor variable, it is simply a function. The values it takes on are both random and variable.

**Definition 3.2.** The set $\{x \in \mathbb{R} \,|\, x = X(s), s \in S\}$ is called the *space of the random variable* $X$.

The space of the random variable $X$ will be denoted by $R_X$. The space of the random variable $X$ is actually the range of the function $X : S \to \mathbb{R}$.

**Example 3.1.** Consider the coin tossing experiment. Construct a random variable $X$ for this experiment. What is the space of this random variable $X$?

**Answer:** The sample space of this experiment is given by

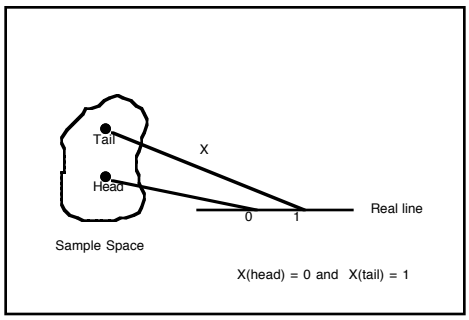
$$S = \{\text{Head},\ \text{Tail}\}.$$

Let us define a function from $S$ into the set of reals as follows

$$X(Head) = 0$$
$$X(Tail) = 1.$$

Then $X$ is a valid map and thus by our definition of random variable, it is a random variable for the coin tossing experiment. The space of this random variable is

$$R_X = \{0, 1\}.$$




**Example 3.2.** Consider an experiment in which a coin is tossed ten times. What is the sample space of this experiment? How many elements are in this sample space? Define a random variable for this sample space and then find the space of the random variable.

**Answer:** The sample space of this experiment is given by

$$S = \{s \,|\, s \text{ is a sequence of 10 heads or tails}\}.$$

The cardinality of $S$ is

$$|S| = 2^{10}.$$

Let $X : S \to \mathbb{R}$ be a function from the sample space $S$ into the set of reals $\mathbb{R}$ defined as follows:

$$X(s) = \text{number of heads in sequence } s.$$

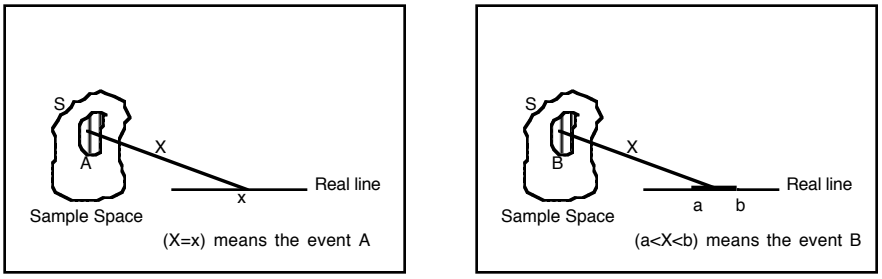
Then $X$ is a random variable. This random variable, for example, maps the sequence $HHTTTHTTHH$ to the real number 5, that is

$$X(HHTTTHTTHH) = 5.$$

The space of this random variable is

$$R_X = \{0, 1, 2, ..., 10\}.$$

Now, we introduce some notations. By $(X = x)$ we mean the event $\{s \in S \,|\, X(s) = x\}$. Similarly, $(a < X < b)$ means the event $\{s \in S \,|\, a < X < b\}$ of the sample space $S$. These are illustrated in the following diagrams.




**Definition 3.3.** If the space of random variable $X$ is countable, then $X$ is called a *discrete random variable*.

**Definition 3.4.** If the space of random variable $X$ is uncountable, then $X$ is called a *continuous random variable*.

In the case of a continuous random variable, the space is either an interval or a union of intervals. A random variable is characterized through its probability density function. First, we consider the discrete case and then we examine the continuous case.

### 3.2. Distribution Functions of Discrete Random Variables

**Definition 3.5.** Let $R_X$ be the space of the random variable $X$. The function $f : R_X \to \mathbb{R}$ defined by

$$f(x) = P(X = x)$$

is called the *probability density function* (pdf) of $X$.

**Example 3.3.** In an introductory statistics class of 50 students, there are 11 freshman, 19 sophomores, 14 juniors and 6 seniors. One student is selected at random. What is the sample space of this experiment? Construct a random variable $X$ for this sample space and then find its space. Further, find the probability density function of this random variable $X$.

**Answer:** The sample space of this random experiment is

$$S = \{Fr,\ So,\ Jr,\ Sr\}.$$

Define a function $X : S \to \mathbb{R}$ as follows:

$$\begin{aligned} X(Fr) &= 1, \qquad X(So) = 2 \\ X(Jr) &= 3, \qquad X(Sr) = 4. \end{aligned}$$

Then clearly $X$ is a random variable in $S$. The space of $X$ is given by

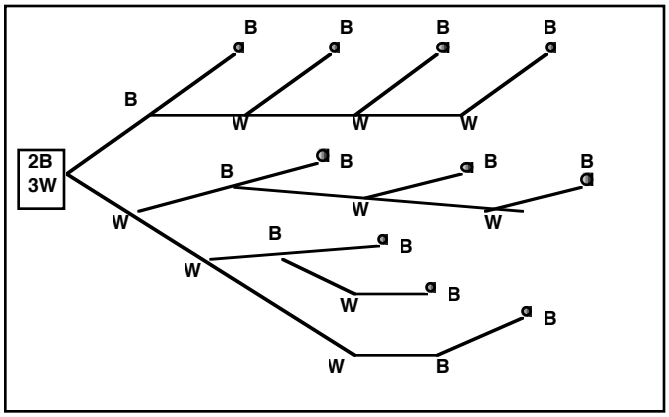
$$R_X = \{1, 2, 3, 4\}.$$

The probability density function of $X$ is given by

$$\begin{aligned} f(1) = P(X = 1) &= \frac{11}{50} \\ f(2) = P(X = 2) &= \frac{19}{50} \\ f(3) = P(X = 3) &= \frac{14}{50} \\ f(4) = P(X = 4) &= \frac{6}{50}. \end{aligned}$$

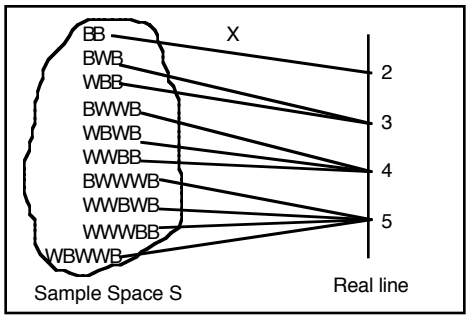
**Example 3.4.** A box contains 5 colored balls, 2 black and 3 white. Balls are drawn successively without replacement. If the random variable $X$ is the number of draws until the last black ball is obtained, find the probability density function for the random variable $X$.

**Answer:** Let 'B' denote the black ball, and 'W' denote the white ball. Then the sample space $S$ of this experiment is given by (see the figure below)




$$S = \{\, BB,\, BWB,\, WBB,\, BWWB,\, WBWB,\, WWBB,$$
$$BWWWB,\, WWBWB,\, WWWBB,\, WBWWB\}.$$

Hence the sample space has 10 points, that is $|S| = 10$. It is easy to see that the space of the random variable $X$ is $\{2, 3, 4, 5\}$.




Therefore, the probability density function of $X$ is given by

$$f(2) = P(X = 2) = \frac{1}{10}, \qquad f(3) = P(X = 3) = \frac{2}{10}$$
$$f(4) = P(X = 4) = \frac{3}{10}, \qquad f(5) = P(X = 5) = \frac{4}{10}.$$

Thus

$$f(x) = \frac{x-1}{10}, \qquad x = 2,\, 3,\, 4,\, 5.$$

**Example 3.5.** A pair of dice consisting of a *six-sided* die and a *four-sided* die is rolled and the sum is determined. Let the random variable $X$ denote this sum. Find the sample space, the space of the random variable, and probability density function of $X$.

**Answer:** The sample space of this random experiment is given by

$$S = \begin{matrix} \{(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6)\} \end{matrix}$$

The space of the random variable $X$ is given by

$$R_X = \{2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\}.$$

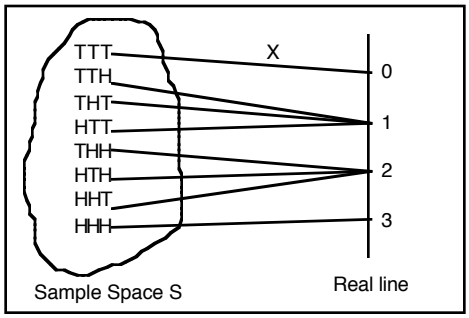
Therefore, the probability density function of $X$ is given by

$$f(2) = P(X = 2) = \frac{1}{24}, \qquad f(3) = P(X = 3) = \frac{2}{24}$$
$$f(4) = P(X = 4) = \frac{3}{24}, \qquad f(5) = P(X = 5) = \frac{4}{24}$$
$$f(6) = P(X = 6) = \frac{4}{24}, \qquad f(7) = P(X = 7) = \frac{4}{24}$$
$$f(8) = P(X = 8) = \frac{3}{24}, \qquad f(9) = P(X = 9) = \frac{2}{24}$$
$$f(10) = P(X = 10) = \frac{1}{24}.$$

**Example 3.6.** A fair coin is tossed 3 times. Let the random variable $X$ denote the number of heads in 3 tosses of the coin. Find the sample space, the space of the random variable, and the probability density function of $X$.

**Answer:** The sample space $S$ of this experiment consists of all binary sequences of length 3, that is

$$S = \{TTT,\ TTH,\ THT,\ HTT,\ THH,\ HTH,\ HHT,\ HHH\}.$$



The space of this random variable is given by

$$R_X = \{0, 1, 2, 3\}.$$

Therefore, the probability density function of $X$ is given by

$$\begin{aligned} f(0) &= P(X = 0) = \frac{1}{8} \\ f(1) &= P(X = 1) = \frac{3}{8} \\ f(2) &= P(X = 2) = \frac{3}{8} \\ f(3) &= P(X = 3) = \frac{1}{8}. \end{aligned}$$

This can be written as follows:

$$f(x) = \binom{3}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{3-x} \qquad x = 0, 1, 2, 3.$$

The probability density function $f(x)$ of a random variable $X$ completely characterizes it. Some basic properties of a discrete probability density function are summarized below.

**Theorem 3.1.** If $X$ is a discrete random variable with space $R_X$ and probability density function $f(x)$, then

(a) $f(x) \geq 0$ for all $x$ in $R_X$, and

(b) $\sum_{x \in R_X} f(x) = 1.$

**Example 3.7.** If the probability of a random variable $X$ with space $R_X = \{1, 2, 3, ..., 12\}$ is given by

$$f(x) = k\,(2x - 1),$$

then, what is the value of the constant $k$?

**Answer:**

$$\begin{aligned}
1 &= \sum_{x \in R_X} f(x) \\
&= \sum_{x \in R_X} k\,(2x-1) \\
&= \sum_{x=1}^{12} k\,(2x-1) \\
&= k\left[2\sum_{x=1}^{12} x - 12\right] \\
&= k\left[2\frac{(12)(13)}{2} - 12\right] \\
&= k\,144.
\end{aligned}$$

Hence

$$k = \frac{1}{144}.$$

**Definition 3.6.** The cumulative distribution function $F(x)$ of a random variable $X$ is defined as

$$F(x) = P(X \leq x)$$

for all real numbers $x$.

**Theorem 3.2.** If $X$ is a random variable with the space $R_X$, then

$$F(x) = \sum_{t \leq x} f(t)$$

for $x \in R_X$.

**Example 3.8.** If the probability density function of the random variable $X$ is given by

$$\frac{1}{144}\,(2x-1) \qquad \text{for } x = 1, 2, 3, ..., 12$$

then find the cumulative distribution function of $X$.

**Answer:** The space of the random variable $X$ is given by

$$R_X = \{1,\ 2,\ 3, ...,\ 12\}.$$

Then

$$F(1) = \sum_{t\leq 1} f(t) = f(1) = \frac{1}{144}$$

$$F(2) = \sum_{t\leq 2} f(t) = f(1) + f(2) = \frac{1}{144} + \frac{3}{144} = \frac{4}{144}$$

$$F(3) = \sum_{t\leq 3} f(t) = f(1) + f(2) + f(3) = \frac{1}{144} + \frac{3}{144} + \frac{5}{144} = \frac{9}{144}$$

.. .....

.. .....

$$F(12) = \sum_{t\leq 12} f(t) = f(1) + f(2) + \cdots + f(12) = 1.$$

$F(x)$ represents the accumulation of $f(t)$ up to $t \leq x$.

**Theorem 3.3.** Let $X$ be a random variable with cumulative distribution function $F(x)$. Then the cumulative distribution function satisfies the followings:

(a) $F(-\infty) = 0$,

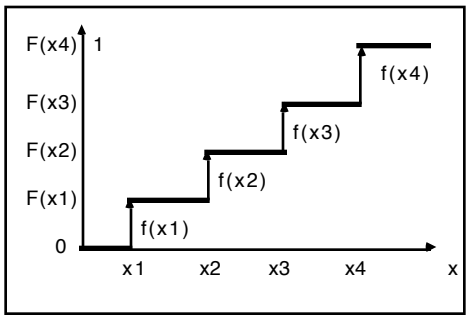
(b) $F(\infty) = 1$, and

(c) $F(x)$ is an increasing function, that is if $x < y$, then $F(x) \leq F(y)$ for all reals $x$, $y$.

The proof of this theorem is trivial and we leave it to the students.

**Theorem 3.4.** If the space $R_X$ of the random variable $X$ is given by $R_X = \{x_1 < x_2 < x_3 < \cdots < x_n\}$, then

$$\begin{aligned}
f(x_1) &= F(x_1)\\
f(x_2) &= F(x_2) - F(x_1)\\
f(x_3) &= F(x_3) - F(x_2)\\
&\text{.. .....}\\
&\text{.. .....}\\
f(x_n) &= F(x_n) - F(x_{n-1}).
\end{aligned}$$




Theorem 3.2 tells us how to find cumulative distribution function from the probability density function, whereas Theorem 3.4 tells us how to find the probability density function given the cumulative distribution function.

**Example 3.9.** Find the probability density function of the random variable $X$ whose cumulative distribution function is

$$F(x) = \begin{cases} 0.00 & \text{if } x < -1 \\ 0.25 & \text{if } -1 \leq x < 1 \\ 0.50 & \text{if } 1 \leq x < 3 \\ 0.75 & \text{if } 3 \leq x < 5 \\ 1.00 & \text{if } x \geq 5 \ . \end{cases}$$

Also, find (a) $P(X \leq 3)$, (b) $P(X = 3)$, and (c) $P(X < 3)$.

**Answer:** The space of this random variable is given by

$$R_X = \{-1, 1, 3, 5\}.$$

By the previous theorem, the probability density function of $X$ is given by

$$\begin{aligned} f(-1) &= 0.25 \\ f(1) &= 0.50 - 0.25 = 0.25 \\ f(3) &= 0.75 - 0.50 = 0.25 \\ f(5) &= 1.00 - 0.75 = 0.25. \end{aligned}$$

The probability $P(X \leq 3)$ can be computed by using the definition of $F$. Hence

$$P(X \leq 3) = F(3) = 0.75.$$

The probability $P(X = 3)$ can be computed from

$$P(X = 3) = F(5) - F(3) = 1 - 0.75 = 0.25.$$

Finally, we get $P(X < 3)$ from

$$P(X < 3) = P(X \leq 1) = F(1) = 0.5.$$

We close this section with an example showing that there is no one-to-one correspondence between a random variable and its distribution function. Consider a coin tossing experiment with the sample space consisting of a head and a tail, that is $S = \{\, head,\ tail \,\}$. Define two random variables $X_1$ and $X_2$ from $S$ as follows:

$$X_1(\,\text{head}\,) = 0 \qquad \text{and} \qquad X_1(\,\text{tail}\,) = 1$$

and

$$X_2(\,\text{head}\,) = 1 \qquad \text{and} \qquad X_2(\,\text{tail}\,) = 0.$$

It is easy to see that both these random variables have the same distribution function, namely

$$F_{X_i}(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x \end{cases}$$

for $i = 1, 2$. Hence there is no one-to-one correspondence between a random variable and its distribution function.

### 3.3. Distribution Functions of Continuous Random Variables

Recall that a random variable $X$ is said to be continuous if its space is either an interval or a union of intervals.

**Definition 3.7.** Let $X$ be a continuous random variable whose space is the set of real numbers $\mathbb{R}$. A nonnegative real valued function $f : \mathbb{R} \to \mathbb{R}$ is said to be the probability density function for the continuous random variable $X$ if it satisfies:

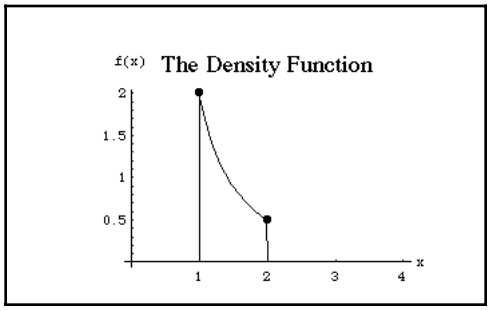
(a) $\int_{-\infty}^{\infty} f(x)\, dx = 1$, and

(b) if $A$ is an event, then $P(A) = \int_A f(x)\, dx$.

**Example 3.10.** Is the real valued function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 2\,x^{-2} & \text{if } 1 < x < 2 \\ 0 & \text{otherwise,} \end{cases}$$

a probability density function for some random variable $X$?




**Answer:** We have to show that $f$ is nonnegative and the area under $f(x)$ is unity. Since the domain of $f$ is the interval $(0, 1)$, it is clear that $f$ is nonnegative. Next, we calculate

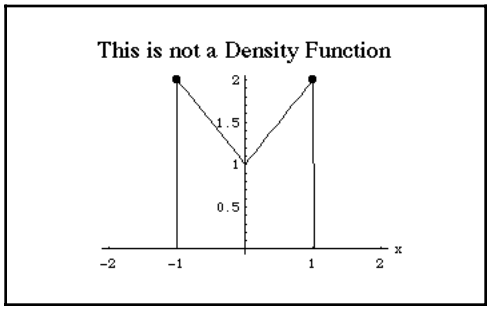
$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \int_{1}^{2} 2\,x^{-2}\,dx \\ &= -2\,\left[\frac{1}{x}\right]_{1}^{2} \\ &= -2\,\left[\frac{1}{2}-1\right] \\ &= 1.\end{aligned}$$

Thus $f$ is a probability density function.

**Example 3.11.** Is the real valued function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1+|x| & \text{if } -1 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

a probability density function for some random variable $X$?



**Answer:** It is easy to see that $f$ is nonnegative, that is $f(x) \geq 0$, since $f(x) = 1 + |x|$. Next we show that the area under $f$ is not unity. For this we compute

$$\begin{aligned}
\int_{-1}^{1} f(x)\,dx &= \int_{-1}^{1} (1 + |x|)\,dx \\
&= \int_{-1}^{0} (1 - x)\,dx + \int_{0}^{1} (1 + x)\,dx \\
&= \left[x - \frac{1}{2}x^2\right]_{-1}^{0} + \left[x + \frac{1}{2}x^2\right]_{0}^{1} \\
&= 1 + \frac{1}{2} + 1 + \frac{1}{2} \\
&= 3.
\end{aligned}$$

Thus $f$ is not a probability density function for some random variable $X$.

**Example 3.12.** For what value of the constant $c$, the real valued function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \frac{c}{1 + (x - \theta)^2}, \qquad -\infty < x < \infty,$$

where $\theta$ is a real parameter, is a probability density function for random variable $X$?

**Answer:** Since $f$ is nonnegative, we see that $c \geq 0$. To find the value of $c$, we use the fact that for pdf the area is unity, that is

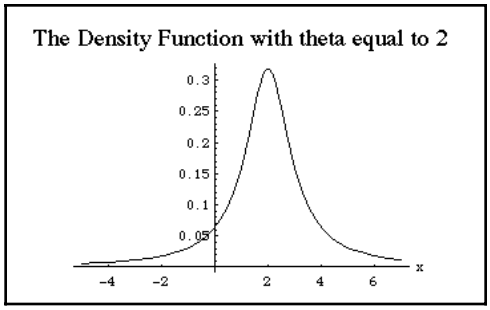
$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x)\,dx \\
&= \int_{-\infty}^{\infty} \frac{c}{1 + (x - \theta)^2}\,dx \\
&= \int_{-\infty}^{\infty} \frac{c}{1 + z^2}\,dz \\
&= c\,\left[\tan^{-1} z\right]_{-\infty}^{\infty} \\
&= c\,\left[\tan^{-1}(\infty) - \tan^{-1}(-\infty)\right] \\
&= c\,\left[\frac{1}{2}\,\pi + \frac{1}{2}\,\pi\right] \\
&= c\,\pi.
\end{aligned}$$

Hence $c = \frac{1}{\pi}$ and the density function becomes

$$f(x) = \frac{1}{\pi\,[1 + (x - \theta)^2]}, \qquad -\infty < x < \infty.$$

This density function is called the Cauchy distribution function with parameter $\theta$. If a random variable $X$ has this pdf then it is called a Cauchy random variable and is denoted by $X \sim CAU(\theta)$.

This distribution is symmetrical about $\theta$. Further, it achieves it maximum at $x = \theta$. The following figure illustrates the symmetry of the distribution for $\theta = 2$.




**Example 3.13.** For what value of the constant $c$, the real valued function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} c & \text{if } a \leq x \leq b \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $a$, $b$ are real constants, is a probability density function for random variable $X$?

**Answer:** Since $f$ is a pdf, $k$ is nonnegative. Further, since the area under $f$ is unity, we get

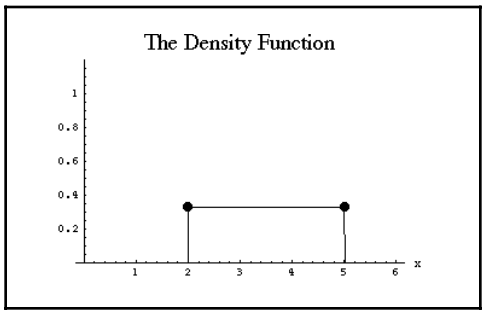
$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)\, dx \\ &= \int_{a}^{b} c\, dx \\ &= c\, [x]_a^b \\ &= c\, [b - a]. \end{aligned}$$

Hence $c = \frac{1}{b-a}$, and the pdf becomes

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ \\ 0 & \text{otherwise.} \end{cases}$$

This probability density function is called the uniform distribution on the interval $[a, b]$. If a random variable $X$ has this pdf then it is called a

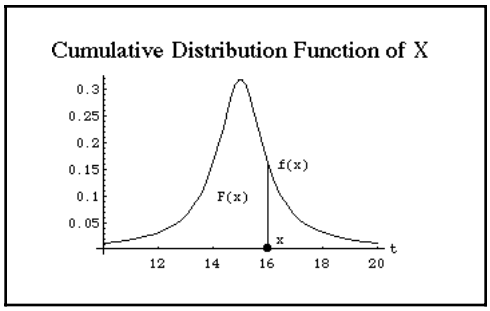
uniform random variable and is denoted by $X \sim UNIF(a, b)$. The following is a graph of the probability density function of a random variable on the interval $[2, 5]$.




**Definition 3.8.** Let $f(x)$ be the probability density function of a continuous random variable $X$. The cumulative distribution function $F(x)$ of $X$ is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

The cumulative distribution function $F(x)$ represents the area under the probability density function $f(x)$ on the interval $(-\infty, x)$ (see figure below).




Like the discrete case, the cdf is an increasing function of $x$, and it takes value 0 at negative infinity and 1 at positive infinity.

**Theorem 3.5.** If $F(x)$ is the cumulative distribution function of a continuous random variable $X$, the probability density function $f(x)$ of $X$ is the derivative of $F(x)$, that is

$$\frac{d}{dx}F(x) = f(x).$$

**Proof:** By Fundamental Theorem of Calculus, we get

$$\begin{aligned}\frac{d}{dx}\left(F(x)\right) &= \frac{d}{dx}\left(\int_{-\infty}^{x} f(t)\,dt\right)\\ &= f(x)\frac{dx}{dx}\\ &= f(x).\end{aligned}$$

This theorem tells us that if the random variable is continuous, then we can find the pdf given cdf by taking the derivative of the cdf. Recall that for discrete random variable, the pdf at a point in space of the random variable can be obtained from the cdf by taking the difference between the cdf at the point and the cdf immediately below the point.

**Example 3.14.** What is the cumulative distribution function of the Cauchy random variable with parameter $\theta$?

**Answer:** The cdf of $X$ is given by

$$\begin{aligned}F(x) &= \int_{-\infty}^{x} f(t)\,dt\\ &= \int_{-\infty}^{x} \frac{1}{\pi\left[1+(t-\theta)^2\right]}\,dt\\ &= \int_{-\infty}^{x-\theta} \frac{1}{\pi\left[1+z^2\right]}\,dz\\ &= \frac{1}{\pi}\,\tan^{-1}\left(x-\theta\right)+\frac{1}{2}.\end{aligned}$$

**Example 3.15.** What is the probability density function of the random variable whose cdf is

$$F(x) = \frac{1}{1+e^{-x}}, \qquad -\infty < x < \infty?$$

**Answer:** The pdf of the random variable is given by

$$\begin{aligned}f(x) &= \frac{d}{dx}F(x)\\ &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)\\ &= \frac{d}{dx}\left(1+e^{-x}\right)^{-1}\\ &= (-1)\left(1+e^{-x}\right)^{-2}\frac{d}{dx}\left(1+e^{-x}\right)\\ &= \frac{e^{-x}}{(1+e^{-x})^2}.\end{aligned}$$

Next, we briefly discuss the problem of finding probability when the cdf is given. We summarize our results in the following theorem.

**Theorem 3.6.** Let $X$ be a continuous random variable whose cdf is $F(x)$. Then followings are true:

(a) $P(X < x) = F(x)$,
(b) $P(X > x) = 1 - F(x)$,
(c) $P(X = x) = 0$ , and
(d) $P(a < X < b) = F(b) - F(a)$.

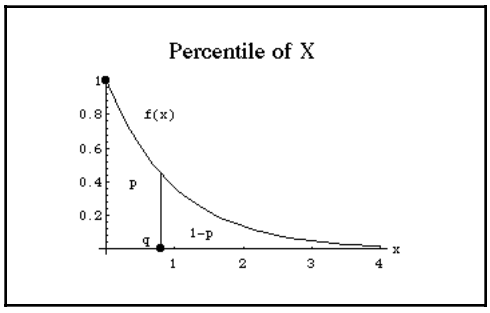
### 3.4. Percentiles for Continuous Random Variables

In this section, we discuss various percentiles of a continuous random variable. If the random variable is discrete, then to discuss percentile, we have to know the order statistics of samples. We shall treat the percentile of discrete random variable in Chapter 13.

**Definition 3.9.** Let $p$ be a real number between 0 and 1. A $100\text{p}^{\text{th}}$ percentile of the distribution of a random variable $X$ is any real number $q$ satisfying

$$P(X \leq q) \leq p \qquad \text{and} \qquad P(X > q) \leq 1 - p.$$

A $100\text{p}^{\text{th}}$ percentile is a measure of location for the probability distribution in the sense that $q$ divides the distribution of the probability mass into two parts, one having probability mass $p$ and other having probability mass $1 - p$ (see diagram below).




**Example 3.16.** If the random variable $X$ has the density function

$$f(x) = \begin{cases} e^{x-2} & \text{for } x < 2 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the 75th percentile of $X$?

**Answer:** Since $100\text{p}^{\text{th}} = 75$, we get $p = 0.75$. By definition of percentile, we have

$$\begin{aligned} 0.75 = p &= \int_{-\infty}^{q} f(x)\,dx \\ &= \int_{-\infty}^{q} e^{x-2}\,dx \\ &= \left[e^{x-2}\right]_{-\infty}^{q} \\ &= e^{q-2}. \end{aligned}$$

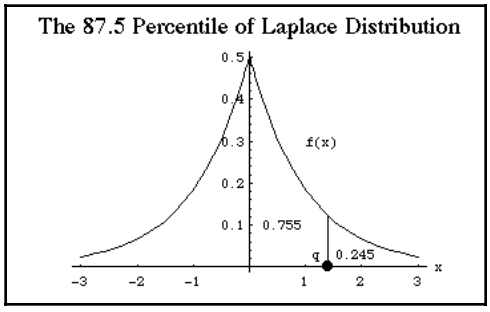
From this solving for $q$, we get the 75th percentile to be

$$q = 2 + \ln\left(\frac{3}{4}\right).$$

**Example 3.17.** What is the 87.5 percentile for the distribution with density function

$$f(x) = \frac{1}{2}e^{-|x|} \qquad -\infty < x < \infty?$$

**Answer:** Note that this density function is symmetric about the $y$-axis, that is $f(x) = f(-x)$.




Hence

$$\int_{-\infty}^{0} f(x)\,dx = \frac{1}{2}.$$

Now we compute the 87.5th percentile $q$ of the above distribution.

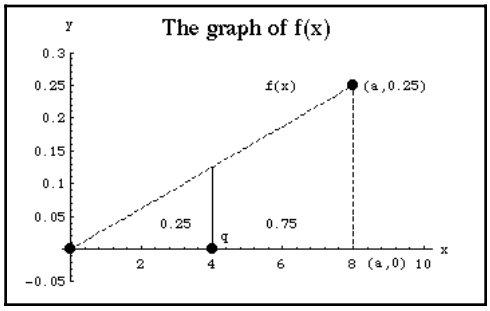
$$\begin{aligned}\frac{87.5}{100} &= \int_{-\infty}^{q} f(x)\,dx \\ &= \int_{-\infty}^{0} \frac{1}{2}\,e^{-|x|}\,dx + \int_{0}^{q} \frac{1}{2}\,e^{-|x|}\,dx \\ &= \int_{-\infty}^{0} \frac{1}{2}\,e^{x}\,dx + \int_{0}^{q} \frac{1}{2}\,e^{-x}\,dx \\ &= \frac{1}{2} + \int_{0}^{q} \frac{1}{2}\,e^{-x}\,dx \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2}e^{-q}.\end{aligned}$$

Therefore solving for $q$, we get

$$\begin{aligned}0.125 &= \frac{1}{2}\,e^{-q} \\ q &= -\ln\left(\frac{25}{100}\right) = \ln 4.\end{aligned}$$

Hence the 87.5th percentile of the distribution is $\ln 4$.

**Example 3.18.** Let the continuous random variable $X$ have the density function $f(x)$ as shown in the figure below:




What is the 25th percentile of the distribution of $X$?

**Answer:** Since the line passes through the points $(0,0)$ and $(a,\frac{1}{4})$, the function $f(x)$ is equal to

$$f(x) = \frac{1}{4a}\,x.$$

Since $f(x)$ is a density function the area under $f(x)$ should be unity. Hence

$$\begin{aligned} 1 &= \int_0^a f(x)\,dx \\ &= \int_0^a \frac{1}{4a}\,x\,dx \\ &= \frac{1}{8a}\,a^2 \\ &= \frac{a}{8}. \end{aligned}$$

Thus $a = 8$. Hence the probability density function of $X$ is

$$f(x) = \frac{1}{32}\,x.$$

Now we want to find the 25th percentile.

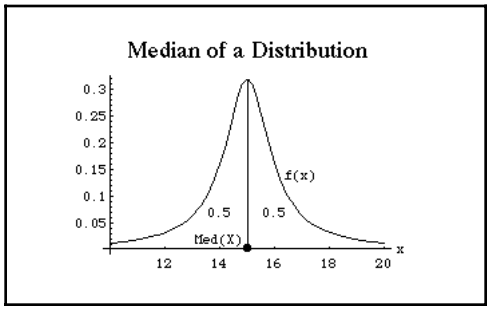
$$\begin{aligned} \frac{25}{100} &= \int_0^q f(x)\,dx \\ &= \int_0^q \frac{1}{32}\,x\,dx \\ &= \frac{1}{64}\,q^2. \end{aligned}$$

Hence $q = \sqrt{16}$, that is the 25th percentile of the above distribution is 4.

**Definition 3.10.** The 25th and 75th percentiles of any distribution are called the *first* and the *third* quartiles, respectively.

**Definition 3.11.** The 50th percentile of any distribution is called the *median* of the distribution.

The median divides the distribution of the probability mass into two equal parts (see the following figure).



If a probability density function $f(x)$ is symmetric about the $y$-axis, then the median is always 0.

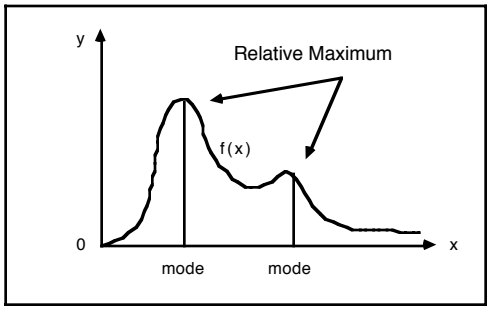
**Example 3.19.** A random variable is called standard normal if its probability density function is of the form

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

What is the median of $X$?

**Answer:** Notice that $f(x) = f(-x)$, hence the probability density function is symmetric about the $y$-axis. Thus the median of $X$ is 0.

**Definition 3.12.** A mode of the distribution of a continuous random variable $X$ is the value of $x$ where the probability density function $f(x)$ attains a relative maximum (see diagram).




A mode of a random variable $X$ is one of its most probable values. A random variable can have more than one mode.

**Example 3.20.** Let $X$ be a uniform random variable on the interval $[0, 1]$, that is $X \sim UNIF(0, 1)$. How many modes does $X$ have?

**Answer:** Since $X \sim UNIF(0, 1)$, the probability density function of $X$ is

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Hence the derivative of $f(x)$ is

$$f'(x) = 0 \qquad x \in (0, 1).$$

Therefore $X$ has infinitely many modes.

**Example 3.21.** Let $X$ be a Cauchy random variable with parameter $\theta = 0$, that is $X \sim CAU(0)$. What is the mode of $X$?

**Answer:** Since $X \sim CAU(0)$, the probability density function of $f(x)$ is

$$f(x) = \frac{1}{\pi\,(1+x^2)} \qquad -\infty < x < \infty.$$

Hence

$$f'(x) = \frac{-2x}{\pi\,(1+x^2)^2}.$$

Setting this derivative to 0, we get $x = 0$. Thus the mode of $X$ is 0.

**Example 3.22.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} x^2\, e^{-bx} & \text{for } x \geq 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $b > 0$. What is the mode of $X$?

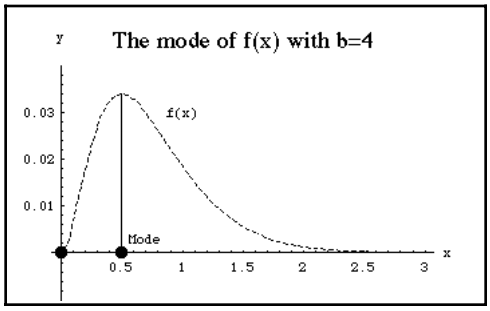
**Answer:**

$$\begin{aligned} 0 &= \frac{df}{dx} \\ &= 2xe^{-bx} - x^2be^{-bx} \\ &= (2-bx)x = 0. \end{aligned}$$

Hence

$$x = 0 \qquad \text{or} \qquad x = \frac{2}{b}.$$

Thus the mode of $X$ is $\frac{2}{b}$. The graph of the $f(x)$ for $b = 4$ is shown below.




**Example 3.23.** A continuous random variable has density function

$$f(x) = \begin{cases} \frac{3x^2}{\theta^3} & \text{for } 0 \leq x \leq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

for some $\theta > 0$. What is the ratio of the mode to the median for this distribution?

**Answer:** For fixed $\theta > 0$, the density function $f(x)$ is an increasing function. Thus, $f(x)$ has maximum at the right end point of the interval $[0, \theta]$. Hence the mode of this distribution is $\theta$.

Next we compute the median of this distribution.

$$\begin{aligned}
\frac{1}{2} &= \int_0^q f(x)\,dx \\
&= \int_0^q \frac{3x^2}{\theta^3}\,dx \\
&= \left[\frac{x^3}{\theta^3}\right]_0^q \\
&= \left[\frac{q^3}{\theta^3}\right].
\end{aligned}$$

Hence

$$q = 2^{-\frac{1}{3}}\,\theta.$$

Thus the ratio of the mode of this distribution to the median is

$$\frac{\text{mode}}{\text{median}} = \frac{\theta}{2^{-\frac{1}{3}}\,\theta} = \sqrt[3]{2}.$$

**Example 3.24.** A continuous random variable has density function

$$f(x) = \begin{cases} \frac{3x^2}{\theta^3} & \text{for } 0 \le x \le \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

for some $\theta > 0$. What is the probability of $X$ less than the ratio of the mode to the median of this distribution?

**Answer:** In the previous example, we have shown that the ratio of the mode to the median of this distribution is given by

$$a := \frac{\text{mode}}{\text{median}} = \sqrt[3]{2}.$$

Hence the probability of $X$ less than the ratio of the mode to the median of this distribution is

$$\begin{aligned} P(X < a) &= \int_0^a f(x)\,dx \\ &= \int_0^a \frac{3x^2}{\theta^3}\,dx \\ &= \left[\frac{x^3}{\theta^3}\right]_0^a \\ &= \frac{a^3}{\theta^3} \\ &= \frac{\left(\sqrt[3]{2}\right)^3}{\theta^3} \\ &= \frac{2}{\theta^3}. \end{aligned}$$

### 3.5. Review Exercises

**1.** Let the random variable $X$ have the density function

$$f(x) = \begin{cases} k\,x & \text{for } 0 \le x \le \sqrt{\frac{2}{k}} \\ 0 & \text{elsewhere.} \end{cases}$$

If the mode of this distribution is at $x = \frac{\sqrt{2}}{4}$, then what is the median of $X$?

**2.** The random variable $X$ has density function

$$f(x) = \begin{cases} c\,x^{k+1}\,(1-x)^k & \text{for } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $c > 0$ and $1 < k < 2$. What is the mode of $X$?

**3.** The random variable $X$ has density function

$$f(x) = \begin{cases} (k+1)\,x^2 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $k$ is a constant. What is the median of $X$?

**4.** What are the median, and mode, respectively, for the density function

$$f(x) = \frac{1}{\pi\,(1+x^2)}, \qquad -\infty < x < \infty?$$

**5.** What is the $10^{\text{th}}$ percentile of the random variable $X$ whose probability density function is

$$f(x) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } x \geq 0, \qquad \theta > 0 \\ \\ 0 & \text{elsewhere?} \end{cases}$$

**6.** What is the median of the random variable $X$ whose probability density function is

$$f(x) = \begin{cases} \frac{1}{2}\, e^{-\frac{x}{2}} & \text{if } x \geq 0 \\ \\ 0 & \text{elsewhere?} \end{cases}$$

**7.** A continuous random variable $X$ has the density

$$f(x) = \begin{cases} \frac{3\,x^2}{8} & \text{for } 0 \leq x \leq 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that $X$ is greater than its $75^{\text{th}}$ percentile?

**8.** What is the probability density function of the random variable $X$ if its cumulative distribution function is given by

$$F(x) = \begin{cases} 0.0 & \text{if } x < 2 \\ 0.5 & \text{if } 2 \leq x < 3 \\ 0.7 & \text{if } 3 \leq x < \pi \\ 1.0 & \text{if } x \geq \pi? \end{cases}$$

**9.** Let the distribution of $X$ for $x > 0$ be

$$F(x) = 1 - \sum_{k=0}^{3} \frac{x^k\, e^{-x}}{k!}.$$

What is the density function of $X$ for $x > 0$?

**10.** Let $X$ be a random variable with cumulative distribution function

$$F(x) = \begin{cases} 1 - e^{-x} & \text{for } x > 0 \\ \\ 0 & \text{for } x \leq 0. \end{cases}$$

What is the $P\left(0 \leq e^X \leq 4\right)$?

**11.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} a\,x^2\,e^{-10\,x} & \text{for } x \geq 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $a > 0$. What is the probability of $X$ greater than or equal to the mode of $X$?

**12.** Let the random variable $X$ have the density function

$$f(x) = \begin{cases} k\,x & \text{for } 0 \leq x \leq \sqrt{\frac{2}{k}} \\ \\ 0 & \text{elsewhere.} \end{cases}$$

If the mode of this distribution is at $x = \frac{\sqrt{2}}{4}$, then what is the probability of $X$ less than the median of $X$?

**13.** The random variable $X$ has density function

$$f(x) = \begin{cases} (k+1)\,x^2 & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $k$ is a constant. What is the probability of $X$ between the first and third quartiles?

**14.** Let $X$ be a random variable having continuous cumulative distribution function $F(x)$. What is the cumulative distribution function $Y = \max(0, -X)$?

**15.** Let $X$ be a random variable with probability density function

$$f(x) = \frac{2}{3^x} \qquad \text{for} \quad x = 1, 2, 3, ....$$

What is the probability that $X$ is even?

**16.** An urn contains 5 balls numbered 1 through 5. Two balls are selected at random without replacement from the urn. If the random variable $X$ denotes the sum of the numbers on the 2 balls, then what are the space and the probability density function of $X$?

**17.** A pair of six-sided dice is rolled and the sum is determined. If the random variable $X$ denotes the sum of the numbers rolled, then what are the space and the probability density function of $X$?

**18.** Five digit codes are selected at random from the set $\{0, 1, 2, ..., 9\}$ with replacement. If the random variable $X$ denotes the number of zeros in randomly chosen codes, then what are the space and the probability density function of $X$?

**19.** A urn contains 10 coins of which 4 are counterfeit. Coins are removed from the urn, one at a time, until all counterfeit coins are found. If the random variable $X$ denotes the number of coins removed to find the first counterfeit one, then what are the space and the probability density function of $X$?

**20.** Let $X$ be a random variable with probability density function

$$f(x) = \frac{2c}{3^x} \qquad \text{for} \quad x = 1, 2, 3, 4, ..., \infty$$

for some constant $c$. What is the value of $c$? What is the probability that $X$ is even?

**21.** If the random variable $X$ possesses the density function

$$f(x) = \begin{cases} cx & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise,} \end{cases}$$

then what is the value of $c$ for which $f(x)$ is a probability density function? What is the cumulative distribution function of $X$. Graph the functions $f(x)$ and $F(x)$. Use $F(x)$ to compute $P(1 \leq X \leq 2)$.

**22.** The length of time required by students to complete a 1-hour exam is a random variable with a pdf given by

$$f(x) = \begin{cases} cx^2 + x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

then what the probability a student finishes in less than a half hour?

**23.** What is the probability of, when blindfolded, hitting a circle inscribed on a square wall?

**24.** Let $f(x)$ be a continuous probability density function. Show that, for every $-\infty < \mu < \infty$ and $\sigma > 0$, the function $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ is also a probability density function.

**25.** Let $X$ be a random variable with probability density function $f(x)$ and cumulative distribution function $F(x)$. True or False?

(a) $f(x)$ can't be larger than 1. (b) $F(x)$ can't be larger than 1. (c) $f(x)$ can't decrease. (d) $F(x)$ can't decrease. (e) $f(x)$ can't be negative. (f) $F(x)$ can't be negative. (g) Area under $f$ must be 1. (h) Area under $F$ must be 1. (i) $f$ can't jump. (j) $F$ can't jump.

# Chapter 4

# MOMENTS OF RANDOM VARIABLES AND CHEBYCHEV INEQUALITY

## 4.1. Moments of Random Variables

In this chapter, we introduce the concepts of various moments of a random variable. Further, we examine the expected value and the variance of random variables in detail. We shall conclude this chapter with a discussion of Chebychev's inequality.

**Definition 4.1.** The $n^{\text{th}}$ moment about the origin of a random variable $X$, as denoted by $E(X^n)$, is defined to be

$$E\left(X^n\right) = \begin{cases} \sum\limits_{x \in R_X} x^n f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

for $n = 0, 1, 2, 3, ....$, provided the right side converges absolutely.

If $n = 1$, then $E(X)$ is called the first moment about the origin. If $n = 2$, then $E(X^2)$ is called the second moment of $X$ about the origin. In general, these moments may or may not exist for a given random variable. If for a random variable, a particular moment does not exist, then we say that the random variable does not have that moment. For these moments to exist one requires absolute convergence of the sum or the integral. Next, we shall define two important characteristics of a random variable, namely the expected value and variance. Occasionally $E\left(X^n\right)$ will be written as $E\left[X^n\right]$.

### 4.2. Expected Value of Random Variables

A random variable $X$ is characterized by its probability density function, which defines the relative likelihood of assuming one value over the others. In Chapter 3, we have seen that given a probability density function $f$ of a random variable $X$, one can construct the distribution function $F$ of it through summation or integration. Conversely, the density function $f(x)$ can be obtained as the marginal value or derivative of $F(x)$. The density function can be used to infer a number of characteristics of the underlying random variable. The two most important attributes are measures of location and dispersion. In this section, we treat the measure of location and treat the other measure in the next section.

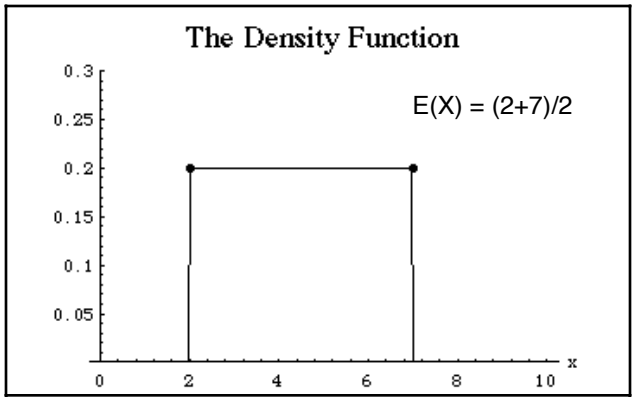
**Definition 4.2.** Let $X$ be a random variable with space $R_X$ and probability density function $f(x)$. The mean $\mu_X$ of the random variable $X$ is defined as

$$\mu_X = \begin{cases} \sum_{x \in R_X} x\, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x\, f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

if the right hand side exists.

The mean of a random variable is a composite of its values weighted by the corresponding probabilities. The mean is a measure of central tendency: the value that the random variable takes "on average." The mean is also called the expected value of the random variable $X$ and is denoted by $E(X)$. The symbol $E$ is called the expectation operator. The expected value of a random variable may or may not exist.

**Example 4.1.** If $X$ is a uniform random variable on the interval $(2, 7)$, then what is the mean of $X$?



**Answer:** The density function of $X$ is

$$f(x) = \begin{cases} \frac{1}{5} & \text{if } 2 < x < 7 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Thus the mean or the expected value of $X$ is

$$\begin{aligned} \mu_X &= E(X) \\ &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ \\ &= \int_2^7 x\, \frac{1}{5}\, dx \\ \\ &= \left[\frac{1}{10}\, x^2\right]_2^7 \\ \\ &= \frac{1}{10}\,(49 - 4) \\ \\ &= \frac{45}{10} \\ \\ &= \frac{9}{2} \\ \\ &= \frac{2+7}{2}. \end{aligned}$$

In general, if $X \sim UNIF(a, b)$, then $E(X) = \frac{1}{2}(a + b)$.

**Example 4.2.** If $X$ is a Cauchy random variable with parameter $\theta$, that is $X \sim CAU(\theta)$, then what is the expected value of $X$?

**Answer:** We want to find the expected value of $X$ if it exists. The expected value of $X$ will exist if the integral $\int_{\mathbb{R}} x f(x) dx$ converges absolutely, that is

$$\int_{\mathbb{R}} |x\, f(x)|\, dx < \infty.$$

If this integral diverges, then the expected value of $X$ does not exist. Hence, let us find out if $\int_{\mathbb{R}} |x\, f(x)|\, dx$ converges or not.

$$\begin{aligned}
&\int_{\mathbb{R}} |x\, f(x)|\, dx \\
&\qquad = \int_{-\infty}^{\infty} |x\, f(x)|\, dx \\
&\qquad = \int_{-\infty}^{\infty} \left| x\, \frac{1}{\pi[1+(x-\theta)^2]} \right| dx \\
&\qquad = \int_{-\infty}^{\infty} \left| (z+\theta)\, \frac{1}{\pi[1+z^2]} \right| dz \\
&\qquad = \theta + 2\int_{0}^{\infty} z\, \frac{1}{\pi[1+z^2]}\, dz \\
&\qquad = \theta + \left[ \frac{1}{\pi}\, \ln(1+z^2) \right]_0^{\infty} \\
&\qquad = \theta + \frac{1}{\pi} \lim_{b\to\infty} \ln(1+b^2) \\
&\qquad = \theta + \infty \\
&\qquad = \infty.
\end{aligned}$$

Since, the above integral does not exist, the expected value for the Cauchy distribution also does not exist.

**Remark 4.1.** Indeed, it can be shown that a random variable $X$ with the Cauchy distribution, $E\left(X^n\right)$, does not exist for any natural number $n$. Thus, Cauchy random variables have no moments at all.

**Example 4.3.** If the probability density function of the random variable $X$ is

$$f(x) = \begin{cases} (1-p)^{x-1}\, p & \text{if } x = 1, 2, 3, 4, ..., \infty \\ 0 & \text{otherwise,} \end{cases}$$

then what is the expected value of $X$?

**Answer:** The expected value of $X$ is

$$\begin{aligned}
E(X) &= \sum_{x \in R_X} x\, f(x) \\
&= \sum_{x=1}^{\infty} x\,(1-p)^{x-1}\, p \\
&= p\,\frac{d}{dp}\left\{ \int \left[ \sum_{x=1}^{\infty} x\,(1-p)^{x-1} \right] dp \right\} \\
&= p\,\frac{d}{dp}\left\{ \left[ \sum_{x=1}^{\infty} \int x\,(1-p)^{x-1} \right] dp \right\} \\
&= -p\,\frac{d}{dp}\left\{ \sum_{x=1}^{\infty} (1-p)^{x} \right\} \\
&= -p\,\frac{d}{dp}\left\{ (1-p)\,\frac{1}{1-(1-p)} \right\} \\
&= -p\,\frac{d}{dp}\left\{ \frac{1}{p} \right\} \\
&= p\left(\frac{1}{p}\right)^{2} \\
&= \frac{1}{p}
\end{aligned}$$

Hence the expected value of $X$ is the reciprocal of the parameter $p$.

**Definition 4.3.** If a random variable $X$ whose probability density function is given by

$$f(x) = \begin{cases} (1-p)^{x-1}\, p & \text{if } x = 1, 2, 3, 4, ..., \infty \\ 0 & \text{otherwise} \end{cases}$$

is called a geometric random variable and is denoted by $X \sim GEO(p)$.

**Example 4.4.** A couple decides to have 3 children. If none of the 3 is a girl, they will try again; and if they still don't get a girl, they will try once more. If the random variable $X$ denotes the number of children the couple will have following this scheme, then what is the expected value of $X$?

**Answer:** Since the couple can have 3 or 4 or 5 children, the space of the random variable $X$ is

$$R_X = \{3,\, 4,\, 5\}.$$

The probability density function of $X$ is given by

$$\begin{aligned} f(3) &= P(X=3) \\ &= P(\text{at least one girl}) \\ &= 1 - P(\text{no girls}) \\ &= 1 - P(3 \text{ boys in 3 tries}) \\ &= 1 - (P(1 \text{ boy in each try}))^3 \\ &= 1 - \left(\frac{1}{2}\right)^3 \\ &= \frac{7}{8}. \end{aligned}$$

$$\begin{aligned} f(4) &= P(X=4) \\ &= P(3 \text{ boys and } 1 \text{ girl in last try}) \\ &= (P(1 \text{ boy in each try}))^3 \; P(1 \text{ girl in last try}) \\ &= \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right) \\ &= \frac{1}{16}. \end{aligned}$$

$$\begin{aligned} f(5) &= P(X=5) \\ &= P(4 \text{ boys and } 1 \text{ girl in last try}) + P(5 \text{ boys in } 5 \text{ tries}) \\ &= P(1 \text{ boy in each try})^4 \, P(1 \text{ girl in last try}) + P(1 \text{ boy in each try})^5 \\ &= \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^5 \\ &= \frac{1}{16}. \end{aligned}$$

Hence, the expected value of the random variable is

$$\begin{aligned} E(X) &= \sum_{x \in R_X} x \, f(x) \\ &= \sum_{x=3}^{5} x \, f(x) \\ &= 3 \, f(3) + 4 \, f(4) + 5 \, f(5) \\ &= 3 \, \frac{14}{16} + 4 \, \frac{1}{16} + 5 \, \frac{1}{16} \\ &= \frac{42 + 4 + 5}{16} \\ &= \frac{51}{16} = 3 \, \frac{3}{16}. \end{aligned}$$

**Remark 4.2.** We interpret this physically as meaning that if many couples have children according to this scheme, it is likely that the average family size would be near $3\frac{3}{16}$ children.

**Example 4.5.** A lot of 8 TV sets includes 3 that are defective. If 4 of the sets are chosen at random for shipment to a hotel, how many defective sets can they expect?

**Answer:** Let $X$ be the random variable representing the number of defective TV sets in a shipment of 4. Then the space of the random variable $X$ is

$$R_X = \{0, 1, 2, 3\}.$$

Then the probability density function of $X$ is given by

$$\begin{aligned} f(x) &= P(X = x) \\ &= P(x \text{ defective TV sets in a shipment of four}) \\ &= \frac{\binom{3}{x}\binom{5}{4-x}}{\binom{8}{4}} \qquad x = 0, 1, 2, 3. \end{aligned}$$

Hence, we have

$$\begin{aligned} f(0) &= \frac{\binom{3}{0}\binom{5}{4}}{\binom{8}{4}} = \frac{5}{70} \\ f(1) &= \frac{\binom{3}{1}\binom{5}{3}}{\binom{8}{4}} = \frac{30}{70} \\ f(2) &= \frac{\binom{3}{2}\binom{5}{2}}{\binom{8}{4}} = \frac{30}{70} \\ f(3) &= \frac{\binom{3}{3}\binom{5}{1}}{\binom{8}{4}} = \frac{5}{70}. \end{aligned}$$

Therefore, the expected value of $X$ is given by

$$\begin{aligned} E(X) &= \sum_{x \in R_X} x\, f(x) \\ &= \sum_{0}^{3} x\, f(x) \\ &= f(1) + 2\, f(2) + 3\, f(3) \\ &= \frac{30}{70} + 2\,\frac{30}{70} + 3\,\frac{5}{70} \\ &= \frac{30 + 60 + 15}{70} \\ &= \frac{105}{70} = 1.5. \end{aligned}$$

**Remark 4.3.** Since they cannot possibly get 1.5 defective TV sets, it should be noted that the term "expect" is not used in its colloquial sense. Indeed, it should be interpreted as an average pertaining to repeated shipments made under given conditions.

Now we prove a result concerning the expected value operator $E$.

**Theorem 4.1.** Let $X$ be a random variable with pdf $f(x)$. If $a$ and $b$ are any two real numbers, then

$$E(aX + b) = a\,E(X) + b.$$

**Proof:** We will prove only for the continuous case.

$$\begin{aligned} E(aX + b) &= \int_{-\infty}^{\infty} (a\,x + b)\,f(x)\,dx \\ &= \int_{-\infty}^{\infty} a\,x\,f(x)\,dx + \int_{-\infty}^{\infty} b\,f(x)\,dx \\ &= a\int_{-\infty}^{\infty} x\,f(x)\,dx + b \\ &= aE(X) + b. \end{aligned}$$

To prove the discrete case, replace the integral by summation. This completes the proof.

### 4.3. Variance of Random Variables

The spread of the distribution of a random variable $X$ is its variance.

**Definition 4.4.** Let $X$ be a random variable with mean $\mu_X$. The variance of $X$, denoted by $Var(X)$, is defined as

$$Var(X) = E\left(\,[\,X - \mu_X\,]^2\,\right).$$

It is also denoted by $\sigma_X^2$. The positive square root of the variance is called the standard deviation of the random variable $X$. Like variance, the standard deviation also measures the spread. The following theorem tells us how to compute the variance in an alternative way.

**Theorem 4.2.** If $X$ is a random variable with mean $\mu_X$ and variance $\sigma_X^2$, then

$$\sigma_X^2 = E(X^2) - (\,\mu_X\,)^2.$$

**Proof:**

$$\begin{aligned}
\sigma_X^2 &= E\left(\,[\,X-\mu_X\,]^2\,\right)\\
&= E\left(\,X^2-2\,\mu_X\,X+\mu_X^2\,\right)\\
&= E(X^2)-2\,\mu_X\,E(X)+(\,\mu_X\,)^2\\
&= E(X^2)-2\,\mu_X\,\mu_X+(\,\mu_X\,)^2\\
&= E(X^2)-(\,\mu_X\,)^2.
\end{aligned}$$

**Theorem 4.3.** If $X$ is a random variable with mean $\mu_X$ and variance $\sigma_X^2$, then

$$Var(aX+b)=a^2\,Var(X),$$

where $a$ and $b$ are arbitrary real constants.

**Proof:**

$$\begin{aligned}
Var(a\,X+b) &= E\left(\,[\,(a\,X+b)-\mu_{aX+b}\,]^2\right)\\
&= E\left(\,[\,a\,X+b-E(a\,X+b)\,]^2\right)\\
&= E\left(\,[\,a\,X+b-a\,\mu_{X+}-b\,]^2\right)\\
&= E\left(a^2\;[\,X-\mu_X\,]^2\right)\\
&= a^2\,E\left(\,[\,X-\mu_X\,]^2\right)\\
&= a^2\,Var(X).
\end{aligned}$$

**Example 4.6.** Let $X$ have the density function

$$f(x)=\begin{cases}\frac{2\,x}{k^2} & \text{for } 0\le x\le k\\[2ex] 0 & \text{otherwise.}\end{cases}$$

For what value of $k$ is the variance of $X$ equal to 2?

**Answer:** The expected value of $X$ is

$$\begin{aligned}
E(X) &= \int_0^k x\,f(x)\,dx\\
&= \int_0^k x\,\frac{2\,x}{k^2}\,dx\\
&= \frac{2}{3}\,k.
\end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_0^k x^2 \, f(x) \, dx \\ &= \int_0^k x^2 \, \frac{2\,x}{k^2} \, dx \\ &= \frac{2}{4} \, k^2. \end{aligned}$$

Hence, the variance is given by

$$\begin{aligned} Var(X) &= E(X^2) - (\,\mu_X\,)^2 \\ &= \frac{2}{4} \, k^2 - \frac{4}{9} \, k^2 \\ &= \frac{1}{18} \, k^2. \end{aligned}$$

Since this variance is given to be 2, we get

$$\frac{1}{18} \, k^2 = 2$$

and this implies that $k = \pm 6$. But $k$ is given to be greater than 0, hence $k$ must be equal to 6.

**Example 4.7.** If the probability density function of the random variable is

$$f(x) = \begin{cases} 1 - |x| & \text{for } |x| < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the variance of $X$?

**Answer:** Since $Var(X) = E(X^2) - \mu_X^2$, we need to find the first and second moments of $X$. The first moment of $X$ is given by

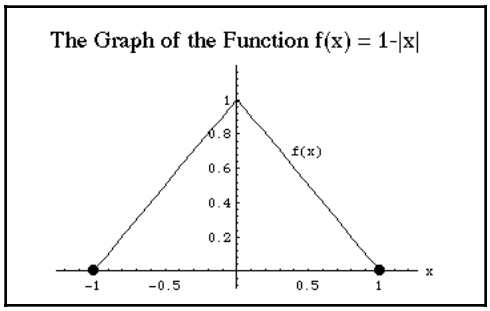
$$\begin{aligned} \mu_X &= E(X) \\ &= \int_{-\infty}^{\infty} x \, f(x) \, dx \\ &= \int_{-1}^{1} x \, (1 - |x|) \, dx \\ &= \int_{-1}^{0} x \, (1 + x) \, dx + \int_{0}^{1} x \, (1 - x) \, dx \\ &= \int_{-1}^{0} (x + x^2) \, dx + \int_{0}^{1} (x - x^2) \, dx \\ &= \frac{1}{3} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} \\ &= 0. \end{aligned}$$

The second moment $E(X^2)$ of $X$ is given by

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)\, dx \\ &= \int_{-1}^{1} x^2 (1 - |x|)\, dx \\ &= \int_{-1}^{0} x^2 (1 + x)\, dx + \int_{0}^{1} x^2 (1 - x)\, dx \\ &= \int_{-1}^{0} (x^2 + x^3)\, dx + \int_{0}^{1} (x^2 - x^3)\, dx \\ &= \frac{1}{3} - \frac{1}{4} + \frac{1}{3} - \frac{1}{4} \\ &= \frac{1}{6}. \end{aligned}$$

Thus, the variance of $X$ is given by

$$Var(X) = E(X^2) - \mu_X^2 = \frac{1}{6} - 0 = \frac{1}{6}.$$




**Example 4.8.** Suppose the random variable $X$ has mean $\mu$ and variance $\sigma^2 > 0$. What are the values of the numbers $a$ and $b$ such that $a + bX$ has mean 0 and variance 1?

**Answer:** The mean of the random variable is 0. Hence

$$\begin{aligned} 0 &= E(a + bX) \\ &= a + b\, E(X) \\ &= a + b\, \mu. \end{aligned}$$

Thus $a = -b\,\mu$. Similarly, the variance of $a + bX$ is 1. That is

$$\begin{aligned} 1 &= Var(a + bX) \\ &= b^2\, Var(X) \\ &= b^2\, \sigma^2. \end{aligned}$$

Hence

$$b = \frac{1}{\sigma} \qquad \text{and} \qquad a = -\frac{\mu}{\sigma}$$

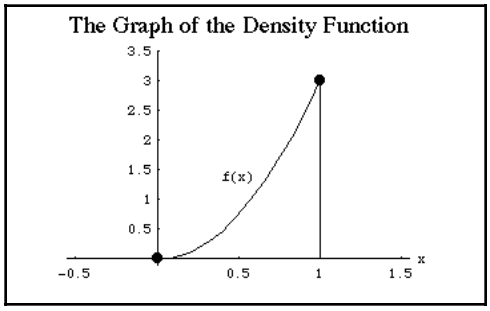
or

$$b = -\frac{1}{\sigma} \qquad \text{and} \qquad a = \frac{\mu}{\sigma}.$$

**Example 4.9.** Suppose $X$ has the density function

$$f(x) = \begin{cases} 3\,x^2 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected area of a random isosceles right triangle with hypotenuse $X$?



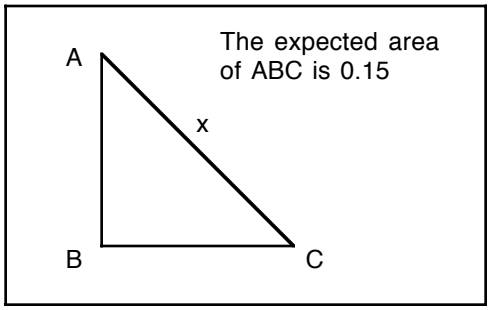

**Answer:** Let $ABC$ denote this random isosceles right triangle. Let $AC = x$. Then

$$AB = BC = \frac{x}{\sqrt{2}}$$

$$\text{Area of ABC} = \frac{1}{2}\,\frac{x}{\sqrt{2}}\,\frac{x}{\sqrt{2}} = \frac{x^2}{4}$$

The expected area of this random triangle is given by

$$E(\text{area of random ABC}) = \int_0^1 \frac{x^2}{4}\,3\,x^2\,dx = \frac{3}{20}.$$



For the next example, we need these following results. For $-1 < x < 1$, let

$$g(x) = \sum_{k=0}^{\infty} a\, x^k = \frac{a}{1-x}.$$

Then

$$g'(x) = \sum_{k=1}^{\infty} a\, k\, x^{k-1} = \frac{a}{(1-x)^2},$$

and

$$g''(x) = \sum_{k=2}^{\infty} a\, k\, (k-1)\, x^{k-2} = \frac{2\, a}{(1-x)^3}.$$

**Example 4.10.** If the probability density function of the random variable $X$ is

$$f(x) = \begin{cases} (1-p)^{x-1}\, p & \text{if } x = 1, 2, 3, 4, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the variance of $X$?

**Answer:** We want to find the variance of $X$. But variance of $X$ is defined as

$$\begin{aligned} Var(X) &= E\left(X^2\right) - \left[\, E(X)\,\right]^2 \\ &= E(X(X-1)) + E(X) - \left[\, E(X)\,\right]^2. \end{aligned}$$

We write the variance in the above manner because $E(X^2)$ has no closed form solution. However, one can find the closed form solution of $E(X(X-1))$. From Example 4.3, we know that $E(X) = \frac{1}{p}$. Hence, we now focus on finding the second factorial moment of $X$, that is $E(X(X-1))$.

$$\begin{aligned} E(X(X-1)) &= \sum_{x=1}^{\infty} x\,(x-1)\,(1-p)^{x-1}\, p \\ &= \sum_{x=2}^{\infty} x\,(x-1)\,(1-p)\,(1-p)^{x-2}\, p \\ &= \frac{2\, p\,(1-p)}{(1-(1-p))^3} = \frac{2\,(1-p)}{p^2}. \end{aligned}$$

Hence

$$Var(X) = E(X(X-1)) + E(X) - \left[\, E(X)\,\right]^2 = \frac{2\,(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

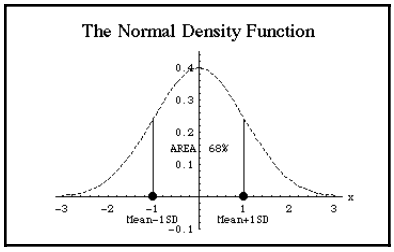
### 4.4. Chebychev Inequality

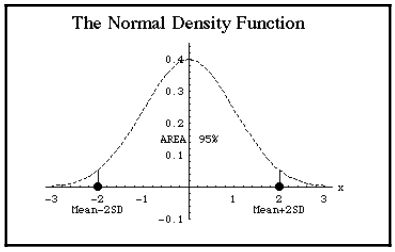
We have taken it for granted, in section 4.2, that the standard deviation (which is the positive square root of the variance) measures the spread of a distribution of a random variable. The spread is measured by the area between "two values". The area under the pdf between two values is the probability of $X$ between the two values. If the standard deviation $\sigma$ measures the spread, then $\sigma$ should control the area between the "two values".

It is well known that if the probability density function is standard normal, that is

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty,$$

then the mean $\mu = 0$ and the standard deviation $\sigma = 1$, and the area between the values $\mu - \sigma$ and $\mu + \sigma$ is 68%.



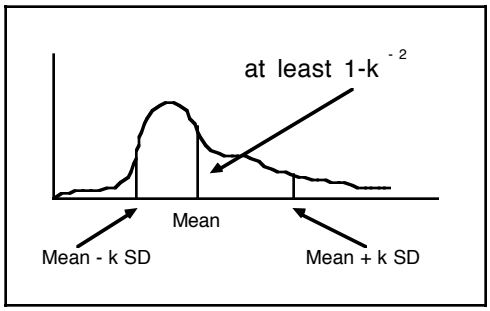




Similarly, the area between the values $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95%. In this way, the standard deviation controls the area between the values $\mu - k\sigma$ and $\mu + k\sigma$ for some $k$ if the distribution is standard normal. If we do not know the probability density function of a random variable, can we find an estimate of the area between the values $\mu - k\sigma$ and $\mu + k\sigma$ for some given $k$? This problem was solved by Chebychev, a well known Russian mathematician. He proved that the area under $f(x)$ on the interval $[\mu - k\sigma,\ \mu + k\sigma]$ is at least $1 - k^{-2}$. This is equivalent to saying the probability that a random variable is within $k$ standard deviations of the mean is at least $1 - k^{-2}$.

**Theorem 4.4 (Chebychev Inequality).** Let $X$ be a random variable with probability density function $f(x)$. If $\mu$ and $\sigma > 0$ are the mean and standard deviation of $X$, then

$$P(|X - \mu| < k\,\sigma) \geq 1 - \frac{1}{k^2}$$

for any nonzero real positive constant $k$.




**Proof:** We assume that the random variable $X$ is continuous. If $X$ is not continuous we replace the integral by summation in the following proof. From the definition of variance, we have the following:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

$$= \int_{-\infty}^{\mu-k\,\sigma} (x-\mu)^2 f(x)\,dx + \int_{\mu-k\,\sigma}^{\mu+k\,\sigma} (x-\mu)^2 f(x)\,dx$$

$$+ \int_{\mu+k\,\sigma}^{\infty} (x-\mu)^2 f(x)\,dx.$$

Since, $\int_{\mu-k\,\sigma}^{\mu+k\,\sigma} (x-\mu)^2 f(x)\,dx$ is positive, we get from the above

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\,\sigma} (x-\mu)^2 f(x)\,dx + \int_{\mu+k\,\sigma}^{\infty} (x-\mu)^2 f(x)\,dx. \tag{4.1}$$

If $x \in (-\infty,\, \mu - k\,\sigma)$, then

$$x \leq \mu - k\,\sigma.$$

Hence

$$k\,\sigma \leq \mu - x$$

for

$$k^2\,\sigma^2 \leq (\mu - x)^2.$$

That is $(\mu - x)^2 \geq k^2\,\sigma^2$. Similarly, if $x \in (\mu + k\,\sigma,\, \infty)$, then

$$x \geq \mu + k\,\sigma$$

Therefore

$$k^2\,\sigma^2 \leq (\mu - x)^2.$$

Thus if $x \notin (\mu - k\,\sigma,\, \mu + k\,\sigma)$, then

$$(\mu - x)^2 \geq k^2\,\sigma^2. \tag{4.2}$$

Using (4.2) and (4.1), we get

$$\sigma^2 \geq k^2\sigma^2 \left[\int_{-\infty}^{\mu - k\,\sigma} f(x)\,dx + \int_{\mu + k\,\sigma}^{\infty} f(x)\,dx\right].$$

Hence

$$\frac{1}{k^2} \geq \left[\int_{-\infty}^{\mu - k\,\sigma} f(x)\,dx + \int_{\mu + k\,\sigma}^{\infty} f(x)\,dx\right].$$

Therefore

$$\frac{1}{k^2} \geq P(X \leq \mu - k\,\sigma) + P(X \geq \mu + k\,\sigma).$$

Thus

$$\frac{1}{k^2} \geq P(|X - \mu| \geq k\,\sigma)$$

which is

$$P(|X - \mu| < k\,\sigma) \geq 1 - \frac{1}{k^2}.$$

This completes the proof of this theorem.

The following integration formula

$$\int_0^1 x^n\,(1 - x)^m\,dx = \frac{n!\,m!}{(n + m + 1)!}$$

will be used in the next example. In this formula $m$ and $n$ represent any two positive integers.

**Example 4.11.** Let the probability density function of a random variable $X$ be

$$f(x) = \begin{cases} 630\,x^4\,(1 - x)^4 & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the exact value of $P(|X - \mu| \leq 2\,\sigma)$? What is the approximate value of $P(|X - \mu| \leq 2\,\sigma)$ when one uses the Chebychev inequality?

**Answer:** First, we find the mean and variance of the above distribution. The mean of $X$ is given by

$$\begin{aligned} E(X) &= \int_0^1 x\, f(x)\, dx \\ &= \int_0^1 630\, x^5\, (1-x)^4\, dx \\ &= 630\, \frac{5!\, 4!}{(5+4+1)!} \\ &= 630\, \frac{5!\, 4!}{10!} \\ &= 630\, \frac{2880}{3628800} \\ &= \frac{630}{1260} \\ &= \frac{1}{2}. \end{aligned}$$

Similarly, the variance of $X$ can be computed from

$$\begin{aligned} Var(X) &= \int_0^1 x^2\, f(x)\, dx - \mu_X^2 \\ &= \int_0^1 630\, x^6\, (1-x)^4\, dx - \frac{1}{4} \\ &= 630\, \frac{6!\, 4!}{(6+4+1)!} - \frac{1}{4} \\ &= 630\, \frac{6!\, 4!}{11!} - \frac{1}{4} \\ &= 630\, \frac{6}{22} - \frac{1}{4} \\ &= \frac{12}{44} - \frac{11}{44} \\ &= \frac{1}{44}. \end{aligned}$$

Therefore, the standard deviation of $X$ is

$$\sigma = \sqrt{\frac{1}{44}} = 0.15.$$

Thus

$$P(|X-\mu| \leq 2\,\sigma) = P(|X-0.5| \leq 0.3)$$

$$= P(-0.3 \leq X - 0.5 \leq 0.3)$$

$$= P(0.2 \leq X \leq 0.8)$$

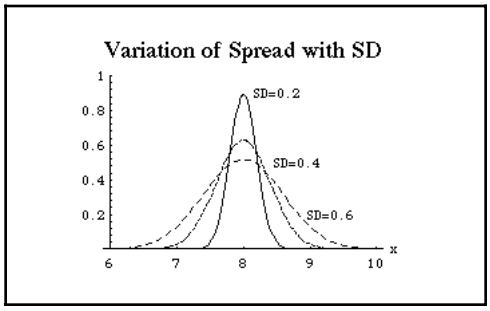
$$= \int_{0.2}^{0.8} 630\, x^4\,(1-x)^4\, dx$$

$$= 0.96.$$

If we use the Chebychev inequality, then we get an approximation of the exact value we have. This approximate value is

$$P(|X-\mu| \leq 2\,\sigma) \geq 1 - \frac{1}{4} = 0.75$$

Hence, Chebychev inequality tells us that if we do not know the distribution of $X$, then $P(|X-\mu| \leq 2\,\sigma)$ is at least 0.75.




Lower the standard deviation, and the smaller is the spread of the distribution. If the standard deviation is zero, then the distribution has no spread. This means that the distribution is concentrated at a single point. In the literature, such distributions are called degenerate distributions. The above figure shows how the spread decreases with the decrease of the standard deviation.

### 4.5. Moment Generating Functions

We have seen in Section 3 that there are some distributions, such as geometric, whose moments are difficult to compute from the definition. A

moment generating function is a real valued function from which one can generate all the moments of a given random variable. In many cases, it is easier to compute various moments of $X$ using the moment generating function.

**Definition 4.5.** Let $X$ be a random variable whose probability density function is $f(x)$. A real valued function $M : \mathbb{R} \to \mathbb{R}$ defined by

$$M(t) = E\left(e^{t\,X}\right)$$

is called the moment generating function of $X$ if this expected value exists for all $t$ in the interval $-h < t < h$ for some $h > 0$.

In general, not every random variable has a moment generating function. But if the moment generating function of a random variable exists, then it is unique. At the end of this section, we will give an example of a random variable which does not have a moment generating function.

Using the definition of expected value of a random variable, we obtain the explicit representation for $M(t)$ as

$$M(t) = \begin{cases} \sum\limits_{x \in R_X} e^{t\,x}\, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{t\,x}\, f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

**Example 4.12.** Let $X$ be a random variable whose moment generating function is $M(t)$ and $n$ be any natural number. What is the $n^{\text{th}}$ derivative of $M(t)$ at $t = 0$?

**Answer:**

$$\begin{aligned} \frac{d}{dt}M(t) &= \frac{d}{dt}E\left(e^{t\,X}\right) \\ &= E\left(\frac{d}{dt}e^{t\,X}\right) \\ &= E\left(X\,e^{t\,X}\right). \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{d^2}{dt^2}M(t) &= \frac{d^2}{dt^2}E\left(e^{t\,X}\right) \\ &= E\left(\frac{d^2}{dt^2}e^{t\,X}\right) \\ &= E\left(X^2\,e^{t\,X}\right). \end{aligned}$$

Hence, in general we get

$$\begin{aligned}\frac{d^n}{dt^n}M(t) &= \frac{d^n}{dt^n}E\left(e^{t\,X}\right)\\ &= E\left(\frac{d^n}{dt^n}e^{t\,X}\right)\\ &= E\left(X^n\,e^{t\,X}\right).\end{aligned}$$

If we set $t=0$ in the $n^{\text{th}}$ derivative, we get

$$\left.\frac{d^n}{dt^n}M(t)\right|_{t=0} = E\left(X^n\,e^{t\,X}\right)\big|_{t=0} = E\left(X^n\right).$$

Hence the $n^{\text{th}}$ derivative of the moment generating function of $X$ evaluated at $t=0$ is the $n^{\text{th}}$ moment of $X$ about the origin.

This example tells us if we know the moment generating function of a random variable; then we can generate all the moments of $X$ by taking derivatives of the moment generating function and then evaluating them at zero.

**Example 4.13.** What is the moment generating function of the random variable $X$ whose probability density function is given by

$$f(x) = \begin{cases} e^{-x} & \text{for } x>0\\ \\ 0 & \text{otherwise?}\end{cases}$$

What are the mean and variance of $X$?

**Answer:** The moment generating function of $X$ is

$$\begin{aligned}M(t) &= E\left(e^{t\,X}\right)\\ &= \int_0^\infty e^{t\,x}\,f(x)\,dx\\ &= \int_0^\infty e^{t\,x}\,e^{-x}\,dx\\ &= \int_0^\infty e^{-(1-t)\,x}\,dx\\ &= \frac{1}{1-t}\,\left[-e^{-(1-t)\,x}\right]_0^\infty\\ &= \frac{1}{1-t} \qquad \text{if } 1-t>0.\end{aligned}$$

The expected value of $X$ can be computed from $M(t)$ as

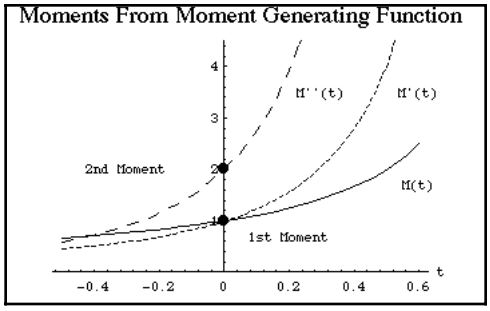
$$\begin{aligned} E(X) &= \frac{d}{dt}M(t)\Big|_{t=0} \\ &= \frac{d}{dt}(1-t)^{-1}\Big|_{t=0} \\ &= (1-t)^{-2}\Big|_{t=0} \\ &= \frac{1}{(1-t)^2}\Big|_{t=0} \\ &= 1. \end{aligned}$$

Similarly

$$\begin{aligned} E(X^2) &= \frac{d^2}{dt^2}M(t)\Big|_{t=0} \\ &= \frac{d^2}{dt^2}(1-t)^{-1}\Big|_{t=0} \\ &= 2\,(1-t)^{-3}\Big|_{t=0} \\ &= \frac{2}{(1-t)^3}\Big|_{t=0} \\ &= 2. \end{aligned}$$

Therefore, the variance of $X$ is

$$Var(X) = E(X^2) - (\mu)^2 = 2 - 1 = 1.$$




**Example 4.14.** Let $X$ have the probability density function

$$f(x) = \begin{cases} \frac{1}{9}\left(\frac{8}{9}\right)^x & \text{for } x = 0, 1, 2, ..., \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is the moment generating function of the random variable $X$?

**Answer:**

$$\begin{aligned} M(t) &= E\left(e^{t\,X}\right) \\ &= \sum_{x=0}^{\infty} e^{t\,x}\, f(x) \\ &= \sum_{x=0}^{\infty} e^{t\,x} \left(\frac{1}{9}\right) \left(\frac{8}{9}\right)^{x} \\ &= \left(\frac{1}{9}\right) \sum_{x=0}^{\infty} \left(e^{t}\,\frac{8}{9}\right)^{x} \\ &= \left(\frac{1}{9}\right) \frac{1}{1 - e^{t}\,\frac{8}{9}} \qquad \text{if} \quad e^{t}\,\frac{8}{9} < 1 \\ &= \frac{1}{9 - 8\,e^{t}} \qquad \text{if} \quad t < \ln\left(\frac{9}{8}\right). \end{aligned}$$

**Example 4.15.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} b\,e^{-b\,x} & \text{for } x > 0 \\ 0 & \text{otherwise}\,, \end{cases}$$

where $b > 0$. If $M(t)$ is the moment generating function of $X$, then what is $M(-6\,b)$?

**Answer:**

$$\begin{aligned} M(t) &= E\left(e^{t\,X}\right) \\ &= \int_0^{\infty} b\,e^{t\,x}\,e^{-b\,x}\,dx \\ &= b\int_0^{\infty} e^{-(b-t)\,x}\,dx \\ &= \frac{b}{b-t}\left[-e^{-(b-t)\,x}\right]_0^{\infty} \\ &= \frac{b}{b-t} \qquad \text{if} \quad b - t > 0. \end{aligned}$$

Hence $M(-6\,b) = \frac{b}{7b} = \frac{1}{7}$.

**Example 4.16.** Let the random variable $X$ have moment generating function $M(t) = (1-t)^{-2}$ for $t < 1$. What is the third moment of $X$ about the origin?

**Answer:** To compute the third moment $E(X^3)$ of $X$ about the origin, we

need to compute the third derivative of $M(t)$ at $t = 0$.

$$\begin{aligned} M(t) &= (1-t)^{-2} \\ M'(t) &= 2\,(1-t)^{-3} \\ M''(t) &= 6\,(1-t)^{-4} \\ M'''(t) &= 24\,(1-t)^{-5}. \end{aligned}$$

Thus the third moment of $X$ is given by

$$E\left(X^3\right) = \frac{24}{(1-0)^5} = 24.$$

**Theorem 4.5.** Let $M(t)$ be the moment generating function of the random variable $X$. If

$$M(t) = a_0 + a_1\,t + a_2\,t^2 + \cdots + a_n\,t^n + \cdots \tag{4.3}$$

is the Taylor series expansion of $M(t)$, then

$$E\left(X^n\right) = (n!)\;a_n$$

for all natural number $n$.

**Proof:** Let $M(t)$ be the moment generating function of the random variable $X$. The Taylor series expansion of $M(t)$ about 0 is given by

$$M(t) = M(0) + \frac{M'(0)}{1!}\,t + \frac{M''(0)}{2!}\,t^2 + \frac{M'''(0)}{3!}\,t^3 + \cdots + \frac{M^{(n)}(0)}{n!}\,t^n + \cdots$$

Since $E(X^n) = M^{(n)}(0)$ for $n \geq 1$ and $M(0) = 1$, we have

$$M(t) = 1 + \frac{E(X)}{1!}\,t + \frac{E(X^2)}{2!}\,t^2 + \frac{E(X^3)}{3!}\,t^3 + \cdots + \frac{E(X^n)}{n!}\,t^n + \cdots \tag{4.4}$$

From (4.3) and (4.4), equating the coefficients of the like powers of $t$, we obtain

$$a_n = \frac{E\left(X^n\right)}{n!}$$

which is

$$E\left(X^n\right) = (n!)\;a_n.$$

This proves the theorem.

**Example 4.17.** What is the $479^{\text{th}}$ moment of $X$ about the origin, if the moment generating function of $X$ is $\frac{1}{1+t}$?

**Answer** The Taylor series expansion of $M(t) = \frac{1}{1+t}$ can be obtained by using long division (a technique we have learned in high school).

$$\begin{aligned} M(t) &= \frac{1}{1+t} \\ &= \frac{1}{1-(-t)} \\ &= 1 + (-t) + (-t)^2 + (-t)^3 + \cdots + (-t)^n + \cdots \\ &= 1 - t + t^2 - t^3 + t^4 + \cdots + (-1)^n t^n + \cdots \end{aligned}$$

Therefore $a_n = (-1)^n$ and from this we obtain $a_{479} = -1$. By Theorem 4.5,

$$E\left(X^{479}\right) = (479!)\, a_{479} = -\,479!$$

**Example 4.18.** If the moment generating of a random variable $X$ is

$$M(t) = \sum_{j=0}^{\infty} \frac{e^{(t\,j-1)}}{j!},$$

then what is the probability of the event $X = 2$?

**Answer:** By examining the given moment generating function of $X$, it is easy to note that $X$ is a discrete random variable with space $R_X = \{0, 1, 2, \cdots, \infty\}$. Hence by definition, the moment generating function of $X$ is

$$M(t) = \sum_{j=0}^{\infty} e^{t\,j}\, f(j). \tag{4.5}$$

But we are given that

$$\begin{aligned} M(t) &= \sum_{j=0}^{\infty} \frac{e^{(t\,j-1)}}{j!} \\ &= \sum_{j=0}^{\infty} \frac{e^{-1}}{j!}\, e^{t\,j}. \end{aligned}$$

From (4.5) and the above, equating the coefficients of $e^{tj}$, we get

$$f(j) = \frac{e^{-1}}{j!} \qquad \text{for} \qquad j = 0, 1, 2, ..., \infty.$$

Thus, the probability of the event $X = 2$ is given by

$$P(X = 2) = f(2) = \frac{e^{-1}}{2!} = \frac{1}{2\,e}.$$

**Example 4.19.** Let $X$ be a random variable with

$$E\,(X^n) = 0.8 \qquad \text{for} \quad n = 1, 2, 3, ..., \infty.$$

What are the moment generating function and probability density function of $X$?

**Answer:**

$$\begin{aligned} M(t) &= M(0) + \sum_{n=1}^{\infty} M^{(n)}(0)\,\left(\frac{t^n}{n!}\right) \\ &= M(0) + \sum_{n=1}^{\infty} E\,(X^n)\,\left(\frac{t^n}{n!}\right) \\ &= 1 + 0.8\,\sum_{n=1}^{\infty}\left(\frac{t^n}{n!}\right) \\ &= 0.2 + 0.8 + 0.8\,\sum_{n=1}^{\infty}\left(\frac{t^n}{n!}\right) \\ &= 0.2 + 0.8\,\sum_{n=0}^{\infty}\left(\frac{t^n}{n!}\right) \\ &= 0.2\,e^{0\,t} + 0.8\,e^{1\,t}. \end{aligned}$$

Therefore, we get $f(0) = P(X = 0) = 0.2$ and $f(1) = P(X = 1) = 0.8$. Hence the moment generating function of $X$ is

$$M(t) = 0.2 + 0.8\,e^{t},$$

and the probability density function of $X$ is

$$f(x) = \begin{cases} |x - 0.2| & \text{for } x = 0, 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

**Example 4.20.** If the moment generating function of a random variable $X$ is given by

$$M(t) = \frac{5}{15}\,e^{t} + \frac{4}{15}\,e^{2\,t} + \frac{3}{15}\,e^{3\,t} + \frac{2}{15}\,e^{4\,t} + \frac{1}{15}\,e^{5\,t},$$

then what is the probability density function of $X$? What is the space of the random variable $X$?

**Answer:** The moment generating function of $X$ is given to be

$$M(t) = \frac{5}{15}\, e^{t} + \frac{4}{15}\, e^{2\,t} + \frac{3}{15}\, e^{3\,t} + \frac{2}{15}\, e^{4\,t} + \frac{1}{15}\, e^{5\,t}.$$

This suggests that $X$ is a discrete random variable. Since $X$ is a discrete random variable, by definition of the moment generating function, we see that

$$\begin{aligned} M(t) &= \sum_{x \in R_X} e^{t\,x}\, f(x) \\ &= e^{t\,x_1}\, f(x_1) + e^{t\,x_2}\, f(x_2) + e^{t\,x_3}\, f(x_3) + e^{t\,x_4}\, f(x_4) + e^{t\,x_5}\, f(x_5). \end{aligned}$$

Hence we have

$$\begin{aligned} f(x_1) = f(1) &= \frac{5}{15} \\ f(x_2) = f(2) &= \frac{4}{15} \\ f(x_3) = f(3) &= \frac{3}{15} \\ f(x_4) = f(4) &= \frac{2}{15} \\ f(x_5) = f(5) &= \frac{1}{15}. \end{aligned}$$

Therefore the probability density function of $X$ is given by

$$f(x) = \frac{6 - x}{15} \qquad \text{for} \qquad x = 1, 2, 3, 4, 5$$

and the space of the random variable $X$ is

$$R_X = \{1, 2, 3, 4, 5\}.$$

**Example 4.21.** If the probability density function of a discrete random variable $X$ is

$$f(x) = \frac{6}{\pi^2\, x^2}, \qquad \text{for} \qquad x = 1, 2, 3, ..., \infty,$$

then what is the moment generating function of $X$?

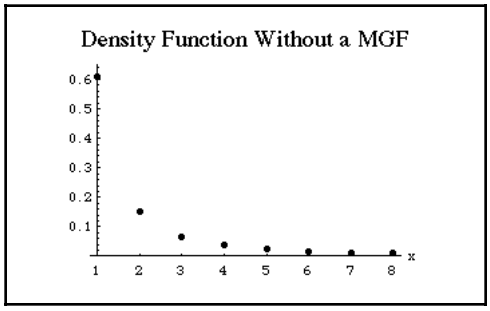
**Answer:** If the moment generating function of $X$ exists, then

$$\begin{aligned} M(t) &= \sum_{x=1}^{\infty} e^{tx} f(x) \\ &= \sum_{x=1}^{\infty} e^{tx} \left(\frac{\sqrt{6}}{\pi\, x}\right)^2 \\ &= \sum_{x=1}^{\infty} \left(\frac{e^{tx}\, 6}{\pi^2\, x^2}\right) \\ &= \frac{6}{\pi^2} \sum_{x=1}^{\infty} \frac{e^{tx}}{x^2}. \end{aligned}$$

Now we show that the above infinite series diverges if $t$ belongs to the interval $(-h,\, h)$ for any $h > 0$. To prove that this series is divergent, we do the ratio test, that is

$$\begin{aligned} \lim_{n\to\infty} \left(\frac{a_{n+1}}{a_n}\right) &= \lim_{n\to\infty} \left(\frac{e^{t\,(n+1)}}{(n+1)^2}\, \frac{n^2}{e^{t\,n}}\right) \\ &= \lim_{n\to\infty} \left(\frac{e^{t\,n}\, e^t}{(n+1)^2}\, \frac{n^2}{e^{t\,n}}\right) \\ &= \lim_{n\to\infty} \left(e^t \left(\frac{n}{n+1}\right)^2\right) \\ &= e^t. \end{aligned}$$

For any $h > 0$, since $e^t$ is not always less than 1 for all $t$ in the interval $(-h,\, h)$, we conclude that the above infinite series diverges and hence for this random variable $X$ the moment generating function does not exist.




Notice that for the above random variable, $E\,[X^n]$ does not exist for any natural number $n$. Hence the discrete random variable $X$ in Example 4.21 has no moments. Similarly, the continuous random variable $X$ whose

probability density function is

$$f(x) = \begin{cases} \frac{1}{x^2} & \text{for } 1 \leq x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

has no moment generating function and no moments.

In the following theorem we summarize some important properties of the moment generating function of a random variable.

**Theorem 4.6.** Let $X$ be a random variable with the moment generating function $M_X(t)$. If $a$ and $b$ are any two real constants, then

$$M_{X+a}(t) = e^{a\,t}\, M_X(t) \tag{4.6}$$

$$M_{b\,X}(t) = M_X(b\,t) \tag{4.7}$$

$$M_{\frac{X+a}{b}}(t) = e^{\frac{a}{b}t}\, M_X\left(\frac{t}{b}\right). \tag{4.8}$$

**Proof:** First, we prove (4.6).

$$\begin{aligned} M_{X+a}(t) &= E\left(e^{t\,(X+a)}\right) \\ &= E\left(e^{t\,X+t\,a}\right) \\ &= E\left(e^{t\,X}\, e^{t\,a}\right) \\ &= e^{t\,a}\, E\left(e^{t\,X}\right) \\ &= e^{t\,a}\, M_X(t). \end{aligned}$$

Similarly, we prove (4.7).

$$\begin{aligned} M_{b\,X}(t) &= E\left(e^{t\,(b\,X)}\right) \\ &= E\left(e^{(t\,b)\,X}\right) \\ &= M_X(t\,b). \end{aligned}$$

By using (4.6) and (4.7), we easily get (4.8).

$$\begin{aligned} M_{\frac{X+a}{b}}(t) &= M_{\frac{X}{b}+\frac{a}{b}}(t) \\ &= e^{\frac{a}{b}\,t}\, M_{\frac{X}{b}}(t) \\ &= e^{\frac{a}{b}\,t}\, M_X\left(\frac{t}{b}\right). \end{aligned}$$

This completes the proof of this theorem.

**Definition 4.6.** The $n^{\text{th}}$ factorial moment of a random variable $X$ is $E(X(X-1)(X-2)\cdots(X-n+1))$.

**Definition 4.7.** The factorial moment generating function (FMGF) of $X$ is denoted by $G(t)$ and defined as

$$G(t) = E\left(t^X\right).$$

It is not difficult to establish a relationship between the moment generating function (MGF) and the factorial moment generating function (FMGF). The relationship between them is the following:

$$G(t) = E\left(t^X\right) = E\left(e^{\ln t^X}\right) = E\left(e^{X\,\ln t}\right) = M(\ln t).$$

Thus, if we know the MGF of a random variable, we can determine its FMGF and conversely.

**Definition 4.8.** Let $X$ be a random variable. The characteristic function $\phi(t)$ of $X$ is defined as

$$\begin{aligned}\phi(t) &= E\left(e^{i\,t\,X}\right)\\ &= E\left(\,cos(tX) + i\,sin(tX)\,\right)\\ &= E\left(\,cos(tX)\,\right) + i\,E\left(\,sin(tX)\,\right).\end{aligned}$$

The probability density function can be recovered from the characteristic function by using the following formula

$$f(x) = \frac{1}{2\,\pi}\int_{-\infty}^{\infty} e^{-i\,t\,x}\,\phi(t)\,dt.$$

Unlike the moment generating function, the characteristic function of a random variable always exists. For example, the Cauchy random variable $X$ with probability density $f(x) = \frac{1}{\pi(1+x^2)}$ has no moment generating function. However, the characteristic function is

$$\begin{aligned}\phi(t) &= E\left(e^{i\,t\,X}\right)\\ &= \int_{-\infty}^{\infty}\frac{e^{itx}}{\pi(1+x^2)}\,dx\\ &= e^{-|t|}.\end{aligned}$$

To evaluate the above integral one needs the theory of residues from the complex analysis.

The characteristic function $\phi(t)$ satisfies the same set of properties as the moment generating functions as given in Theorem 4.6.

The following integrals

$$\int_0^\infty x^m\, e^{-x}\, dx = m! \qquad \text{if } m \text{ is a positive integer}$$

and

$$\int_0^\infty \sqrt{x}\, e^{-x}\, dx = \frac{\sqrt{\pi}}{2}$$

are needed for some problems in the Review Exercises of this chapter. These formulas will be discussed in Chapter 6 while we describe the properties and usefulness of the gamma distribution.

We end this chapter with the following comment about the Taylor's series. Taylor's series was discovered to mimic the decimal expansion of real numbers. For example

$$125 = 1\,(10)^2 + 2\,(10)^1 + 5\,(10)^0$$

is an expansion of the number 125 with respect to base 10. Similarly,

$$125 = 1\,(9)^2 + 4\,(9)^1 + 8\,(9)^0$$

is an expansion of the number 125 in base 9 and it is 148. Since given a function $f : \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}$, $f(x)$ is a real number and it can be expanded with respect to the base $x$. The expansion of $f(x)$ with respect to base $x$ will have a form

$$f(x) = a_0x^0 + a_1x^1 + a_2x^2 + a_3x^3 + \cdots$$

which is

$$f(x) = \sum_{k=0}^{\infty} a_k x^k.$$

If we know the coefficients $a_k$ for $k = 0, 1, 2, 3, ...$, then we will have the expansion of $f(x)$ in base $x$. Taylor found the remarkable fact that the the coefficients $a_k$ can be computed if $f(x)$ is sufficiently differentiable. He proved that for $k = 1, 2, 3, ...$

$$a_k = \frac{f^{(k)}(0)}{k!} \qquad \text{with} \quad f^{(0)} = f(0).$$

### 4.6. Review Exercises

**1.** In a state lottery a five-digit integer is selected at random. If a player bets 1 dollar on a particular number, the payoff (if that number is selected) is \$500 minus the \$1 paid for the ticket. Let $X$ equal the payoff to the better. Find the expected value of $X$.

**2.** A discrete random variable $X$ has probability density function of the form

$$f(x) = \begin{cases} c\,(8-x) & \text{for } x = 0,1,2,3,4,5 \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the constant $c$. (b) Find $P(X > 2)$. (c) Find the expected value $E(X)$ for the random variable $X$.

**3.** A random variable $X$ has a cumulative distribution function

$$F(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 < x \leq 1 \\ \\ x - \frac{1}{2} & \text{if } 1 < x \leq \frac{3}{2}\ . \end{cases}$$

(a) Graph $F(x)$. (b) Graph $f(x)$. (c) Find $P(X \leq 0.5)$. (d) Find $P(X \geq 0.5)$. (e) Find $P(X \leq 1.25)$. (f) Find $P(X = 1.25)$.

**4.** Let $X$ be a random variable with probability density function

$$f(x) = \begin{cases} \frac{1}{8}x & \text{for } \ x = 1,2,5 \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the expected value of $X$. (b) Find the variance of $X$. (c) Find the expected value of $2X + 3$. (d) Find the variance of $2X + 3$. (e) Find the expected value of $3X - 5X^2 + 1$.

**5.** The measured radius of a circle, $R$, has probability density function

$$f(r) = \begin{cases} 6\,r\,(1-r) & \text{if } 0 < r < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the expected value of the radius. (b) Find the expected circumference. (c) Find the expected area.

**6.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} \theta\,x + \frac{3}{2}\,\theta^{\frac{3}{2}}\,x^2 & \text{for } 0 < x < \frac{1}{\sqrt{\theta}} \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. What is the expected value of $X$?

**7.** Suppose $X$ is a random variable with mean $\mu$ and variance $\sigma^2 > 0$. For what value of $a$, where $a > 0$ is $E\left(\left[a\,X - \frac{1}{a}\right]^2\right)$ minimized?

**8.** A rectangle is to be constructed having dimension $X$ by $2X$, where $X$ is a random variable with probability density function

$$f(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x < 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected area of the rectangle?

**9.** A box is to be constructed so that the height is 10 inches and its base is $X$ inches by $X$ inches. If $X$ has a uniform distribution over the interval $[2, 8]$, then what is the expected volume of the box in cubic inches?

**10.** If $X$ is a random variable with density function

$$f(x) = \begin{cases} 1.4\,e^{-2x} + 0.9\,e^{-3x} & \text{for } x > 0 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

then what is the expected value of $X$?

**11.** A fair coin is tossed. If a head occurs, 1 die is rolled; if a tail occurs, 2 dice are rolled. Let $X$ be the total on the die or dice. What is the expected value of $X$?

**12.** If velocities of the molecules of a gas have the probability density (Maxwell's law)

$$f(v) = \begin{cases} a\,v^2 e^{-h^2\,v^2} & \text{for } v \geq 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what are the expectation and the variance of the velocity of the molecules and also the magnitude of $a$ for some given $h$?

**13.** A couple decides to have children until they get a girl, but they agree to stop with a maximum of 3 children even if they haven't gotten a girl. If $X$ and $Y$ denote the number of children and number of girls, respectively, then what are $E(X)$ and $E(Y)$?

**14.** In roulette, a wheel stops with equal probability at any of the 38 numbers 0, 00, 1, 2, ..., 36. If you bet \$1 on a number, then you win \$36 (net gain is

$35) if the number comes up; otherwise, you lose your dollar. What are your expected winnings?

**15.** If the moment generating function for the random variable $X$ is $M_X(t) = \frac{1}{1+t}$, what is the third moment of $X$ about the point $x = 2$?

**16.** If the mean and the variance of a certain distribution are 2 and 8, what are the first three terms in the series expansion of the moment generating function?

**17.** Let $X$ be a random variable with density function

$$f(x) = \begin{cases} a\, e^{-ax} & \text{for } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $a > 0$. If $M(t)$ denotes the moment generating function of $X$, what is $M(-3a)$?

**18.** Suppose the random variable $X$ has moment generating

$$M(t) = \frac{1}{(1-\beta t)^k}, \qquad \text{for} \quad t < \frac{1}{\beta}.$$

What is the $n^{\text{th}}$ moment of $X$?

**19.** Two balls are dropped in such a way that each ball is equally likely to fall into any one of four holes. Both balls may fall into the same hole. Let $X$ denote the number of unoccupied holes at the end of the experiment. What is the moment generating function of $X$?

**20.** If the moment generating function of $X$ is $M(t) = \frac{1}{(1-t)^2}$ for $t < 1$, then what is the fourth moment of $X$?

**21.** Let the random variable $X$ have the moment generating function

$$M(t) = \frac{e^{3t}}{1-t^2}, \qquad -1 < t < 1.$$

What are the mean and the variance of $X$, respectively?

**22.** Let the random variable $X$ have the moment generating function

$$M(t) = e^{3t+t^2}.$$

What is the second moment of $X$ about $x = 0$?

**23.** Suppose the random variable $X$ has the cumulative density function $F(x)$. Show that the expected value of the random variable $(X - c)^2$ is minimum if $c$ equals the expected value of $X$.

**24.** Suppose the continuous random variable $X$ has the cumulative density function $F(x)$. Show that the expected value of the random variable $|X - c|$ is minimum if $c$ equals the median of $X$ (that is, $F(c) = 0.5$).

**25.** Let the random variable $X$ have the probability density function

$$f(x) = \frac{1}{2}\, e^{-|x|} \qquad -\infty < x < \infty.$$

What are the expected value and the variance of $X$?

**26.** If $M_X(t) = k\ (2 + 3e^t)^4$, what is the value of $k$?

**27.** Given the moment generating function of $X$ as

$$M(t) = 1 + t + 4t^2 + 10t^3 + 14t^4 + \cdots$$

what is the third moment of $X$ about its mean?

**28.** A set of measurements $X$ has a mean of 7 and standard deviation of 0.2. For simplicity, a linear transformation $Y = aX + b$ is to be applied to make the mean and variance both equal to 1. What are the values of the constants $a$ and $b$?

**29.** A fair coin is to be tossed 3 times. The player receives 10 dollars if all three turn up heads and pays 3 dollars if there is one or no heads. No gain or loss is incurred otherwise. If $Y$ is the gain of the player, what the expected value of $Y$?

**30.** If $X$ has the probability density function

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

then what is the expected value of the random variable $Y = e^{\frac{3}{4}X} + 6$?

**31.** If the probability density function of the random variable $X$ if

$$f(x) = \begin{cases} (1-p)^{x-1}\, p & \text{if } x = 1, 2, 3, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the expected value of the random variable $X^{-1}$ ?

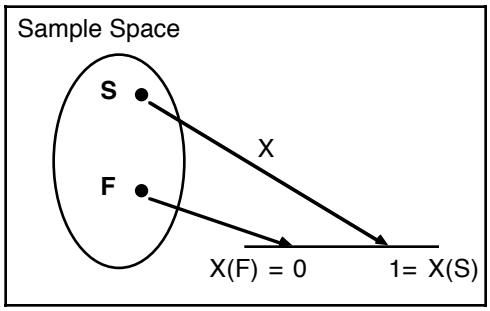
# Chapter 5
# SOME SPECIAL DISCRETE DISTRIBUTIONS

Given a random experiment, we can find the set of all possible outcomes which is known as the sample space. Objects in a sample space may not be numbers. Thus, we use the notion of random variable to quantify the qualitative elements of the sample space. A random variable is characterized by either its probability density function or its cumulative distribution function. The other characteristics of a random variable are its mean, variance and moment generating function. In this chapter, we explore some frequently encountered discrete distributions and study their important characteristics.

## 5.1. Bernoulli Distribution

A Bernoulli trial is a random experiment in which there are precisely two possible outcomes, which we conveniently call 'failure' (F) and 'success' (S). We can define a random variable from the sample space $\{S, F\}$ into the set of real numbers as follows:

$$X(F) = 0 \qquad X(S) = 1.$$




The probability density function of this random variable is

$$f(0) = P(X = 0) = 1 - p$$
$$f(1) = P(X = 1) = p,$$

where $p$ denotes the probability of success. Hence

$$f(x) = p^x (1-p)^{1-x}, \qquad x = 0, 1.$$

**Definition 5.1.** The random variable $X$ is called the Bernoulli random variable if its probability density function is of the form

$$f(x) = p^x (1-p)^{1-x}, \qquad x = 0, 1$$

where $p$ is the probability of success.

We denote the Bernoulli random variable by writing $X \sim BER(p)$.

**Example 5.1.** What is the probability of getting a score of not less than 5 in a throw of a six-sided die?

**Answer:** Although there are six possible scores {1, 2, 3, 4, 5, 6}, we are grouping them into two sets, namely {1, 2, 3, 4} and {5, 6}. Any score in {1, 2, 3, 4} is a failure and any score in {5, 6} is a success. Thus, this is a Bernoulli trial with

$$P(X = 0) = P(\text{failure}) = \frac{4}{6} \qquad \text{and} \qquad P(X = 1) = P(\text{success}) = \frac{2}{6}.$$

Hence, the probability of getting a score of not less than 5 in a throw of a six-sided die is $\frac{2}{6}$.

**Theorem 5.1.** If $X$ is a Bernoulli random variable with parameter $p$, then the mean, variance and moment generating functions are respectively given by

$$\begin{aligned} \mu_X &= p \\ \sigma_X^2 &= p\,(1-p) \\ M_X(t) &= (1-p) + p\,e^t. \end{aligned}$$

**Proof:** The mean of the Bernoulli random variable is

$$\begin{aligned} \mu_X &= \sum_{x=0}^{1} x\,f(x) \\ &= \sum_{x=0}^{1} x\,p^x\,(1-p)^{1-x} \\ &= p. \end{aligned}$$

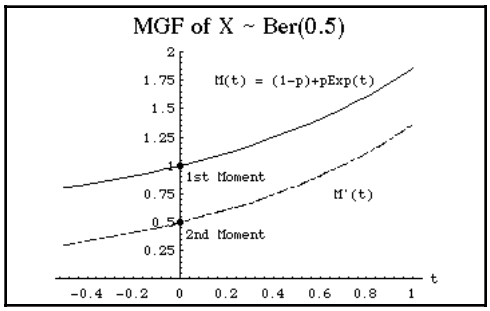
Similarly, the variance of $X$ is given by

$$\begin{aligned} \sigma_X^2 &= \sum_{x=0}^{1} (x-\mu_X)^2\,f(x) \\ &= \sum_{x=0}^{1} (x-p)^2\,p^x\,(1-p)^{1-x} \\ &= p^2\,(1-p) + p\,(1-p)^2 \\ &= p\,(1-p)\,[p + (1-p)] \\ &= p\,(1-p). \end{aligned}$$

Next, we find the moment generating function of the Bernoulli random variable

$$\begin{aligned} M(t) &= E\left(e^{tX}\right) \\ &= \sum_{x=0}^{1} e^{tx}\,p^x\,(1-p)^{1-x} \\ &= (1-p) + e^t\,p. \end{aligned}$$

This completes the proof. The moment generating function of $X$ and all the moments of $X$ are shown below for $p = 0.5$. Note that for the Bernoulli distribution all its moments about zero are same and equal to $p$.




## 5.2. Binomial Distribution

Consider a fixed number $n$ of mutually independent Bernoulli trails. Suppose these trials have same probability of success, say $p$. A random variable $X$ is called a binomial random variable if it represents the total number of successes in $n$ independent Bernoulli trials.

Now we determine the probability density function of a binomial random variable. Recall that the probability density function of $X$ is defined as

$$f(x) = P(X = x).$$

Thus, to find the probability density function of $X$ we have to find the probability of $x$ successes in $n$ independent trails.

If we have $x$ successes in $n$ trails, then the probability of each $n$-tuple with $x$ successes and $n - x$ failures is

$$p^x (1-p)^{n-x}.$$

However, there are $\binom{n}{x}$ tuples with $x$ successes and $n - x$ failures in $n$ trials. Hence

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Therefore, the probability density function of $X$ is

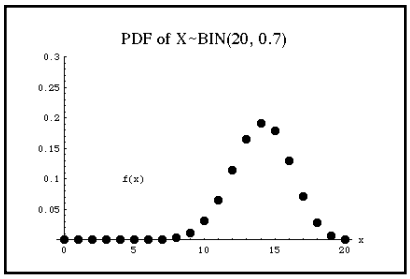
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n.$$

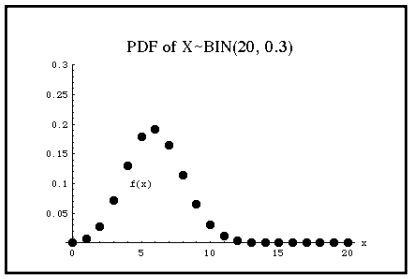
**Definition 5.2.** The random variable $X$ is called the binomial random variable with parameters $p$ and $n$ if its probability density function is of the form

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n$$

where $0 < p < 1$ is the probability of success.

We will denote a binomial random variable with parameters $p$ and $n$ as $X \sim BIN(n, p)$.




**Example 5.2.** Is the real valued function $f(x)$ given by

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n$$

where $n$ and $p$ are parameters, a probability density function?

**Answer:** To answer this question, we have to check that $f(x)$ is nonnegative and $\sum_{x=0}^{n} f(x)$ is 1. It is easy to see that $f(x) \geq 0$. We show that sum is one.

$$\begin{aligned} \sum_{x=0}^{n} f(x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p + 1 - p)^n \\ &= 1. \end{aligned}$$

Hence $f(x)$ is really a probability density function.

**Example 5.3.** On a five-question multiple-choice test there are five possible answers, of which one is correct. If a student guesses randomly and independently, what is the probability that she is correct only on questions 1 and 4?

**Answer:** Here the probability of success is $p = \frac{1}{5}$, and thus $1 - p = \frac{4}{5}$. Therefore, the probability that she is correct on questions 1 and 4 is

$$\begin{aligned} P(\text{correct on questions 1 and 4}) =& p^2(1-p)^3 \\ &= \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^3 \\ &= \frac{64}{5^5} = 0.02048. \end{aligned}$$

**Example 5.4.** On a five-question multiple-choice test there are five possible answers, of which one is correct. If a student guesses randomly and independently, what is the probability that she is correct only on two questions?

**Answer:** Here the probability of success is $p = \frac{1}{5}$, and thus $1-p = \frac{4}{5}$. There are $\binom{5}{2}$ different ways she can be correct on two questions. Therefore, the probability that she is correct on two questions is

$$\begin{aligned} P(\text{correct on two questions}) &= \binom{5}{2} p^2(1-p)^3 \\ &= 10 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^3 \\ &= \frac{640}{5^5} = 0.2048. \end{aligned}$$

**Example 5.5.** What is the probability of rolling two sixes and three nonsixes in 5 independent casts of a fair die?

**Answer:** Let the random variable $X$ denote the number of sixes in 5 independent casts of a fair die. Then $X$ is a binomial random variable with probability of success $p$ and $n = 5$. The probability of getting a six is $p = \frac{1}{6}$. Hence

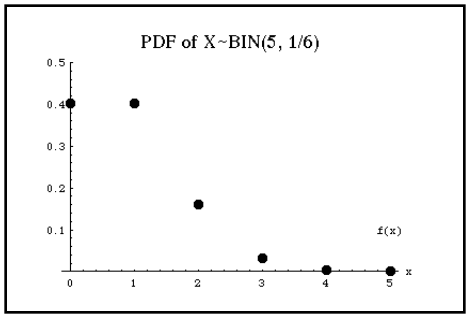
$$\begin{aligned} P(X = 2) = f(2) &= \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\ &= 10 \left(\frac{1}{36}\right) \left(\frac{125}{216}\right) \\ &= \frac{1250}{7776} = 0.160751. \end{aligned}$$

**Example 5.6.** What is the probability of rolling at most two sixes in 5 independent casts of a fair die?

**Answer:** Let the random variable $X$ denote number of sixes in 5 independent casts of a fair die. Then $X$ is a binomial random variable with probability of success $p$ and $n = 5$. The probability of getting a six is $p = \frac{1}{6}$. Hence, the

probability of rolling at most two sixes is

$$\begin{aligned}
P(X \le 2) &= F(2) = f(0) + f(1) + f(2) \\
&= \binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 + \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 + \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 \\
&= \sum_{k=0}^{2}\binom{5}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{5-k} \\
&= \frac{1}{2}\left(0.9421 + 0.9734\right) = 0.9577 \qquad \text{(from binomial table)}
\end{aligned}$$




**Theorem 5.2.** If $X$ is binomial random variable with parameters $p$ and $n$, then the mean, variance and moment generating functions are respectively given by

$$\begin{aligned}
\mu_X &= n\,p \\
\sigma_X^2 &= n\,p\,(1-p) \\
M_X(t) &= \left[(1-p) + p\,e^t\right]^n.
\end{aligned}$$

**Proof:** First, we determine the moment generating function $M(t)$ of $X$ and then we generate mean and variance from $M(t)$.

$$\begin{aligned}
M(t) &= E\left(e^{tX}\right) \\
&= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x\,(1-p)^{n-x} \\
&= \sum_{x=0}^{n}\binom{n}{x}\left(p\,e^t\right)^x\,(1-p)^{n-x} \\
&= \left(p\,e^t + 1 - p\right)^n.
\end{aligned}$$

Hence

$$M'(t) = n\,\left(p\,e^t + 1 - p\right)^{n-1}\,p\,e^t.$$

Therefore

$$\mu_X = M'(0) = n\,p.$$

Similarly

$$M''(t) = n\,\left(p\,e^t + 1 - p\right)^{n-1}\,p\,e^t + n\,(n-1)\,\left(p\,e^t + 1 - p\right)^{n-2}\,\left(p\,e^t\right)^2.$$

Therefore

$$E(X^2) = M''(0) = n\,(n-1)\,p^2 + n\,p.$$

Hence

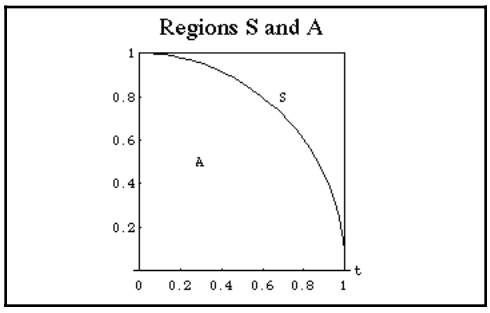
$$Var(X) = E(X^2) - \mu_X^2 = n\,(n-1)\,p^2 + n\,p - n^2\,p^2 = n\,p\,(1-p).$$

This completes the proof.

**Example 5.7.** Suppose that 2000 points are selected independently and at random from the unit squares $S = \{(x, y)\,|\,0 \le x, y \le 1\}$. Let $X$ equal the number of points that fall in $A = \{(x, y)\,|\,x^2 + y^2 < 1\}$. How is $X$ distributed? What are the mean, variance and standard deviation of $X$?

**Answer:** If a point falls in $A$, then it is a success. If a point falls in the complement of $A$, then it is a failure. The probability of success is

$$p = \frac{\text{area of A}}{\text{area of S}} = \frac{1}{4}\,\pi.$$




Since, the random variable represents the number of successes in 2000 independent trials, the random variable $X$ is a binomial with parameters $p = \frac{\pi}{4}$ and $n = 2000$, that is $X \sim BIN(2000,\, \frac{\pi}{4})$.

Hence by Theorem 5.2,

$$\mu_X = 2000\,\frac{\pi}{4} = 1570.8,$$

and

$$\sigma_X^2 = 2000\,\left(1-\frac{\pi}{4}\right)\,\frac{\pi}{4} = 337.1.$$

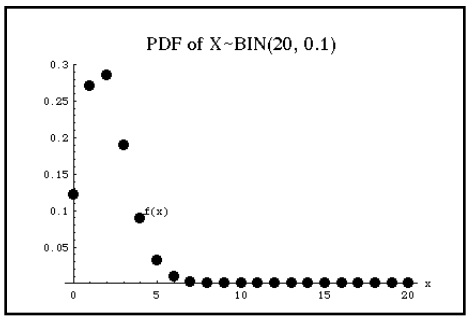
The standard deviation of $X$ is

$$\sigma_X = \sqrt{337.1} = 18.36.$$

**Example 5.8.** Let the probability that the birth weight (in grams) of babies in America is less than 2547 grams be 0.1. If $X$ equals the number of babies that weigh less than 2547 grams at birth among 20 of these babies selected at random, then what is $P(X \leq 3)$?

**Answer:** If a baby weighs less than 2547, then it is a success; otherwise it is a failure. Thus $X$ is a binomial random variable with probability of success $p$ and $n = 20$. We are given that $p = 0.1$. Hence

$$\begin{aligned} P(X \leq 3) &= \sum_{k=0}^{3} \binom{20}{k}\,\left(\frac{1}{10}\right)^k\,\left(\frac{9}{10}\right)^{20-k} \\ &= 0.867 \qquad\qquad \text{(from table)}. \end{aligned}$$

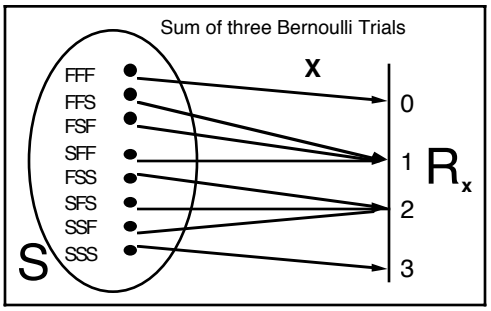



**Example 5.9.** Let $X_1$, $X_2$, $X_3$ be three independent Bernoulli random variables with the same probability of success $p$. What is the probability density function of the random variable $X = X_1 + X_2 + X_3$?

**Answer:** The sample space of the three independent Bernoulli trials is

$$S = \{FFF,\, FFS,\, FSF,\, SFF,\, FSS,\, SFS,\, SSF,\, SSS\}.$$

The random variable $X = X_1 + X_2 + X_3$ represents the number of successes in each element of $S$. The following diagram illustrates this.




Let $p$ be the probability of success. Then

$$f(0) = P(X = 0) = P(FFF) = (1-p)^3$$
$$f(1) = P(X = 1) = P(FFS) + P(FSF) + P(SFF) = 3\,p\,(1-p)^2$$
$$f(2) = P(X = 2) = P(FSS) + P(SFS) + P(SSF) = 3\,p^2\,(1-p)$$
$$f(3) = P(X = 3) = P(SSS) = p^3.$$

Hence

$$f(x) = \binom{3}{x} p^x\,(1-p)^{3-x}, \qquad x = 0,\, 1,\, 2,\, 3.$$

Thus

$$X \sim BIN(3,\, p).$$

In general, if $X_i \sim BER(p)$, then $\sum_{i=1}^{n} X_i \sim BIN(n, p)$ and hence

$$E\left(\sum_{i=1}^{n} X_i\right) = n\,p$$

and

$$Var\left(\sum_{i=1}^{n} X_i\right) = n\,p\,(1-p).$$

The binomial distribution can arise whenever we select a random sample of $n$ units with replacement. Each unit in the population is classified into one of two categories according to whether it does or does not possess a certain property. For example, the unit may be a person and the property may be

whether he intends to vote "yes". If the unit is a machine part, the property may be whether the part is defective and so on. If the proportion of units in the population possessing the property of interest is $p$, and if $Z$ denotes the number of units in the sample of size $n$ that possess the given property, then

$$Z \sim BIN(n,\, p).$$

### 5.3. Geometric Distribution

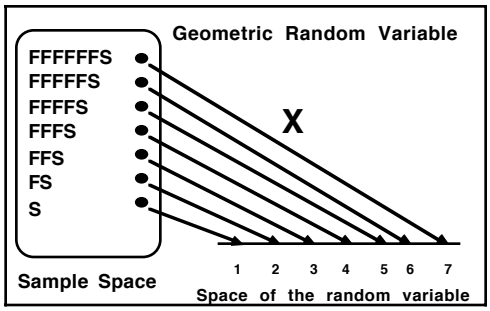
If $X$ represents the total number of successes in $n$ independent Bernoulli trials, then the random variable

$$X \sim BIN(n, p),$$

where $p$ is the probability of success of a single Bernoulli trial and the probability density function of $X$ is given by

$$f(x) = \binom{n}{x} p^x \, (1-p)^{n-x}, \qquad x = 0,\, 1,\, ...,\, n.$$

Let $X$ denote the trial number on which the first success occurs.




Hence the probability that the first success occurs on $x^{\text{th}}$ trial is given by

$$f(x) = P(X = x) = (1-p)^{x-1}\, p.$$

Hence, the probability density function of $X$ is

$$f(x) = (1-p)^{x-1}\, p \qquad\qquad x = 1,\, 2,\, 3,\, ...,\, \infty,$$

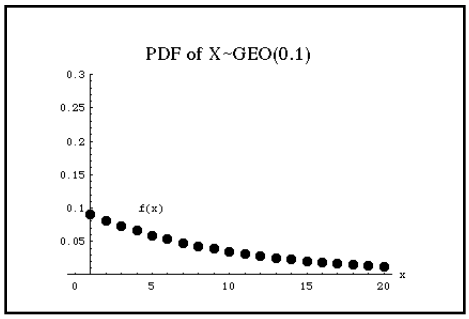
where $p$ denotes the probability of success in a single Bernoulli trial.

**Definition 5.3.** A random variable $X$ has a geometric distribution if its probability density function is given by

$$f(x) = (1-p)^{x-1}\, p \qquad\qquad x = 1,\, 2,\, 3, ...,\, \infty,$$

where $p$ denotes the probability of success in a single Bernoulli trial.

If $X$ has a geometric distribution we denote it as $X \sim GEO(p)$.




**Example 5.10.** Is the real valued function $f(x)$ defined by

$$f(x) = (1-p)^{x-1}\, p \qquad\qquad x = 1,\, 2,\, 3,\, ..., \infty$$

where $0 < p < 1$ is a parameter, a probability density function?

**Answer:** It is easy to check that $f(x) \geq 0$. Thus we only show that the sum is one.

$$\begin{aligned}\sum_{x=1}^{\infty} f(x) &= \sum_{x=1}^{\infty}(1-p)^{x-1}\, p \\ &= p\sum_{y=0}^{\infty}(1-p)^{y}, \qquad \text{where } y = x-1 \\ &= p\,\frac{1}{1-(1-p)} = 1.\end{aligned}$$

Hence $f(x)$ is a probability density function.

**Example 5.11.** The probability that a machine produces a defective item is 0.02. Each item is checked as it is produced. Assuming that these are independent trials, what is the probability that at least 100 items must be checked to find one that is defective?

**Answer:** Let $X$ denote the trial number on which the first defective item is observed. We want to find

$$\begin{aligned} P(X \geq 100) &= \sum_{x=100}^{\infty} f(x) \\ &= \sum_{x=100}^{\infty} (1-p)^{x-1}\, p \\ &= (1-p)^{99} \sum_{y=0}^{\infty} (1-p)^{y}\, p \\ &= (1-p)^{99} \\ &= (0.98)^{99} = 0.1353. \end{aligned}$$

Hence the probability that at least 100 items must be checked to find one that is defective is 0.1353.

**Example 5.12.** A gambler plays roulette at Monte Carlo and continues gambling, wagering the same amount each time on "Red", until he wins for the first time. If the probability of "Red" is $\frac{18}{38}$ and the gambler has only enough money for 5 trials, then (a) what is the probability that he will win before he exhausts his funds; (b) what is the probability that he wins on the second trial?

**Answer:**

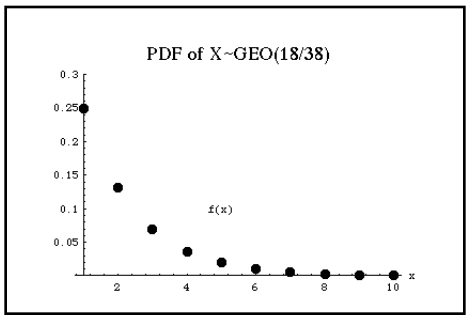
$$p = P(Red) = \frac{18}{38}.$$

(a) Hence the probability that he will win before he exhausts his funds is given by

$$\begin{aligned} P(X \leq 5) &= 1 - P(X \geq 6) \\ &= 1 - (1-p)^5 \\ &= 1 - \left(1 - \frac{18}{38}\right)^5 \\ &= 1 - (0.5263)^5 = 1 - 0.044 = 0.956. \end{aligned}$$

(b) Similarly, the probability that he wins on the second trial is given by

$$\begin{aligned} P(X = 2) &= f(2) \\ &= (1-p)^{2-1}\, p \\ &= \left(1 - \frac{18}{38}\right)\left(\frac{18}{38}\right) \\ &= \frac{360}{1444} = 0.2493. \end{aligned}$$




The following theorem provides us with the mean, variance and moment generating function of a random variable with the geometric distribution.

**Theorem 5.3.** If $X$ is a geometric random variable with parameter $p$, then the mean, variance and moment generating functions are respectively given by

$$\mu_X = \frac{1}{p}$$
$$\sigma_X^2 = \frac{1-p}{p^2}$$
$$M_X(t) = \frac{p\,e^t}{1-(1-p)\,e^t}, \qquad \text{if } t < -ln(1-p).$$

**Proof:** First, we compute the moment generating function of $X$ and then we generate all the mean and variance of $X$ from it.

$$\begin{aligned} M(t) &= \sum_{x=1}^{\infty} e^{tx}\,(1-p)^{x-1}\,p \\ &= p\sum_{y=0}^{\infty} e^{t(y+1)}\,(1-p)^{y}, \qquad \text{where } y = x-1 \\ &= p\,e^t \sum_{y=0}^{\infty} \left(e^t\,(1-p)\right)^y \\ &= \frac{p\,e^t}{1-(1-p)\,e^t}, \qquad \text{if } t < -ln(1-p). \end{aligned}$$

Differentiating $M(t)$ with respect to $t$, we obtain

$$M'(t) = \frac{(1-(1-p)\,e^t)\;p\,e^t + p\,e^t\,(1-p)\,e^t}{\left[1-(1-p)e^t\right]^2}$$

$$= \frac{p\,e^t\;\left[1-(1-p)\,e^t + (1-p)\,e^t\right]}{\left[1-(1-p)e^t\right]^2}$$

$$= \frac{p\,e^t}{\left[1-(1-p)e^t\right]^2}.$$

Hence

$$\mu_X = E(X) = M'(0) = \frac{1}{p}.$$

Similarly, the second derivative of $M(t)$ can be obtained from the first derivative as

$$M''(t) = \frac{\left[1-(1-p)\,e^t\right]^2 p\,e^t + p\,e^t\,2\left[1-(1-p)\,e^t\right]\,(1-p)\,e^t}{\left[1-(1-p)e^t\right]^4}.$$

Hence

$$M''(0) = \frac{p^3 + 2\,p^2(1-p)}{p^4} = \frac{2-p}{p^2}.$$

Therefore, the variance of $X$ is

$$\begin{aligned}\sigma_X^2 &= M''(0) - (\,M'(0)\,)^2\\ &= \frac{2-p}{p^2} - \frac{1}{p^2}\\ &= \frac{1-p}{p^2}.\end{aligned}$$

This completes the proof of the theorem.

**Theorem 5.4.** The random variable $X$ is geometric if and only if it satisfies the memoryless property, that is

$$P(X > m+n\,/\,X > n) = P(X > m)$$

for all natural numbers $n$ and $m$.

**Proof:** It is easy to check that the geometric distribution satisfies the lack of memory property

$$P(X > m+n\,/\,X > n) = P(X > m)$$

which is

$$P(X > m + n \text{ and } X > n) = P(X > m)\, P(X > n). \tag{5.1}$$

If $X$ is geometric, that is $X \sim (1-p)^{x-1}\, p$, then

$$\begin{aligned} P(X > n + m) &= \sum_{x=n+m+1}^{\infty} (1-p)^{x-1}\, p \\ &= (1-p)^{n+m} \\ &= (1-p)^n\, (1-p)^m \\ &= P(X > n)\, P(X > m). \end{aligned}$$

Hence the geometric distribution has the lack of memory property. Let $X$ be a random variable which satisfies the lack of memory property, that is

$$P(X > m + n \text{ and } X > n) = P(X > m)\, P(X > n).$$

We want to show that $X$ is geometric. Define $g : \mathbf{N} \to \mathbb{R}$ by

$$g(n) := P(X > n) \tag{5.2}$$

Using (5.2) in (5.1), we get

$$g(m + n) = g(m)\, g(n) \qquad \forall\, m, n \in \mathbf{N}, \tag{5.3}$$

since $P(X > m + n \text{ and } X > n) = P(X > m + n)$. Letting $m = 1$ in (5.3), we see that

$$\begin{aligned} g(n+1) &= g(n)\, g(1) \\ &= g(n-1)\, g(1)^2 \\ &= g(n-2)\, g(1)^3 \\ &= \cdots \qquad \cdots \\ &= g(n-(n-1))\, g(1)^n \\ &= g(1)^{n+1} \\ &= a^{n+1}, \end{aligned}$$

where $a$ is an arbitrary constant. Hence $g(n) = a^n$. From (5.2), we get

$$1 - F(n) = P(X > n) = a^n$$

and thus

$$F(n) = 1 - a^n.$$

Since $F(n)$ is a distribution function

$$1 = \lim_{n\to\infty} F(n) = \lim_{n\to\infty} \left(1 - a^n\right).$$

From the above, we conclude that $0 < a < 1$. We rename the constant $a$ as $(1-p)$. Thus,

$$F(n) = 1 - (1-p)^n.$$

The probability density function of $X$ is hence

$$\begin{aligned}
f(1) &= F(1) = p \\
f(2) &= F(2) - F(1) = 1 - (1-p)^2 - 1 + (1-p) = (1-p)\,p \\
f(3) &= F(3) - F(2) = 1 - (1-p)^3 - 1 + (1-p)^2 = (1-p)^2\,p \\
&\cdots \qquad \cdots \\
f(x) &= F(x) - F(x-1) = (1-p)^{x-1}\,p.
\end{aligned}$$

Thus $X$ is geometric with parameter $p$. This completes the proof.

The difference between the binomial and the geometric distributions is the following. In binomial distribution, the number of trials was predetermined, whereas in geometric it is the random variable.

### 5.4. Negative Binomial Distribution

Let $X$ denote the trial number on which the $r^{\text{th}}$ success occurs. Here $r$ is a positive integer greater than or equal to one. This is equivalent to saying that the random variable $X$ denotes the number of trials needed to observe the $r^{\text{th}}$ successes. Suppose we want to find the probability that the fifth head is observed on the $10^{\text{th}}$ independent flip of an unbiased coin. This is a case of finding $P(X = 10)$. Let us find the general case $P(X = x)$.

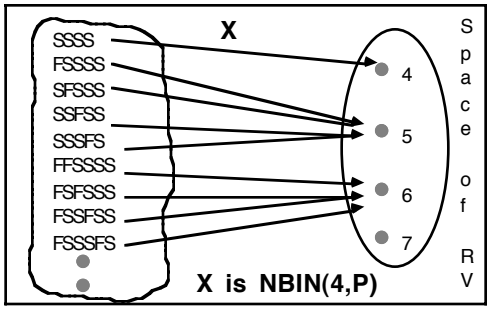
$$\begin{aligned}
P(X = x) &= P(\text{first } x-1 \text{ trials contain } x-r \text{ failures and } r-1 \text{ successes}) \\
&\quad \cdot P(r^{\text{th}} \text{ success in } x^{\text{th}} \text{ trial}) \\
&= \binom{x-1}{r-1} p^{r-1}\,(1-p)^{x-r}\,p \\
&= \binom{x-1}{r-1} p^{r}\,(1-p)^{x-r}, \qquad x = r,\, r+1, ..., \infty.
\end{aligned}$$

Hence the probability density function of the random variable $X$ is given by

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, ..., \infty.$$

Notice that this probability density function $f(x)$ can also be expressed as

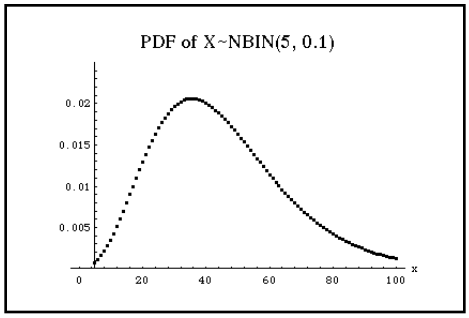
$$f(x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \qquad x = 0, 1, ..., \infty.$$




**Definition 5.4.** A random variable $X$ has the negative binomial (or Pascal) distribution if its probability density function is of the form

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, ..., \infty,$$

where $p$ is the probability of success in a single Bernoulli trial. We denote the random variable $X$ whose distribution is negative binomial distribution by writing $X \sim NBIN(r, p)$.




We need the following technical result to show that the above function is really a probability density function. The technical result we are going to establish is called the negative binomial theorem.

**Theorem 5.5.** Let $r$ be a nonzero positive integer. Then

$$(1-y)^{-r} = \sum_{x=r}^{\infty} \binom{x-1}{r-1} y^{x-r}$$

where $|y| < 1$.

**Proof:** Define

$$h(y) = (1-y)^{-r}.$$

Now expanding $h(y)$ by Taylor series method about $y = 0$, we get

$$(1-y)^{-r} = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} y^k,$$

where $h^{(k)}(y)$ is the $k^{\text{th}}$ derivative of $h$. This $k^{\text{th}}$ derivative of $h(y)$ can be directly computed and direct computation gives

$$h^{(k)}(y) = r\,(r+1)\,(r+2)\cdots(r+k-1)\,(1-y)^{-(r+k)}.$$

Hence, we get

$$h^{(k)}(0) = r\,(r+1)\,(r+2)\cdots(r+k-1) = \frac{(r+k-1)!}{(r-1)!}.$$

Letting this into the Taylor's expansion of $h(y)$, we get

$$\begin{aligned}(1-y)^{-r} &= \sum_{k=0}^{\infty} \frac{(r+k-1)!}{(r-1)!\,k!} y^k \\ &= \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} y^k.\end{aligned}$$

Letting $x = k + r$, we get

$$(1-y)^{-r} = \sum_{x=r}^{\infty} \binom{x-1}{r-1} y^{x-r}.$$

This completes the proof of the theorem.

Theorem 5.5 can also be proved using the geometric series

$$\sum_{n=0}^{\infty} y^n = \frac{1}{1-y} \tag{5.4}$$

where $|y| < 1$. Differentiating $k$ times both sides of the equality (5.4) and then simplifying we have

$$\sum_{n=k}^{\infty} \binom{n}{k} y^{n-k} = \frac{1}{(1-y)^{k+1}}. \tag{5.5}$$

Letting $n = x - 1$ and $k = r - 1$ in (5.5), we have the asserted result.

**Example 5.13.** Is the real valued function defined by

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, ..., \infty,$$

where $0 < p < 1$ is a parameter, a probability density function?

**Answer:** It is easy to check that $f(x) \geq 0$. Now we show that $\sum_{x=r}^{\infty} f(x)$ is equal to one.

$$\begin{aligned}
\sum_{x=r}^{\infty} f(x) &= \sum_{x=r}^{\infty} \binom{x-1}{r-1} p^r (1-p)^{x-r} \\
&= p^r \sum_{x=r}^{\infty} \binom{x-1}{r-1} (1-p)^{x-r} \\
&= p^r (1-(1-p))^{-r} \\
&= p^r p^{-r} \\
&= 1.
\end{aligned}$$

Computing the mean and variance of the negative binomial distribution using definition is difficult. However, if we use the moment generating approach, then it is not so difficult. Hence in the next example, we determine the moment generating function of this negative binomial distribution.

**Example 5.14.** What is the moment generating function of the random variable $X$ whose probability density function is

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, ..., \infty?$$

**Answer:** The moment generating function of this negative binomial random

variable is

$$\begin{aligned}
M(t) &= \sum_{x=r}^{\infty} e^{tx} f(x) \\
&= \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r} \\
&= p^r \sum_{x=r}^{\infty} e^{t(x-r)} e^{tr} \binom{x-1}{r-1} (1-p)^{x-r} \\
&= p^r e^{tr} \sum_{x=r}^{\infty} \binom{x-1}{r-1} e^{t(x-r)} (1-p)^{x-r} \\
&= p^r e^{tr} \sum_{x=r}^{\infty} \binom{x-1}{r-1} \left[e^t (1-p)\right]^{x-r} \\
&= p^r e^{tr} \left[1-(1-p)e^t\right]^{-r} \\
&= \left(\frac{p\, e^t}{1-(1-p)e^t}\right)^r, \qquad \text{if } t < -ln(1-p).
\end{aligned}$$

The following theorem can easily be proved.

**Theorem 5.6.** If $X \sim NBIN(r,p)$, then

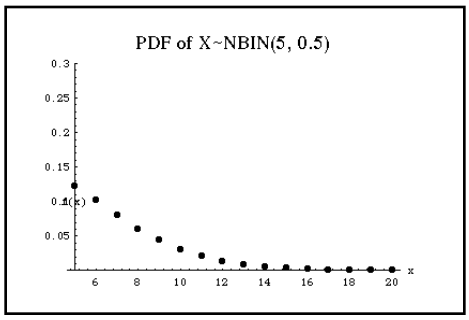
$$\begin{aligned}
E(X) &= \frac{r}{p} \\
Var(X) &= \frac{r\,(1-p)}{p^2} \\
M(t) &= \left(\frac{p\, e^t}{1-(1-p)e^t}\right)^r, \qquad \text{if } t < -ln(1-p).
\end{aligned}$$

**Example 5.15.** What is the probability that the fifth head is observed on the $10^{\text{th}}$ independent flip of a coin?

**Answer:** Let $X$ denote the number of trials needed to observe $5^{\text{th}}$ head. Hence $X$ has a negative binomial distribution with $r = 5$ and $p = \frac{1}{2}$.

We want to find

$$\begin{aligned}
P(X = 10) &= f(10) \\
&= \binom{9}{4} p^5 (1-p)^5 \\
&= \binom{9}{4} \left(\frac{1}{2}\right)^{10} \\
&= \frac{63}{512}.
\end{aligned}$$




We close this section with the following comment. In the negative binomial distribution the parameter $r$ is a positive integer. One can generalize the negative binomial distribution to allow noninteger values of the parameter $r$. To do this let us write the probability density function of the negative binomial distribution as

$$\begin{aligned} f(x) &= \binom{x-1}{r-1} p^r (1-p)^{x-r} \\ &= \frac{(x-1)!}{(r-1)!\,(x-r)!}\, p^r (1-p)^{x-r} \\ &= \frac{\Gamma(x)}{\Gamma(r)\,\Gamma(x-r-1)}\, p^r (1-p)^{x-r}, \quad \text{for } x = r, r+1, ..., \infty, \end{aligned}$$

where

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$$

is the well known gamma function. The gamma function generalizes the notion of factorial and it will be treated in the next chapter.

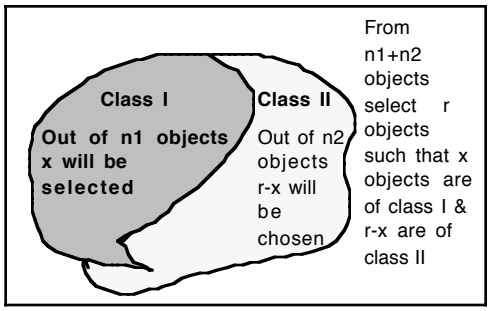
### 5.5. Hypergeometric Distribution

Consider a collection of $n$ objects which can be classified into two classes, say class 1 and class 2. Suppose that there are $n_1$ objects in class 1 and $n_2$ objects in class 2. A collection of $r$ objects is selected from these $n$ objects at random and without replacement. We are interested in finding out the probability that exactly $x$ of these $r$ objects are from class 1. If $x$ of these $r$ objects are from class 1, then the remaining $r-x$ objects must be from class 2. We can select $x$ objects from class 1 in any one of $\binom{n_1}{x}$ ways. Similarly, the remaining $r-x$ objects can be selected in $\binom{n_2}{r-x}$ ways. Thus, the number of ways one can select a subset of $r$ objects from a set of $n$ objects, such that

$x$ number of objects will be from class 1 and $r-x$ number of objects will be from class 2, is given by $\binom{n_1}{x}\binom{n_2}{r-x}$. Hence,

$$P(X=x)=\frac{\binom{n_1}{x}\binom{n_2}{r-x}}{\binom{n}{r}},$$

where $x \leq r,\ x \leq n_1$ and $r-x \leq n_2$.




**Definition 5.5.** A random variable $X$ is said to have a hypergeometric distribution if its probability density function is of the form

$$f(x)=\frac{\binom{n_1}{x}\binom{n_2}{r-x}}{\binom{n_1+n_2}{r}}, \qquad x=0,1,2,...,r$$

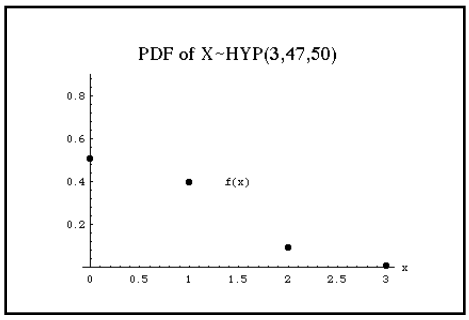
where $x \leq n_1$ and $r-x \leq n_2$ with $n_1$ and $n_2$ being two positive integers. We shall denote such a random variable by writing

$$X \sim HYP(n_1, n_2, r).$$

**Example 5.16.** Suppose there are 3 defective items in a lot of 50 items. A sample of size 10 is taken at random and without replacement. Let $X$ denote the number of defective items in the sample. What is the probability that the sample contains at most one defective item?

**Answer:** Clearly, $X \sim HYP(3, 47, 10)$. Hence the probability that the sample contains at most one defective item is

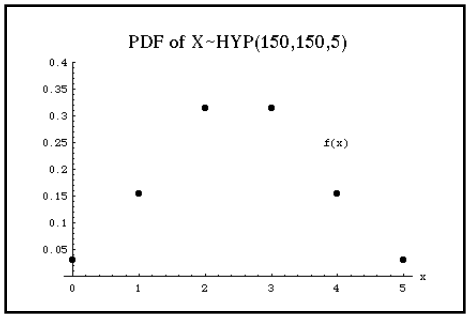
$$\begin{aligned} P(X \leq 1) &= P(X=0)+P(X=1) \\ &= \frac{\binom{3}{0}\binom{47}{10}}{\binom{50}{10}}+\frac{\binom{3}{1}\binom{47}{9}}{\binom{50}{10}} \\ &= 0.504+0.4 \\ &= 0.904. \end{aligned}$$




**Example 5.17.** A random sample of 5 students is drawn without replacement from among 300 seniors, and each of these 5 seniors is asked if she/he has tried a certain drug. Suppose 50% of the seniors actually have tried the drug. What is the probability that two of the students interviewed have tried the drug?

**Answer:** Let $X$ denote the number of students interviewed who have tried the drug. Hence the probability that two of the students interviewed have tried the drug is

$$P(X=2) = \frac{\binom{150}{2}\binom{150}{3}}{\binom{300}{5}}$$
$$= 0.3146.$$




**Example 5.18.** A radio supply house has 200 transistor radios, of which 3 are improperly soldered and 197 are properly soldered. The supply house randomly draws 4 radios without replacement and sends them to a customer. What is the probability that the supply house sends 2 improperly soldered radios to its customer?

**Answer:** The probability that the supply house sends 2 improperly soldered

radios to its customer is

$$P(X=2) = \frac{\binom{3}{2}\binom{197}{2}}{\binom{200}{4}} = 0.000895.$$

**Theorem 5.7.** If $X \sim HYP(n_1, n_2, r)$, then

$$E(X) = r\,\frac{n_1}{n_1+n_2}$$
$$Var(X) = r\,\left(\frac{n_1}{n_1+n_2}\right)\left(\frac{n_2}{n_1+n_2}\right)\left(\frac{n_1+n_2-r}{n_1+n_2-1}\right).$$

**Proof:** Let $X \sim HYP(n_1, n_2, r)$. We compute the mean and variance of $X$ by computing the first and the second factorial moments of the random variable $X$. First, we compute the first factorial moment (which is same as the expected value) of $X$. The expected value of $X$ is given by

$$\begin{aligned}
E(X) &= \sum_{x=0}^{r} x\, f(x) \\
&= \sum_{x=0}^{r} x\,\frac{\binom{n_1}{x}\binom{n_2}{r-x}}{\binom{n_1+n_2}{r}} \\
&= n_1\,\sum_{x=1}^{r}\,\frac{(n_1-1)!}{(x-1)!\,(n_1-x)!}\,\frac{\binom{n_2}{r-x}}{\binom{n_1+n_2}{r}} \\
&= n_1\,\sum_{x=1}^{r}\,\frac{\binom{n_1-1}{x-1}\binom{n_2}{r-x}}{\frac{n_1+n_2}{r}\binom{n_1+n_2-1}{r-1}} \\
&= r\,\frac{n_1}{n_1+n_2}\sum_{y=0}^{r-1}\,\frac{\binom{n_1-1}{y}\binom{n_2}{r-1-y}}{\binom{n_1+n_2-1}{r-1}}, \qquad \text{where } y = x-1 \\
&= r\,\frac{n_1}{n_1+n_2}.
\end{aligned}$$

The last equality is obtained since

$$\sum_{y=0}^{r-1}\,\frac{\binom{n_1-1}{y}\binom{n_2}{r-1-y}}{\binom{n_1+n_2-1}{r-1}} = 1.$$

Similarly, we find the second factorial moment of $X$ to be

$$E(X(X-1)) = \frac{r(r-1)\,n_1\,(n_1-1)}{(n_1+n_2)\,(n_1+n_2-1)}.$$

Therefore, the variance of $X$ is

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= E(X(X-1)) + E(X) - E(X)^2 \\ &= \frac{r(r-1)\, n_1\, (n_1-1)}{(n_1+n_2)\,(n_1+n_2-1)} + r\, \frac{n_1}{n_1+n_2} - \left( r\, \frac{n_1}{n_1+n_2} \right)^2 \\ &= r \left( \frac{n_1}{n_1+n_2} \right) \left( \frac{n_2}{n_1+n_2} \right) \left( \frac{n_1+n_2-r}{n_1+n_2-1} \right). \end{aligned}$$

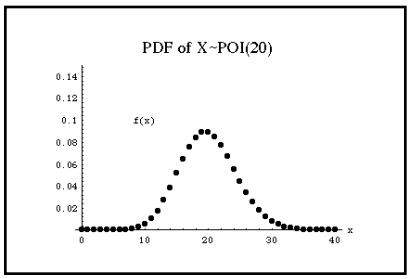
### 5.6. Poisson Distribution

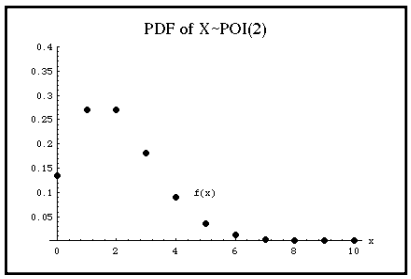
In this section, we define an important discrete distribution which is widely used for modeling many real life situations. First, we define this distribution and then we present some of its important properties.

**Definition 5.6.** A random variable $X$ is said to have a Poisson distribution if its probability density function is given by

$$f(x) = \frac{e^{-\lambda}\, \lambda^x}{x!}, \qquad x = 0,\, 1,\, 2, \cdots, \infty,$$

where $0 < \lambda < \infty$ is a parameter. We denote such a random variable by $X \sim POI(\lambda)$.




The probability density function $f$ is called the Poisson distribution after Simeon D. Poisson (1781-1840).

**Example 5.19.** Is the real valued function defined by

$$f(x) = \frac{e^{-\lambda}\, \lambda^x}{x!}, \qquad x = 0,\, 1,\, 2, \cdots, \infty,$$

where $0 < \lambda < \infty$ is a parameter, a probability density function?

**Answer:** It is easy to check $f(x) \geq 0$. We show that $\sum_{x=0}^{\infty} f(x)$ is equal to one.

$$\begin{aligned}\sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\,\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda}\, e^{\lambda} = 1.\end{aligned}$$

**Theorem 5.8.** If $X \sim POI(\lambda)$, then

$$\begin{aligned}E(X) &= \lambda \\ Var(X) &= \lambda \\ M(t) &= e^{\lambda\,(e^t-1)}.\end{aligned}$$

**Proof:** First, we find the moment generating function of $X$.

$$\begin{aligned}M(t) &= \sum_{x=0}^{\infty} e^{tx}\, f(x) \\ &= \sum_{x=0}^{\infty} e^{tx}\, \frac{e^{-\lambda}\,\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} e^{tx}\, \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!} \\ &= e^{-\lambda}\, e^{\lambda e^t} \\ &= e^{\lambda\,(e^t-1)}.\end{aligned}$$

Thus,

$$M'(t) = \lambda\, e^t\, e^{\lambda\,(e^t-1)},$$

and

$$E(X) = M'(0) = \lambda.$$

Similarly,

$$M''(t) = \lambda\, e^t\, e^{\lambda\,(e^t-1)} + \left(\lambda\, e^t\right)^2\, e^{\lambda\,(e^t-1)}.$$

Hence

$$M''(0) = E(X^2) = \lambda^2 + \lambda.$$

Therefore

$$Var(X) = E(X^2) - (\,E(X)\,)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

This completes the proof.

**Example 5.20.** A random variable $X$ has a Poisson distribution with a mean of 3. What is the probability that $X$ is bounded by 1 and 3, that is, $P(1 \leq X \leq 3)$?

**Answer:**

$$\mu_X = 3 = \lambda$$

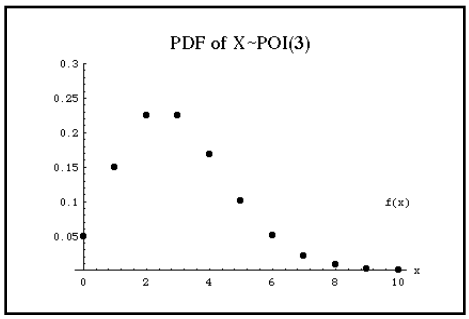
$$f(x) = \frac{\lambda^x\, e^{-\lambda}}{x!}$$

Hence

$$f(x) = \frac{3^x\, e^{-3}}{x!}, \qquad x = 0,\, 1,\, 2,\, ...$$

Therefore

$$\begin{aligned} P(1 \leq X \leq 3) &= f(1) + f(2) + f(3) \\ &= 3\, e^{-3} + \frac{9}{2}\, e^{-3} + \frac{27}{6}\, e^{-3} \\ &= 12\, e^{-3}. \end{aligned}$$




**Example 5.21.** The number of traffic accidents per week in a small city has a Poisson distribution with mean equal to 3. What is the probability of exactly 2 accidents occur in 2 weeks?

**Answer:** The mean traffic accident is 3. Thus, the mean accidents in two weeks are

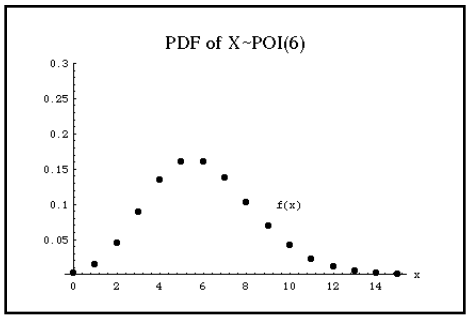
$$\lambda = (3)\,(2) = 6.$$

Since

$$f(x) = \frac{\lambda^x \, e^{-\lambda}}{x!}$$

we get

$$f(2) = \frac{6^2 \, e^{-6}}{2!} = 18 \, e^{-6}.$$




**Example 5.22.** Let $X$ have a Poisson distribution with parameter $\lambda = 1$. What is the probability that $X \geq 2$ given that $X \leq 4$?

**Answer:**

$$P(X \geq 2 \,/\, X \leq 4) = \frac{P(2 \leq X \leq 4)}{P(X \leq 4)}.$$

$$\begin{aligned} P(2 \leq X \leq 4) &= \sum_{x=2}^{4} \frac{\lambda^x \, e^{-\lambda}}{x!} \\ &= \frac{1}{e} \sum_{x=2}^{4} \frac{1}{x!} \\ &= \frac{17}{24 \, e}. \end{aligned}$$

Similarly

$$\begin{aligned} P(X \leq 4) &= \frac{1}{e} \sum_{x=0}^{4} \frac{1}{x!} \\ &= \frac{65}{24 \, e}. \end{aligned}$$

Therefore, we have

$$P(X \geq 2 \,/\, X \leq 4) = \frac{17}{65}.$$

**Example 5.23.** If the moment generating function of a random variable $X$ is $M(t) = e^{4.6\,(e^t - 1)}$, then what are the mean and variance of $X$? What is the probability that $X$ is between 3 and 6, that is $P(3 < X < 6)$?

**Answer:** Since the moment generating function of $X$ is given by

$$M(t) = e^{4.6\,(e^t-1)}$$

we conclude that $X \sim POI(\lambda)$ with $\lambda = 4.6$. Thus, by Theorem 5.8, we get

$$E(X) = 4.6 = Var(X).$$

$$\begin{aligned} P(3 < X < 6) &= f(4) + f(5) \\ &= F(5) - F(3) \\ &= 0.686 - 0.326 \\ &= 0.36. \end{aligned}$$

### 5.7. Riemann Zeta Distribution

The zeta distribution was used by the Italian economist Vilfredo Pareto (1848-1923) to study the distribution of family incomes of a given country.

**Definition 5.7.** A random variable $X$ is said to have Riemann zeta distribution if its probability density function is of the form

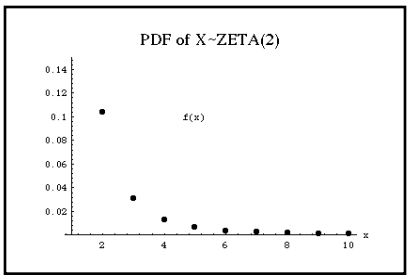
$$f(x) = \frac{1}{\zeta(\alpha+1)}\, x^{-(\alpha+1)}, \qquad x = 1, 2, 3, ..., \infty$$

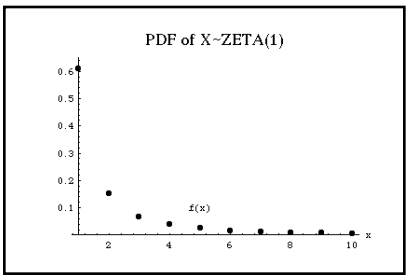
where $\alpha > 0$ is a parameter and

$$\zeta(s) = 1 + \left(\frac{1}{2}\right)^s + \left(\frac{1}{3}\right)^s + \left(\frac{1}{4}\right)^s + \cdots + \left(\frac{1}{x}\right)^s + \cdots$$

is the well known the Riemann zeta function. A random variable having a Riemann zeta distribution with parameter $\alpha$ will be denoted by $X \sim RIZ(\alpha)$.

The following figures illustrate the Riemann zeta distribution for the case $\alpha = 2$ and $\alpha = 1$.



The following theorem is easy to prove and we leave its proof to the reader.

**Theorem 5.9.** If $X \sim RIZ(\alpha)$, then

$$E(X) = \frac{\zeta(\alpha)}{\zeta(\alpha+1)}$$
$$Var(X) = \frac{\zeta(\alpha-1)\zeta(\alpha+1) - (\zeta(\alpha))^2}{(\zeta(\alpha+1))^2}.$$

**Remark 5.1.** If $0 < \alpha \leq 1$, then $\zeta(\alpha) = \infty$. Hence if $X \sim RIZ(\alpha)$ and the parameter $\alpha \leq 1$, then the variance of $X$ is infinite.

### 5.8. Review Exercises

**1.** What is the probability of getting exactly 3 heads in 5 flips of a fair coin?

**2.** On six successive flips of a fair coin, what is the probability of observing 3 heads and 3 tails?

**3.** What is the probability that in 3 rolls of a pair of six-sided dice, exactly one total of 7 is rolled?

**4.** What is the probability of getting exactly four "sixes" when a die is rolled 7 times?

**5.** In a family of 4 children, what is the probability that there will be exactly two boys?

**6.** If a fair coin is tossed 4 times, what is the probability of getting at least two heads?

**7.** In Louisville the probability that a thunderstorm will occur on any day during the spring is 0.05. Assuming independence, what is the probability that the first thunderstorm occurs on April 5? (Assume spring begins on March 1.)

**8.** A ball is drawn from an urn containing 3 white and 3 black balls. After the ball is drawn, it is then replaced and another ball is drawn. This goes on indefinitely. What is the probability that, of the first 4 balls drawn, exactly 2 are white?

**9.** What is the probability that a person flipping a fair coin requires four tosses to get a head?

**10.** Assume that hitting oil at one drilling location is independent of another, and that, in a particular region, the probability of success at any individual

location is 0.3. Suppose the drilling company believes that a venture will be profitable if the number of wells drilled until the second success occurs is less than or equal to 7. What is the probability that the venture will be profitable?

**11.** Suppose an experiment consists of tossing a fair coin until three heads occur. What is the probability that the experiment ends after exactly six flips of the coin with a head on the fifth toss as well as on the sixth?

**12.** Customers at Fred's Cafe wins a \$100 prize if their cash register receipts show a star on each of the five consecutive days Monday, Tuesday, ..., Friday in any one week. The cash register is programmed to print stars on a randomly selected 10% of the receipts. If Mark eats at Fred's Cafe once each day for four consecutive weeks, and if the appearance of the stars is an independent process, what is the probability that Mark will win at least \$100?

**13.** If a fair coin is tossed repeatedly, what is the probability that the third head occurs on the $n^{\text{th}}$ toss?

**14.** Suppose 30 percent of all electrical fuses manufactured by a certain company fail to meet municipal building standards. What is the probability that in a random sample of 10 fuses, exactly 3 will fail to meet municipal building standards?

**15.** A bin of 10 light bulbs contains 4 that are defective. If 3 bulbs are chosen without replacement from the bin, what is the probability that exactly $k$ of the bulbs in the sample are defective?

**16.** Let $X$ denote the number of independent rolls of a fair die required to obtain the first "3". What is $P(X \geq 6)$?

**17.** The number of automobiles crossing a certain intersection during any time interval of length $t$ minutes between 3:00 P.M. and 4:00 P.M. has a Poisson distribution with mean $t$. Let $W$ be time elapsed after 3:00 P.M. before the first automobile crosses the intersection. What is the probability that $W$ is less than 2 minutes?

**18.** In rolling one die repeatedly, what is the probability of getting the third six on the $x^{\text{th}}$ roll?

**19.** A coin is tossed 6 times. What is the probability that the number of heads in the first 3 throws is the same as the number in the last 3 throws?

**20.** One hundred pennies are being distributed independently and at random into 30 boxes, labeled 1, 2, ..., 30. What is the probability that there are exactly 3 pennies in box number 1?

**21.** The density function of a certain random variable is

$$f(x) = \begin{cases} \binom{22}{4x} (0.2)^{4x} (0.8)^{22-4x} & \text{if } x = 0, \frac{1}{4}, \frac{2}{4}, \cdots, \frac{22}{4} \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of $X^2$ ?

**22.** If $M_X(t) = k\ (2 + 3e^t)^{100}$, what is the value of $k$? What is the variance of the random variable $X$?

**23.** If $M_X(t) = k\ \left(\frac{e^t}{7-5e^t}\right)^3$, what is the value of $k$? What is the variance of the random variable $X$?

**24.** If for a Poisson distribution $2f(0) + f(2) = 2f(1)$, what is the mean of the distribution?

**25.** The number of hits, $X$, per baseball game, has a Poisson distribution. If the probability of a no-hit game is $\frac{1}{3}$, what is the probability of having 2 or more hits in specified game?

**26.** Suppose $X$ has a Poisson distribution with a standard deviation of 4. What is the conditional probability that $X$ is exactly 1 given that $X \geq 1$ ?

**27.** A die is loaded in such a way that the probability of the face with $j$ dots turning up is proportional to $j^2$ for $j = 1, 2, 3, 4, 5, 6$. What is the probability of rolling at most three sixes in 5 independent casts of this die?

**28.** A die is loaded in such a way that the probability of the face with $j$ dots turning up is proportional to $j^2$ for $j = 1, 2, 3, 4, 5, 6$. What is the probability of getting the third six on the $7^{\text{th}}$ roll of this loaded die?

# Chapter 6

# SOME SPECIAL CONTINUOUS DISTRIBUTIONS

In this chapter, we study some well known continuous probability density functions. We want to study them because they arise in many applications. We begin with the simplest probability density function.

## 6.1. Uniform Distribution

Let the random variable $X$ denote the outcome when a point is selected at random from an interval $[a, b]$. We want to find the probability of the event $X \leq x$, that is we would like to determine the probability that the point selected from $[a, b]$ would be less than or equal to $x$. To compute this probability, we need a probability measure $\mu$ that satisfies the three axioms of Kolmogorov (namely nonnegativity, normalization and countable additivity). For continuous variables, the events are interval or union of intervals. The length of the interval when normalized satisfies all the three axioms and thus it can be used as a probability measure for one-dimensional random variables. Hence

$$P(X \leq x) = \frac{\text{length of } [\text{a}, \text{x}]}{\text{length of } [\text{a}, \text{b}]}.$$

Thus, the cumulative distribution function $F$ is

$$F(x) = P(X \leq x) = \frac{x-a}{b-a}, \qquad a \leq x \leq b,$$

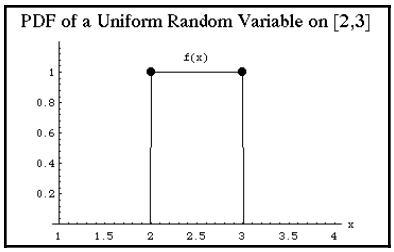
where $a$ and $b$ are any two real constants with $a < b$. To determine the probability density function from cumulative density function, we calculate the derivative of $F(x)$. Hence

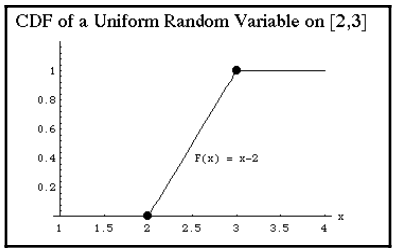
$$f(x) = \frac{d}{dx}F(x) = \frac{1}{b-a}, \qquad a \leq x \leq b.$$

**Definition 6.1.** A random variable $X$ is said to be uniform on the interval $[a, b]$ if its probability density function is of the form

$$f(x) = \frac{1}{b-a}, \qquad a \leq x \leq b,$$

where $a$ and $b$ are constants. We denote a random variable $X$ with the uniform distribution on the interval $[a, b]$ as $X \sim UNIF(a, b)$.




The uniform distribution provides a probability model for selecting points at random from an interval $[a, b]$. An important application of uniform distribution lies in random number generation. The following theorem gives the mean, variance and moment generating function of a uniform random variable.

**Theorem 6.1.** If $X$ is uniform on the interval $[a, b]$ then the mean, variance and moment generating function of $X$ are given by

$$E(X) = \frac{b+a}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

$$M(t) = \begin{cases} 1 & \text{if } t = 0 \\ \frac{e^{tb} - e^{ta}}{t\,(b-a)}, & \text{if } t \neq 0 \end{cases}$$

**Proof:**

$$\begin{aligned} E(X) &= \int_a^b x\, f(x)\, dx \\ &= \int_a^b x\, \frac{1}{b-a}\, dx \\ &= \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b \\ &= \frac{1}{2}\,(b+a). \end{aligned}$$

$$\begin{aligned}
E(X^2) &= \int_a^b x^2 \, f(x) \, dx \\
&= \int_a^b x^2 \, \frac{1}{b-a} \, dx \\
&= \frac{1}{b-a} \, \left[ \frac{x^3}{3} \right]_a^b \\
&= \frac{1}{b-a} \, \frac{b^3 - a^3}{3} \\
&= \frac{1}{(b-a)} \, \frac{(b-a)\,(b^2+ba+a^2)}{3} \\
&= \frac{1}{3} \, (b^2 + ba + a^2).
\end{aligned}$$

Hence, the variance of $X$ is given by

$$\begin{aligned}
Var(X) &= E(X^2) - (\, E(X)\,)^2 \\
&= \frac{1}{3} \, (b^2 + ba + a^2) - \frac{(b+a)^2}{4} \\
&= \frac{1}{12} \, \left[ 4b^2 + 4ba + 4a^2 - 3a^2 - 3b^2 - 6ba \right] \\
&= \frac{1}{12} \, \left[ b^2 - 2ba + a^2 \right] \\
&= \frac{1}{12} \, (b-a)^2 \, .
\end{aligned}$$

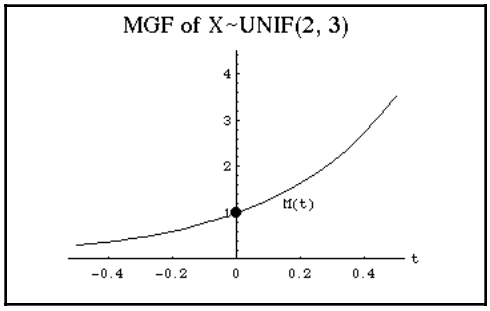
Next, we compute the moment generating function of $X$. First, we handle the case $t \neq 0$. Assume $t \neq 0$. Hence

$$\begin{aligned}
M(t) &= E\left(e^{tX}\right) \\
&= \int_a^b e^{tx} \, \frac{1}{b-a} \, dx \\
&= \frac{1}{b-a} \, \left[ \frac{e^{tx}}{t} \right]_a^b \\
&= \frac{e^{tb} - e^{ta}}{t\,(b-a)}.
\end{aligned}$$

If $t = 0$, we have know that $M(0) = 1$, hence we get

$$M(t) = \begin{cases} 1 & \text{if } t = 0 \\ \frac{e^{tb} - e^{ta}}{t\,(b-a)}, & \text{if } t \neq 0 \end{cases}$$

and this completes the proof.




**Example 6.1.** Suppose $Y \sim UNIF(0,1)$ and $Y = \frac{1}{4}X^2$. What is the probability density function of $X$?

**Answer:** We shall find the probability density function of $X$ through the cumulative distribution function of $Y$. The cumulative distribution function of $X$ is given by

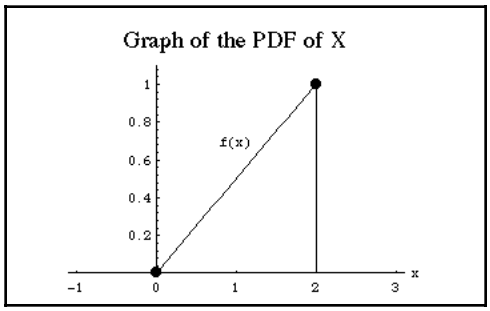
$$\begin{aligned} F(x) &= P\left(X \leq x\right) \\ &= P\left(X^2 \leq x^2\right) \\ &= P\left(\frac{1}{4}X^2 \leq \frac{1}{4}x^2\right) \\ &= P\left(Y \leq \frac{x^2}{4}\right) \\ &= \int_0^{\frac{x^2}{4}} f(y)\,dy \\ &= \int_0^{\frac{x^2}{4}} dy \\ &= \frac{x^2}{4}. \end{aligned}$$

Thus

$$f(x) = \frac{d}{dx}F(x) = \frac{x}{2}.$$

Hence the probability density function of $X$ is given by

$$f(x) = \begin{cases} \frac{x}{2} & \text{for } 0 \leq x \leq 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$




**Example 6.2.** If $X$ has a uniform distribution on the interval from 0 to 10, then what is $P\left(X + \frac{10}{X} \geq 7\right)$?

**Answer:** Since $X \sim UNIF(0, 10)$, the probability density function of $X$ is $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. Hence

$$\begin{aligned} P\left(X + \frac{10}{X} \geq 7\right) &= P\left(X^2 + 10 \geq 7\,X\right) \\ &= P\left(X^2 - 7\,X + 10 \geq 0\right) \\ &= P\left((X-5)\,(X-2) \geq 0\right) \\ &= P\left(X \leq 2 \text{ or } X \geq 5\right) \\ &= 1 - P\left(2 \leq X \leq 5\right) \\ &= 1 - \int_2^5 f(x)\,dx \\ &= 1 - \int_2^5 \frac{1}{10}\,dx \\ &= 1 - \frac{3}{10} = \frac{7}{10}. \end{aligned}$$

**Example 6.3.** If $X$ is uniform on the interval from 0 to 3, what is the probability that the quadratic equation $4t^2 + 4tX + X + 2 = 0$ has real solutions?

**Answer:** Since $X \sim UNIF(0, 3)$, the probability density function of $X$ is

$$f(x) = \begin{cases} \frac{1}{3} & 0 \leq x \leq 3 \\ \\ 0 & \text{otherwise.} \end{cases}$$

The quadratic equation $4t^2 + 4tX + X + 2 = 0$ has real solution if the discriminant of this equation is positive. That is

$$16X^2 - 16(X + 2) \geq 0,$$

which is

$$X^2 - X - 2 \geq 0.$$

From this, we get

$$(X - 2)(X + 1) \geq 0.$$

The probability that the quadratic equation $4t^2 + 4tX + X + 2 = 0$ has real roots is equivalent to

$$\begin{aligned} P\left((X-2)(X+1) \geq 0\right) &= P(X \leq -1 \text{ or } X \geq 2) \\ &= P(X \leq -1) + P(X \geq 2) \\ &= \int_{-\infty}^{-1} f(x)\,dx + \int_{2}^{3} f(x)\,dx \\ &= 0 + \int_{2}^{3} \frac{1}{3}\,dx \\ &= \frac{1}{3} = 0.3333. \end{aligned}$$

**Theorem 6.2.** If $X$ is a continuous random variable with a strictly increasing cumulative distribution function $F(x)$, then the random variable $Y$, defined by

$$Y = F(X)$$

has the uniform distribution on the interval [0, 1].

**Proof:** Since $F$ is strictly increasing, the inverse $F^{-1}(x)$ of $F(x)$ exists. We want to show that the probability density function $g(y)$ of $Y$ is $g(y) = 1$. First, we find the cumulative distribution $G(y)$ function of $Y$.

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(F(X) \leq y) \\ &= P\left(X \leq F^{-1}(y)\right) \\ &= F\left(F^{-1}(y)\right) \\ &= y. \end{aligned}$$

Hence the probability density function of $Y$ is given by

$$g(y) = \frac{d}{dy}G(y) = \frac{d}{dy}y = 1.$$

The following problem can be solved using this theorem but we solve it without this theorem.

**Example 6.4.** If the probability density function of $X$ is

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}, \qquad -\infty < x < \infty,$$

then what is the probability density function of $Y = \frac{1}{1+e^{-X}}$?

**Answer:** The cumulative distribution function of $Y$ is given by

$$\begin{aligned}
G(y) &= P\left(Y \leq y\right) \\
&= P\left(\frac{1}{1+e^{-X}} \leq y\right) \\
&= P\left(1+e^{-X} \geq \frac{1}{y}\right) \\
&= P\left(e^{-X} \geq \frac{1-y}{y}\right) \\
&= P\left(-X \geq \ln\frac{1-y}{y}\right) \\
&= P\left(X \leq -\ln\frac{1-y}{y}\right) \\
&= \int_{-\infty}^{-\ln\frac{1-y}{y}} \frac{e^{-x}}{(1+e^{-x})^2}\,dx \\
&= \left[\frac{1}{1+e^{-x}}\right]_{-\infty}^{-\ln\frac{1-y}{y}} \\
&= \frac{1}{1+\frac{1-y}{y}} \\
&= y.
\end{aligned}$$

Hence, the probability density function of $Y$ is given by

$$f(y) = \begin{cases} 1 & \text{if } 0 < y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

**Example 6.5.** A box to be constructed so that its height is 10 inches and its base is $X$ inches by $X$ inches. If $X$ has a uniform distribution over the interval $(2, 8)$, then what is the expected volume of the box in cubic inches?

**Answer:** Since $X \sim UNIF(2, 8)$,

$$f(x) = \frac{1}{8-2} = \frac{1}{6} \qquad \text{on} \quad (2, 8).$$

The volume $V$ of the box is

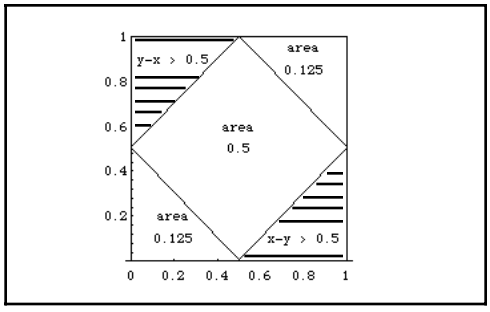
$$V = 10\,X^2.$$

Hence

$$\begin{aligned} E(V) &= E\left(10X^2\right) \\ &= 10\,E\left(X^2\right) \\ &= 10\int_2^8 x^2 \frac{1}{6}\,dx \\ &= \frac{10}{6}\left[\frac{x^3}{3}\right]_2^8 \\ &= \frac{10}{18}\left[\,8^3 - 2^3\,\right] = (5)\,(8)\,(7) = 280 \text{ cubic inches.} \end{aligned}$$

**Example 6.6.** Two numbers are chosen independently and at random from the interval $(0, 1)$. What is the probability that the two numbers differs by more than $\frac{1}{2}$?

**Answer:** See figure below:




Choose $x$ from the $x$-axis between 0 and 1, and choose $y$ from the $y$-axis between 0 and 1. The probability that the two numbers differ by more than

$\frac{1}{2}$ is equal to the area of the shaded region. Thus

$$P\left(|X-Y|>\frac{1}{2}\right)=\frac{\frac{1}{8}+\frac{1}{8}}{1}=\frac{1}{4}.$$

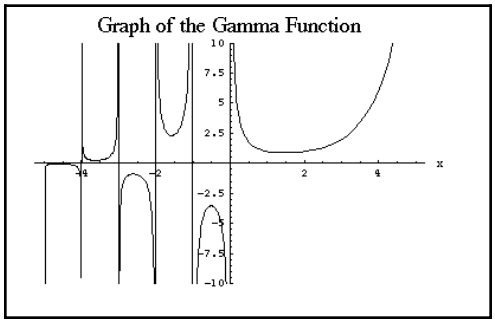
### 6.2. Gamma Distribution

The gamma distribution involves the notion of gamma function. First, we develop the notion of gamma function and study some of its well known properties. The gamma function, $\Gamma(z)$, is a generalization of the notion of factorial. The gamma function is defined as

$$\Gamma(z):=\int_0^\infty x^{z-1}\,e^{-x}\,dx,$$

where $z$ is positive real number (that is, $z>0$). The condition $z>0$ is assumed for the convergence of the integral. Although the integral does not converge for $z<0$, it can be shown by using an alternative definition of gamma function that it is defined for all $z\in\mathbb{R}\setminus\{0,-1,-2,-3,...\}$.

The integral on the right side of the above expression is called Euler's second integral, after the Swiss mathematician Leonhard Euler (1707-1783). The graph of the gamma function is shown below. Observe that the zero and negative integers correspond to vertical asymptotes of the graph of gamma function.




**Lemma 6.1.** $\Gamma(1)=1$.

**Proof:**

$$\Gamma(1)=\int_0^\infty x^0\,e^{-x}\,dx=\left[-e^{-x}\right]_0^\infty=1.$$

**Lemma 6.2.** The gamma function $\Gamma(z)$ satisfies the functional equation $\Gamma(z)=(z-1)\,\Gamma(z-1)$ for all real number $z>1$.

**Proof:** Let $z$ be a real number such that $z > 1$, and consider

$$\begin{aligned}\Gamma(z) &= \int_0^\infty x^{z-1}\, e^{-x}\, dx \\ &= \left[-x^{z-1}\, e^{-x}\right]_0^\infty + \int_0^\infty (z-1)\, x^{z-2}\, e^{-x}\, dx \\ &= (z-1)\int_0^\infty x^{z-2}\, e^{-x}\, dx \\ &= (z-1)\,\Gamma(z-1).\end{aligned}$$

Although, we have proved this lemma for all real $z > 1$, actually this lemma holds also for all real number $z \in \mathbb{R} \setminus \{1, 0, -1, -2, -3, ...\}$.

**Lemma 6.3.** $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

**Proof:** We want to show that

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-x}}{\sqrt{x}}\, dx$$

is equal to $\sqrt{\pi}$. We substitute $y = \sqrt{x}$, hence the above integral becomes

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty \frac{e^{-x}}{\sqrt{x}}\, dx \\ &= 2\int_0^\infty e^{-y^2}\, dy, \qquad \text{where } y = \sqrt{x}.\end{aligned}$$

Hence

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^\infty e^{-u^2}\, du$$

and also

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^\infty e^{-v^2}\, dv.$$

Multiplying the above two expressions, we get

$$\left(\Gamma\left(\frac{1}{2}\right)\right)^2 = 4\int_0^\infty \int_0^\infty e^{-(u^2+v^2)}\, du\, dv.$$

Now we change the integral into polar form by the transformation $u = r\cos(\theta)$ and $v = r\sin(\theta)$. The Jacobian of the transformation is

$$\begin{aligned}J(r,\theta) &= det\begin{pmatrix} \frac{\partial u}{\partial r} & \frac{\partial u}{\partial \theta} \\ \\ \frac{\partial v}{\partial r} & \frac{\partial v}{\partial \theta} \end{pmatrix} \\ &= det\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} \\ &= r\cos^2(\theta) + r\sin^2(\theta) \\ &= r.\end{aligned}$$

Hence, we get

$$\begin{aligned}\left(\Gamma\left(\frac{1}{2}\right)\right)^2 &= 4\int_0^{\frac{\pi}{2}}\int_0^{\infty} e^{-r^2}\, J(r,\theta)\, dr\, d\theta \\ &= 4\int_0^{\frac{\pi}{2}}\int_0^{\infty} e^{-r^2}\, r\, dr\, d\theta \\ &= 2\int_0^{\frac{\pi}{2}}\int_0^{\infty} e^{-r^2}\, 2r\, dr\, d\theta \\ &= 2\int_0^{\frac{\pi}{2}}\left[\int_0^{\infty} e^{-r^2}\, dr^2\right] d\theta \\ &= 2\int_0^{\frac{\pi}{2}} \Gamma(1)\, d\theta \\ &= \pi.\end{aligned}$$

Therefore, we get

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

**Lemma 6.4.** $\Gamma\left(-\frac{1}{2}\right) = -2\sqrt{\pi}$.

**Proof:** By Lemma 6.2, we get

$$\Gamma(z) = (z-1)\,\Gamma(z-1)$$

for all $z \in \mathbb{R} \setminus \{1, 0, -1, -2, -3, ...\}$. Letting $z = \frac{1}{2}$, we get

$$\Gamma\left(\frac{1}{2}\right) = \left(\frac{1}{2} - 1\right)\Gamma\left(\frac{1}{2} - 1\right)$$

which is

$$\Gamma\left(-\frac{1}{2}\right) = -2\,\Gamma\left(\frac{1}{2}\right) = -2\sqrt{\pi}.$$

**Example 6.7.** Evaluate $\Gamma\left(\frac{5}{2}\right)$.

**Answer:**

$$\Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\,\frac{1}{2}\,\Gamma\left(\frac{1}{2}\right) = \frac{3}{4}\,\sqrt{\pi}.$$

**Example 6.8.** Evaluate $\Gamma\left(-\frac{7}{2}\right)$.

**Answer:** Consider

$$\begin{aligned}\Gamma\left(-\frac{1}{2}\right) &= -\frac{3}{2}\,\Gamma\left(-\frac{3}{2}\right)\\ &= \left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\Gamma\left(-\frac{5}{2}\right)\\ &= \left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\left(-\frac{7}{2}\right)\Gamma\left(-\frac{7}{2}\right).\end{aligned}$$

Hence

$$\Gamma\left(-\frac{7}{2}\right) = \left(-\frac{2}{3}\right)\left(-\frac{2}{5}\right)\left(-\frac{2}{7}\right)\Gamma\left(-\frac{1}{2}\right) = \frac{16}{105}\,\sqrt{\pi}.$$

**Example 6.9.** Evaluate $\Gamma\,(7.8)$.

**Answer:**

$$\begin{aligned}\Gamma\,(7.8) &= (6.8)\,(5.8)\,(4.8)\,(3.8)\,(2.8)\,(1.8)\,\Gamma\,(1.8)\\ &= (3625.7)\,\Gamma\,(1.8)\\ &= (3625.7)\,(0.9314) = 3376.9.\end{aligned}$$

Here we have used the gamma table to find $\Gamma\,(1.8)$ to be 0.9314.

**Example 6.10.** If $n$ is a natural number, then $\Gamma(n+1) = n!$.

**Answer:**

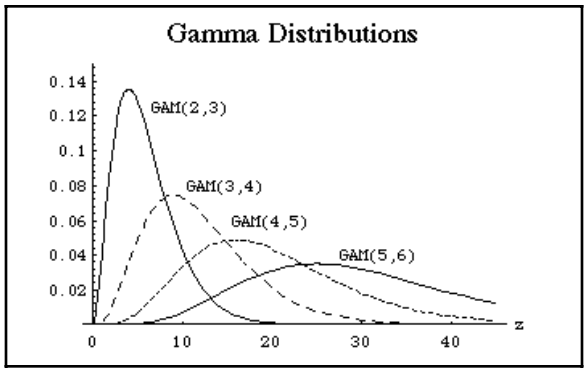
$$\begin{aligned}\Gamma(n+1) &= n\,\Gamma(n)\\ &= n\,(n-1)\,\Gamma(n-1)\\ &= n\,(n-1)\,(n-2)\,\Gamma(n-2)\\ &= \cdots \quad \cdots\\ &= n\,(n-1)\,(n-2)\cdots(1)\,\Gamma(1)\\ &= n!\end{aligned}$$

Now we are ready to define the gamma distribution.

**Definition 6.2.** A continuous random variable $X$ is said to have a gamma distribution if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\,\theta^{\alpha}}\,x^{\alpha-1}\,e^{-\frac{x}{\theta}} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,}\end{cases}$$

where $\alpha > 0$ and $\theta > 0$. We denote a random variable with gamma distribution as $X \sim GAM(\theta,\, \alpha)$. The following diagram shows the graph of the gamma density for various values of values of the parameters $\theta$ and $\alpha$.




The following theorem gives the expected value, the variance, and the moment generating function of the gamma random variable

**Theorem 6.3.** If $X \sim GAM(\theta, \alpha)$, then

$$E(X) = \theta\,\alpha$$

$$Var(X) = \theta^2\,\alpha$$

$$M(t) = \left(\frac{1}{1-\theta\,t}\right)^{\alpha}, \qquad \text{if} \qquad t < \frac{1}{\theta}.$$

**Proof:** First, we derive the moment generating function of $X$ and then we compute the mean and variance of it. The moment generating function

$$\begin{aligned}
M(t) &= E\left(e^{tX}\right) \\
&= \int_0^{\infty} \frac{1}{\Gamma(\alpha)\,\theta^{\alpha}}\, x^{\alpha-1}\, e^{-\frac{x}{\theta}}\, e^{tx}\, dx \\
&= \int_0^{\infty} \frac{1}{\Gamma(\alpha)\,\theta^{\alpha}}\, x^{\alpha-1}\, e^{-\frac{1}{\theta}(1-\theta t)x}\, dx \\
&= \int_0^{\infty} \frac{1}{\Gamma(\alpha)\,\theta^{\alpha}}\, \frac{\theta^{\alpha}}{(1-\theta t)^{\alpha}}\, y^{\alpha-1}\, e^{-y}\, dy, \qquad \text{where} \quad y = \frac{1}{\theta}\,(1-\theta t)x \\
&= \frac{1}{(1-\theta t)^{\alpha}} \int_0^{\infty} \frac{1}{\Gamma(\alpha)}\, y^{\alpha-1}\, e^{-y}\, dy \\
&= \frac{1}{(1-\theta t)^{\alpha}}, \qquad \text{since the integrand is } GAM(1,\alpha).
\end{aligned}$$

The first derivative of the moment generating function is

$$\begin{aligned} M'(t) &= \frac{d}{dt}(1-\theta t)^{-\alpha} \\ &= (-\alpha)\,(1-\theta t)^{-\alpha-1}\,(-\theta) \\ &= \alpha\,\theta\,(1-\theta t)^{-(\alpha+1)}. \end{aligned}$$

Hence from above, we find the expected value of $X$ to be

$$E(X) = M'(0) = \alpha\,\theta.$$

Similarly,

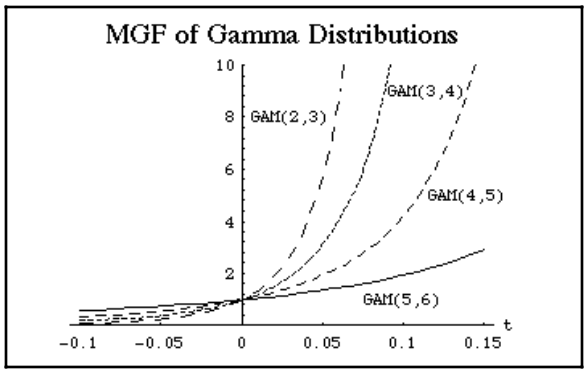
$$\begin{aligned} M''(t) &= \frac{d}{dt}\left(\alpha\,\theta\,(1-\theta t)^{-(\alpha+1)}\right) \\ &= \alpha\,\theta\,(\alpha+1)\,\theta\,(1-\theta t)^{-(\alpha+2)} \\ &= \alpha\,(\alpha+1)\,\theta^2\,(1-\theta t)^{-(\alpha+2)}. \end{aligned}$$

Thus, the variance of $X$ is

$$\begin{aligned} Var(X) &= M''(0) - (M'(0))^2 \\ &= \alpha\,(\alpha+1)\,\theta^2 - \alpha^2\,\theta^2 \\ &= \alpha\,\theta^2 \end{aligned}$$

and proof of the theorem is now complete

In figure below the graphs of moment generating function for various values of the parameters are illustrated.




**Example 6.11.** Let $X$ have the density function

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\,\theta^{\alpha}}\,x^{\alpha-1}\,e^{-\frac{x}{\theta}} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $\theta > 0$. If $\alpha = 4$, what is the mean of $\frac{1}{X^3}$?

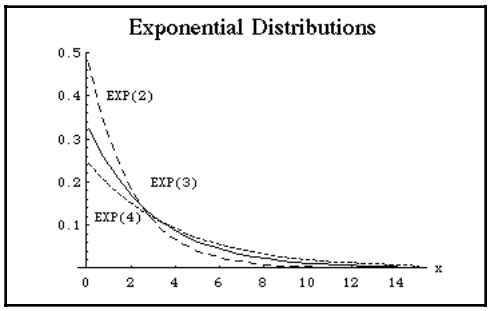
**Answer:**

$$\begin{aligned} E\left(X^{-3}\right) &= \int_0^\infty \frac{1}{x^3}\, f(x)\, dx \\ &= \int_0^\infty \frac{1}{x^3}\, \frac{1}{\Gamma(4)\,\theta^4}\, x^3\, e^{-\frac{x}{\theta}}\, dx \\ &= \frac{1}{3!\,\theta^4} \int_0^\infty e^{-\frac{x}{\theta}}\, dx \\ &= \frac{1}{3!\,\theta^3} \int_0^\infty \frac{1}{\theta}\, e^{-\frac{x}{\theta}}\, dx \\ &= \frac{1}{3!\,\theta^3} \qquad \text{since the integrand is GAM}(\theta, 1). \end{aligned}$$

**Definition 6.3.** A continuous random variable is said to be an exponential random variable with parameter $\theta$ if its probability density function is of the form

$$f(x) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. If a random variable $X$ has an exponential density function with parameter $\theta$, then we denote it by writing $X \sim EXP(\theta)$.

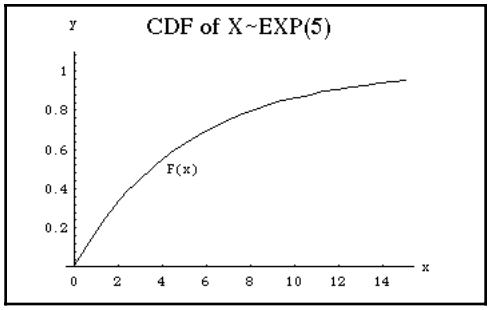



An exponential distribution is a special case of the gamma distribution. If the parameter $\alpha = 1$, then the gamma distribution reduces to the exponential distribution. Hence most of the information about an exponential distribution can be obtained from the gamma distribution.

**Example 6.12.** What is the cumulative density function of a random variable which has an exponential distribution with variance 25?

**Answer:** Since an exponential distribution is a special case of the gamma distribution with $\alpha = 1$, from Theorem 6.3, we get $Var(X) = \theta^2$. But this is given to be 25. Thus, $\theta^2 = 25$ or $\theta = 5$. Hence, the probability density function of $X$ is

$$\begin{aligned} F(x) &= \int_0^x f(t)\,dt \\ &= \int_0^x \frac{1}{5}\, e^{-\frac{t}{5}}\, dt \\ &= \frac{1}{5}\, \left[-5\, e^{-\frac{t}{5}}\right]_0^x \\ &= 1 - e^{-\frac{x}{5}}. \end{aligned}$$




**Example 6.13.** If the random variable $X$ has a gamma distribution with parameters $\alpha = 1$ and $\theta = 1$, then what is the probability that $X$ is between its mean and median?

**Answer:** Since $X \sim GAM(1, 1)$, the probability density function of $X$ is

$$f(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the median $q$ of $X$ can be calculated from

$$\begin{aligned} \frac{1}{2} &= \int_0^q e^{-x}\, dx \\ &= \left[-e^{-x}\right]_0^q \\ &= 1 - e^{-q}. \end{aligned}$$

Hence

$$\frac{1}{2} = 1 - e^{-q}$$

and from this, we get

$$q = \ln 2.$$

The mean of $X$ can be found from the Theorem 6.3.

$$E(X) = \alpha\,\theta = 1.$$

Hence the mean of $X$ is 1 and the median of $X$ is $\ln 2$. Thus

$$\begin{aligned} P(\ln 2 \le X \le 1) &= \int_{\ln 2}^{1} e^{-x}\,dx \\ &= \left[-e^{-x}\right]_{\ln 2}^{1} \\ &= e^{-\ln 2} - \frac{1}{e} \\ &= \frac{1}{2} - \frac{1}{e} \\ &= \frac{e-2}{2\,e}. \end{aligned}$$

**Example 6.14.** If the random variable $X$ has a gamma distribution with parameters $\alpha = 1$ and $\theta = 2$, then what is the probability density function of the random variable $Y = e^X$?

**Answer:** First, we calculate the cumulative distribution function $G(y)$ of $Y$.

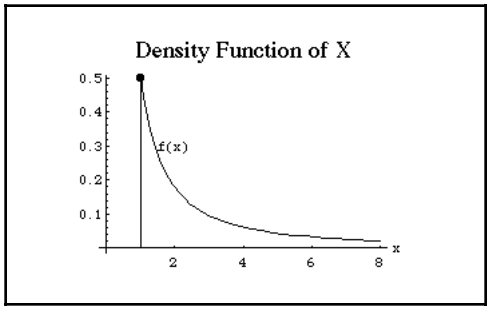
$$\begin{aligned} G(y) &= P\,(\,Y \le y\,) \\ &= P\left(\,e^X \le y\,\right) \\ &= P\,(\,X \le \ln y\,) \\ &= \int_0^{\ln y} \frac{1}{2}\,e^{-\frac{x}{2}}\,dx \\ &= \frac{1}{2}\,\left[-2\,e^{-\frac{x}{2}}\right]_0^{\ln y} \\ &= 1 - \frac{1}{e^{\frac{1}{2}\ln y}} \\ &= 1 - \frac{1}{\sqrt{y}}. \end{aligned}$$

Hence, the probability density function of $Y$ is given by

$$g(y) = \frac{d}{dy}G(y) = \frac{d}{dy}\left(1 - \frac{1}{\sqrt{y}}\right) = \frac{1}{2y\sqrt{y}}.$$

Thus, if $X \sim GAM(1,2)$, then probability density function of $e^X$ is

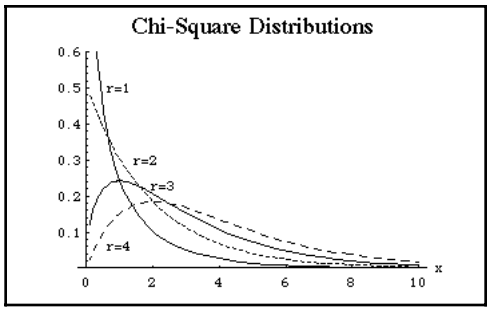
$$f(x) = \begin{cases} \frac{1}{2x\sqrt{x}} & \text{if} \quad 1 \leq x < \infty \\ 0 & \text{otherwise.} \end{cases}$$




**Definition 6.4.** A continuous random variable $X$ is said to have a chi-square distribution with $r$ degrees of freedom if its probability density function is of the form

$$f(x) = \begin{cases} \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}} \, x^{\frac{r}{2}-1} \, e^{-\frac{x}{2}} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $r > 0$. If $X$ has a chi-square distribution, then we denote it by writing $X \sim \chi^2(r)$.




The gamma distribution reduces to the chi-square distribution if $\alpha = \frac{r}{2}$ and $\theta = 2$. Thus, the chi-square distribution is a special case of the gamma distribution. Further, if $r \to \infty$, then the chi-square distribution tends to the normal distribution.

The chi-square distribution was originated in the works of British Statistician Karl Pearson (1857-1936) but it was originally discovered by German physicist F. R. Helmert (1843-1917).

**Example 6.15.** If $X \sim GAM(1,1)$, then what is the probability density function of the random variable $2X$?

**Answer:** We will use the moment generating method to find the distribution of $2X$. The moment generating function of a gamma random variable is given by (see Theorem 6.3)

$$M(t) = (1 - \theta\, t)^{-\alpha}, \qquad \text{if} \qquad t < \frac{1}{\theta}.$$

Since $X \sim GAM(1,1)$, the moment generating function of $X$ is given by

$$M_X(t) = \frac{1}{1-t}, \qquad t < 1.$$

Hence, the moment generating function of $2X$ is

$$\begin{aligned} M_{2X}(t) &= M_X(2t) \\ &= \frac{1}{1-2t} \\ &= \frac{1}{(1-2t)^{\frac{2}{2}}} \\ &= \text{MGF of } \chi^2(2). \end{aligned}$$

Hence, if $X$ is an exponential with parameter 1, then $2X$ is chi-square with 2 degrees of freedom.

**Example 6.16.** If $X \sim \chi^2(5)$, then what is the probability that $X$ is between 1.145 and 12.83?

**Answer:** The probability of $X$ between 1.145 and 12.83 can be calculated from the following:

$$\begin{aligned} &P(1.145 \leq X \leq 12.83) \\ &\qquad = P(X \leq 12.83) - P(X \leq 1.145) \\ &\qquad = \int_0^{12.83} f(x)\, dx - \int_0^{1.145} f(x)\, dx \\ &\qquad = \int_0^{12.83} \frac{1}{\Gamma\left(\frac{5}{2}\right) 2^{\frac{5}{2}}}\, x^{\frac{5}{2}-1}\, e^{-\frac{x}{2}}\, dx - \int_0^{1.145} \frac{1}{\Gamma\left(\frac{5}{2}\right) 2^{\frac{5}{2}}}\, x^{\frac{5}{2}-1}\, e^{-\frac{x}{2}}\, dx \\ &\qquad = 0.975 - 0.050 \qquad \text{(from } \chi^2 \text{ table)} \\ &\qquad = 0.925. \end{aligned}$$

These integrals are hard to evaluate and so their values are taken from the chi-square table.

**Example 6.17.** If $X \sim \chi^2(7)$, then what are values of the constants $a$ and $b$ such that $P(a < X < b) = 0.95$?

**Answer:** Since

$$0.95 = P(a < X < b) = P(X < b) - P(X < a),$$

we get

$$P(X < b) = 0.95 + P(X < a).$$

We choose $a = 1.690$, so that

$$P(X < 1.690) = 0.025.$$

From this, we get

$$P(X < b) = 0.95 + 0.025 = 0.975$$

Thus, from the chi-square table, we get $b = 16.01$.

**Definition 6.5.** A continuous random variable $X$ is said to have a $n$-Erlang distribution if its probability density function is of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!}, & \text{if } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda > 0$ is a parameter.

The gamma distribution reduces to $n$-Erlang distribution if $\alpha = n$, where $n$ is a positive integer, and $\theta = \frac{1}{\lambda}$. The gamma distribution can be generalized to include the Weibull distribution. We call this generalized distribution the unified distribution. The form of this distribution is the following:

$$f(x) = \begin{cases} \frac{\alpha}{\theta^{\alpha\psi}\Gamma(\alpha^{\psi}+1)} x^{\alpha-1} e^{\frac{-x^{-(\alpha^{\psi}-\alpha-1)}}{\theta}}, & \text{if } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$, $\alpha > 0$, and $\psi \in \{0, 1\}$ are parameters.

If $\psi = 0$, the unified distribution reduces

$$f(x) = \begin{cases} \frac{\alpha}{\theta} x^{\alpha-1} e^{-\frac{x^{\alpha}}{\theta}}, & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

which is known as the Weibull distribution. For $\alpha = 1$, the Weibull distribution becomes an exponential distribution. The Weibull distribution provides probabilistic models for life-length data of components or systems. The mean and variance of the Weibull distribution are given by

$$E(X) = \theta^{\frac{1}{\alpha}} \Gamma\left(1 + \frac{1}{\alpha}\right),$$
$$Var(X) = \theta^{\frac{2}{\alpha}} \left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\left(1 + \frac{1}{\alpha}\right)\right]^2\right\}.$$

From this Weibull distribution, one can get the Rayleigh distribution by taking $\theta = 2\sigma^2$ and $\alpha = 2$. The Rayleigh distribution is given by

$$f(x) = \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, & \text{if } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

If $\psi = 1$, the unified distribution reduces to the gamma distribution.

### 6.3. Beta Distribution

The beta distribution is one of the basic distributions in statistics. It has many applications in classical as well as Bayesian statistics. It is a versatile distribution and as such it is used in modeling the behavior of random variables that are positive but bounded in possible values. Proportions and percentages fall in this category.

The beta distribution involves the notion of beta function. First we explain the notion of the beta integral and some of its simple properties. Let $\alpha$ and $\beta$ be any two positive real numbers. The beta function $B(\alpha, \beta)$ is defined as

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

First, we prove a theorem that establishes the connection between the beta function and the gamma function.

**Theorem 6.4.** Let $\alpha$ and $\beta$ be any two positive real numbers. Then

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

where

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

is the gamma function.

**Proof:** We prove this theorem by computing

$$\begin{aligned}
\Gamma(\alpha)\,\Gamma(\beta) &= \left(\int_0^\infty x^{\alpha-1}e^{-x}dx\right)\left(\int_0^\infty y^{\beta-1}e^{-y}dy\right)\\
&= \left(\int_0^\infty u^{2\alpha-2}e^{-u^2}2udu\right)\left(\int_0^\infty v^{2\beta-2}e^{-v^2}2vdv\right)\\
&= 4\int_0^\infty\int_0^\infty u^{2\alpha-1}v^{2\beta-1}e^{-(u^2+v^2)}dudv\\
&= 4\int_0^{\frac{\pi}{2}}\int_0^\infty r^{2\alpha+2\beta-2}(\cos\theta)^{2\alpha-1}(\sin\theta)^{2\beta-1}e^{-r^2}rdrd\theta\\
&= \left(\int_0^\infty (r^2)^{\alpha+\beta-1}e^{-r^2}dr^2\right)\left(2\int_0^{\frac{\pi}{2}}(\cos\theta)^{2\alpha-1}(\sin\theta)^{2\beta-1}d\theta\right)\\
&= \Gamma(\alpha+\beta)\left(2\int_0^{\frac{\pi}{2}}(\cos\theta)^{2\alpha-1}(\sin\theta)^{2\beta-1}d\theta\right)\\
&= \Gamma(\alpha+\beta)\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt\\
&= \Gamma(\alpha+\beta)\,B(\alpha,\beta).
\end{aligned}$$

The second line in the above integral is obtained by substituting $x = u^2$ and $y = v^2$. Similarly, the fourth and seventh lines are obtained by substituting $u = r\cos\theta$, $v = r\sin\theta$, and $t = \cos^2\theta$, respectively. This proves the theorem.

The following two corollaries are consequences of the last theorem.

**Corollary 6.1.** For every positive $\alpha$ and $\beta$, the beta function is symmetric, that is

$$B(\alpha,\beta) = B(\beta,\alpha).$$

**Corollary 6.2.** For every positive $\alpha$ and $\beta$, the beta function can be written as

$$B(\alpha,\beta) = 2\int_0^{\frac{\pi}{2}}(\cos\theta)^{2\alpha-1}(\sin\theta)^{2\beta-1}d\theta.$$

The following corollary is obtained substituting $s = \frac{t}{1-t}$ in the definition of the beta function.

**Corollary 6.3.** For every positive $\alpha$ and $\beta$, the beta function can be expressed as

$$B(\alpha,\beta) = \int_0^\infty \frac{s^{\alpha-1}}{(1+s)^{\alpha+\beta}}\,ds.$$

Using Theorem 6.4 and the property of gamma function, we have the following corollary.

**Corollary 6.4.** For every positive real number $\beta$ and every positive integer $\alpha$, the beta function reduces to

$$B(\alpha, \beta) = \frac{(\alpha - 1)!}{(\alpha - 1 + \beta)(\alpha - 2 + \beta) \cdots (1 + \beta)\beta}.$$

**Corollary 6.5.** For every pair of positive integers $\alpha$ and $\beta$, the beta function satisfies the following recursive relation

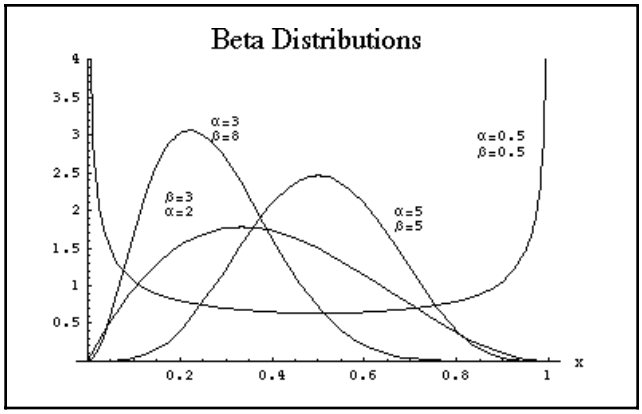
$$B(\alpha, \beta) = \frac{(\alpha - 1)(\beta - 1)}{(\alpha + \beta - 1)(\alpha + \beta - 2)} B(\alpha - 1, \beta - 1).$$

**Definition 6.6.** A random variable $X$ is said to have the beta density function if its probability density function is of the form

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

for every positive $\alpha$ and $\beta$. If $X$ has a beta distribution, then we symbolically denote this by writing $X \sim BETA(\alpha, \beta)$.

The following figure illustrates the graph of the beta distribution for various values of $\alpha$ and $\beta$.




The beta distribution reduces to the uniform distribution over $(0, 1)$ if $\alpha = 1 = \beta$. The following theorem gives the mean and variance of the beta distribution.

**Theorem 6.5.** If $X \sim BETA(\alpha, \beta)$,

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}.$$

**Proof:** The expected value of $X$ is given by

$$\begin{aligned} E(X) &= \int_0^1 x\, f(x)\, dx \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}\,(1-x)^{\beta-1}\, dx \\ &= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\ &= \frac{\Gamma(\alpha+1)\,\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}\ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \\ &= \frac{\alpha\,\Gamma(\alpha)\,\Gamma(\beta)}{(\alpha+\beta)\,\Gamma(\alpha+\beta)}\ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \\ &= \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

Similarly, we can show that

$$E\left(X^2\right) = \frac{\alpha\,(\alpha+1)}{(\alpha+\beta+1)\,(\alpha+\beta)}.$$

Therefore

$$Var(X) = E\left(X^2\right) - E(X) = \frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}$$

and the proof of the theorem is now complete.

**Example 6.18.** The percentage of impurities per batch in a certain chemical product is a random variable $X$ that follows the beta distribution given by

$$f(x) = \begin{cases} 60\,x^3\,(1-x)^2 & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that a randomly selected batch will have more than 25% impurities?

**Proof:** The probability that a randomly selected batch will have more than 25% impurities is given by

$$\begin{aligned} P(X \geq 0.25) &= \int_{0.25}^{1} 60\, x^3\, (1-x)^2\, dx \\ &= 60 \int_{0.25}^{1} \left(x^3 - 2x^4 + x^5\right)\, dx \\ &= 60 \left[\frac{x^4}{4} - \frac{2x^5}{5} + \frac{x^6}{6}\right]_{0.25}^{1} \\ &= 60\, \frac{657}{40960} = 0.9624. \end{aligned}$$

**Example 6.19.** The proportion of time per day that all checkout counters in a supermarket are busy follows a distribution

$$f(x) = \begin{cases} k\, x^2\, (1-x)^9 & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the value of the constant $k$ so that $f(x)$ is a valid probability density function?

**Proof:** Using the definition of the beta function, we get that

$$\int_0^1 x^2\, (1-x)^9\, dx = B(3, 10).$$

Hence by Theorem 6.4, we obtain

$$B(3, 10) = \frac{\Gamma(3)\, \Gamma(10)}{\Gamma(13)} = \frac{1}{660}.$$

Hence $k$ should be equal to 660.

The beta distribution can be generalized to any bounded interval $[a, b]$. This generalized distribution is called the generalized beta distribution. If a random variable $X$ has this generalized beta distribution we denote it by writing $X \sim GBETA(\alpha, \beta, a, b)$. The probability density of the generalized beta distribution is given by

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)}\, \frac{(x-a)^{\alpha-1}\, (b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}} & \text{if } a < x < b \\ \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha, \beta, a > 0$.

If $X \sim GBETA(\alpha, \beta, a, b)$, then

$$E(X) = (b-a)\,\frac{\alpha}{\alpha+\beta} + a$$

$$Var(X) = (b-a)^2\,\frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}.$$

It can be shown that if $X = (b-a)Y + a$ and $Y \sim BETA(\alpha, \beta)$, then $X \sim GBETA(\alpha, \beta, a, b)$. Thus using Theorem 6.5, we get

$$E(X) = E((b-a)Y + a) = (b-a)E(Y) + a = (b-a)\,\frac{\alpha}{\alpha+\beta} + a$$

and

$$Var(X) = Var((b-a)Y+a) = (b-a)^2 Var(Y) = (b-a)^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

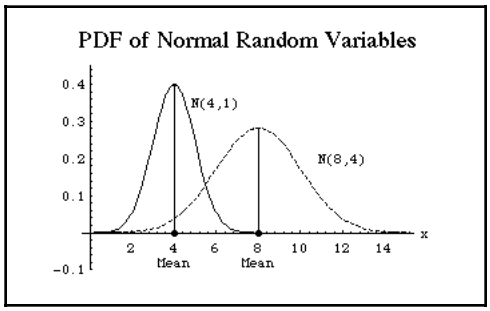
### 6.4. Normal Distribution

Among continuous probability distributions, the normal distribution is very well known since it arises in many applications. Normal distribution was discovered by a French mathematician Abraham DeMoivre (1667-1754). DeMoivre wrote two important books. One is called the *Annuities Upon Lives*, the first book on actuarial sciences and the second book is called the *Doctrine of Chances*, one of the early books on the probability theory. Pierre-Simon Laplace (1749-1827) applied normal distribution to astronomy. Carl Friedrich Gauss (1777-1855) used normal distribution in his studies of problems in physics and astronomy. Adolphe Quetelet (1796-1874) demonstrated that man's physical traits (such as height, chest expansion, weight etc.) as well as social traits follow normal distribution. The main importance of normal distribution lies on the central limit theorem which says that the sample mean has a normal distribution if the sample size is large.

**Definition 6.7.** A random variable $X$ is said to have a normal distribution if its probability density function is given by

$$f(x) = \frac{1}{\sigma\,\sqrt{2\,\pi}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty,$$

where $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ are arbitrary parameters. If $X$ has a normal distribution with parameters $\mu$ and $\sigma^2$, then we write $X \sim N(\mu, \sigma^2)$.




**Example 6.20.** Is the real valued function defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty$$

a probability density function of some random variable $X$?

**Answer:** To answer this question, we must check that $f$ is nonnegative and it integrates to 1. The nonnegative part is trivial since the exponential function is always positive. Hence using property of the gamma function, we show that $f$ integrates to 1 on $\mathbb{R}$.

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= 2\int_{\mu}^{\infty} \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \frac{2}{\sigma\sqrt{2\,\pi}} \int_0^{\infty} e^{-z}\, \frac{\sigma}{\sqrt{2\,z}}\, dz, \qquad \text{where} \quad z = \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$= \frac{1}{\sqrt{\pi}} \int_0^{\infty} \frac{1}{\sqrt{z}}\, e^{-z}\, dz$$

$$= \frac{1}{\sqrt{\pi}}\,\Gamma\left(\frac{1}{2}\right) = \frac{1}{\sqrt{\pi}}\,\sqrt{\pi} = 1.$$

The following theorem tells us that the parameter $\mu$ is the mean and the parameter $\sigma^2$ is the variance of the normal distribution.

**Theorem 6.6.** If $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} E(X) &= \mu \\ Var(X) &= \sigma^2 \\ M(t) &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \end{aligned}$$

**Proof:** We prove this theorem by first computing the moment generating function and finding out the mean and variance of $X$ from it.

$$\begin{aligned} M(t) &= E\left(e^{tX}\right) \\ &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(x^2-2\mu x+\mu^2\right)} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(x^2-2\mu x+\mu^2-2\sigma^2 tx\right)} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(x-\mu-\sigma^2 t\right)^2} e^{\mu t+\frac{1}{2}\sigma^2 t^2} dx \\ &= e^{\mu t+\frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(x-\mu-\sigma^2 t\right)^2} dx \\ &= e^{\mu t+\frac{1}{2}\sigma^2 t^2}. \end{aligned}$$

The last integral integrates to 1 because the integrand is the probability density function of a normal random variable whose mean is $\mu + \sigma^2 t$ and variance $\sigma^2$, that is $N(\mu + \sigma^2 t, \sigma^2)$. Finally, from the moment generating function one determines the mean and variance of the normal distribution. We leave this part to the reader.

**Example 6.21.** If $X$ is any random variable with mean $\mu$ and variance $\sigma^2 > 0$, then what are the mean and variance of the random variable $Y = \frac{X-\mu}{\sigma}$?

**Answer:** The mean of the random variable $Y$ is

$$\begin{aligned} E(Y) &= E\left(\frac{X-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma}\, E\,(X-\mu) \\ &= \frac{1}{\sigma}\, (E(X)-\mu) \\ &= \frac{1}{\sigma}\, (\mu-\mu) \\ &= 0. \end{aligned}$$

The variance of $Y$ is given by

$$\begin{aligned} Var(Y) &= Var\left(\frac{X-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma^2}\, Var\,(X-\mu) \\ &= \frac{1}{\sigma}\, Var(X) \\ &= \frac{1}{\sigma^2}\, \sigma^2 \\ &= 1. \end{aligned}$$

Hence, if we define a new random variable by taking a random variable and subtracting its mean from it and then dividing the resulting by its standard deviation, then this new random variable will have zero mean and unit variance.

**Definition 6.8.** A normal random variable is said to be standard normal, if its mean is zero and variance is one. We denote a standard normal random variable $X$ by $X \sim N(0,1)$.

The probability density function of standard normal distribution is the following:

$$f(x) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty.$$

**Example 6.22.** If $X \sim N(0,1)$, what is the probability of the random variable $X$ less than or equal to $-1.72$?

**Answer:**

$$\begin{aligned} P(X \leq -1.72) &= 1 - P(X \leq 1.72) \\ &= 1 - 0.9573 \qquad \text{(from table)} \\ &= 0.0427. \end{aligned}$$

**Example 6.23.** If $Z \sim N(0,1)$, what is the value of the constant $c$ such that $P(|Z| \leq c) = 0.95$?

**Answer:**

$$\begin{aligned} 0.95 &= P(|Z| \leq c) \\ &= P(-c \leq Z \leq c) \\ &= P(Z \leq c) - P(Z \leq -c) \\ &= 2\,P(Z \leq c) - 1. \end{aligned}$$

Hence

$$P(Z \leq c) = 0.975,$$

and from this using the table we get

$$c = 1.96.$$

The following theorem is very important and allows us to find probabilities by using the standard normal table.

**Theorem 6.7.** If $X \sim N(\mu, \sigma^2)$, then the random variable $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

**Proof:** We will show that $Z$ is standard normal by finding the probability density function of $Z$. We compute the probability density of $Z$ by cumulative distribution function method.

$$\begin{aligned} F(z) &= P(Z \leq z) \\ &= P\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= P(X \leq \sigma\, z + \mu) \\ &= \int_{-\infty}^{\sigma\, z+\mu} \frac{1}{\sigma\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\,\pi}}\, \sigma\, e^{-\frac{1}{2}\,w^2}\, dw, \qquad \text{where } w = \frac{x-\mu}{\sigma}. \end{aligned}$$

Hence

$$f(z) = F'(z) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\,z^2}.$$

The following example illustrates how to use standard normal table to find probability for normal random variables.

**Example 6.24.** If $X \sim N(3, 16)$, then what is $P(4 \leq X \leq 8)$?

**Answer:**

$$\begin{aligned} P(4 \le X \le 8) &= P\left(\frac{4-3}{4} \le \frac{X-3}{4} \le \frac{8-3}{4}\right) \\ &= P\left(\frac{1}{4} \le Z \le \frac{5}{4}\right) \\ &= P(Z \le 1.25) - P(Z \le 0.25) \\ &= 0.8944 - 0.5987 \\ &= 0.2957. \end{aligned}$$

**Example 6.25.** If $X \sim N(25, 36)$, then what is the value of the constant $c$ such that $P(|X - 25| \le c) = 0.9544$?

**Answer:**

$$\begin{aligned} 0.9544 &= P(|X - 25| \le c) \\ &= P(-c \le X - 25 \le c) \\ &= P\left(-\frac{c}{6} \le \frac{X-25}{6} \le \frac{c}{6}\right) \\ &= P\left(-\frac{c}{6} \le Z \le \frac{c}{6}\right) \\ &= P\left(Z \le \frac{c}{6}\right) - P\left(Z \le -\frac{c}{6}\right) \\ &= 2\,P\left(Z \le \frac{c}{6}\right) - 1. \end{aligned}$$

Hence

$$P\left(Z \le \frac{c}{6}\right) = 0.9772$$

and from this, using the normal table, we get

$$\frac{c}{6} = 2 \qquad \text{or} \qquad c = 12.$$

The following theorem can be proved similar to Theorem 6.7.

**Theorem 6.8.** If $X \sim N(\mu, \sigma^2)$, then the random variable $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$.

**Proof:** Let $W = \left(\frac{X-\mu}{\sigma}\right)^2$ and $Z = \frac{X-\mu}{\sigma}$. We will show that the random variable $W$ is chi-square with 1 degree of freedom. This amounts to showing that the probability density function of $W$ to be

$$g(w) = \begin{cases} \frac{1}{\sqrt{2\pi\, w}}\, e^{-\frac{1}{2}w} & \text{if } 0 < w < \infty \\ 0 & \text{otherwise .} \end{cases}$$

We compute the probability density function of $W$ by distribution function method. Let $G(w)$ be the cumulative distribution function $W$, which is

$$\begin{aligned}
G(w) &= P\left(W \leq w\right) \\
&= P\left(\left(\frac{X-\mu}{\sigma}\right)^2 \leq w\right) \\
&= P\left(-\sqrt{w} \leq \frac{X-\mu}{\sigma} \leq \sqrt{w}\right) \\
&= P\left(-\sqrt{w} \leq Z \leq \sqrt{w}\right) \\
&= \int_{-\sqrt{w}}^{\sqrt{w}} f(z)\, dz,
\end{aligned}$$

where $f(z)$ denotes the probability density function of the standard normal random variable $Z$. Thus, the probability density function of $W$ is given by

$$\begin{aligned}
g(w) &= \frac{d}{dw} G(w) \\
&= \frac{d}{dw} \int_{-\sqrt{w}}^{\sqrt{w}} f(z)\, dz \\
&= f\left(\sqrt{w}\right) \frac{d\sqrt{w}}{dw} - f\left(-\sqrt{w}\right) \frac{d\left(-\sqrt{w}\right)}{dw} \\
&= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}w}\, \frac{1}{2\sqrt{w}} + \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}w}\, \frac{1}{2\sqrt{w}} \\
&= \frac{1}{\sqrt{2\pi\, w}}\, e^{-\frac{1}{2}w}.
\end{aligned}$$

Thus, we have shown that $W$ is chi-square with one degree of freedom and the proof is now complete.

**Example 6.26.** If $X \sim N(7, 4)$, what is $P\left(15.364 \leq (X-7)^2 \leq 20.095\right)$?

**Answer:** Since $X \sim N(7, 4)$, we get $\mu = 7$ and $\sigma = 2$. Thus

$$\begin{aligned}
&P\left(15.364 \leq (X-7)^2 \leq 20.095\right) \\
&\qquad = P\left(\frac{15.364}{4} \leq \left(\frac{X-7}{2}\right)^2 \leq \frac{20.095}{4}\right) \\
&\qquad = P\left(3.841 \leq Z^2 \leq 5.024\right) \\
&\qquad = P\left(0 \leq Z^2 \leq 5.024\right) - P\left(0 \leq Z^2 \leq 3.841\right) \\
&\qquad = 0.975 - 0.949 \\
&\qquad = 0.026.
\end{aligned}$$

A generalization of the normal distribution is the following:

$$g(x) = \frac{\nu\,\varphi(\nu)}{2\,\sigma\,\Gamma(1/\nu)}\,e^{-\left(\frac{\varphi(\nu)}{\sigma}\,|x-\mu|\right)^{\nu}}$$

where

$$\varphi(\nu) = \sqrt{\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}}$$

and $\nu$ and $\sigma$ are real positive constants and $-\infty < \mu < \infty$ is a real constant. The constant $\mu$ represents the mean and the constant $\sigma$ represents the standard deviation of the generalized normal distribution. If $\nu = 2$, then generalized normal distribution reduces to the normal distribution. If $\nu = 1$, then the generalized normal distribution reduces to the Laplace distribution whose density function is given by

$$f(x) = \frac{1}{2\,\theta}\,e^{-\frac{|x-\mu|}{\theta}}$$

where $\theta = \frac{\sigma}{\sqrt{2}}$. The generalized normal distribution is very useful in signal processing and in particular modeling of the discrete cosine transform (DCT) coefficients of a digital image.

### 6.5. Lognormal Distribution

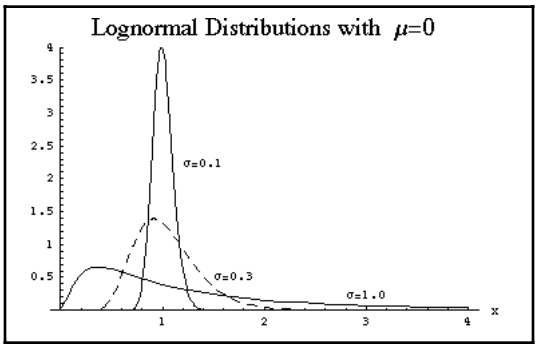
The study lognormal distribution was initiated by Galton and McAlister in 1879. They came across this distribution while studying the use of the geometric mean as an estimate of location. Later, Kapteyn (1903) discussed the genesis of this distribution. This distribution can be defined as the distribution of a random variable whose logarithm is normally distributed. Often the size distribution of organisms, the distribution of species, the distribution of the number of persons in a census occupation class, the distribution of stars in the universe, and the distribution of the size of incomes are modeled by lognormal distributions. The lognormal distribution is used in biology, astronomy, economics, pharmacology and engineering. This distribution is sometimes known as the Galton-McAlister distribution. In economics, the lognormal distribution is called the Cobb-Douglas distribution.

**Definition 6.10.** A random variable $X$ is said to have a lognormal distribution if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{x\,\sigma\,\sqrt{2\,\pi}}\,e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}, & \text{if } 0 < x < \infty \\ 0 & \text{otherwise ,} \end{cases}$$

where $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ are arbitrary parameters.

If $X$ has a lognormal distribution with parameters $\mu$ and $\sigma^2$, then we write $X \sim \bigwedge\!\!\backslash(\mu, \sigma^2)$.




**Example 6.27.** If $X \sim \bigwedge\!\!\backslash(\mu, \sigma^2)$, what is the $100p^{\text{th}}$ percentile of $X$?

**Answer:** Let $q$ be the $100p^{\text{th}}$ percentile of $X$. Then by definition of percentile, we get

$$p = \int_0^q \frac{1}{x\,\sigma\,\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} dx.$$

Substituting $z = \frac{\ln(x)-\mu}{\sigma}$ in the above integral, we have

$$\begin{aligned} p &= \int_{-\infty}^{\frac{\ln(q)-\mu}{\sigma}} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}z^2} dz \\ &= \int_{-\infty}^{z_p} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}z^2} dz, \end{aligned}$$

where $z_p = \frac{\ln(q)-\mu}{\sigma}$ is the $100p^{\text{th}}$ of the standard normal random variable. Hence $100p^{\text{th}}$ percentile of $X$ is

$$q = e^{\sigma\, z_p + \mu},$$

where $z_p$ is the $100p^{\text{th}}$ percentile of the standard normal random variable $Z$.

**Theorem 6.9.** If $X \sim \bigwedge\!\!\backslash(\mu, \sigma^2)$, then

$$\begin{aligned} E(X) &= e^{\mu + \frac{1}{2}\sigma^2} \\ Var(X) &= \left[e^{\sigma^2} - 1\right] e^{2\mu + \sigma^2}. \end{aligned}$$

**Proof:** Let $t$ be a positive integer. We compute the $t^{\text{th}}$ moment of $X$.

$$\begin{aligned} E\left(X^t\right) &= \int_0^\infty x^t\, f(x)\, dx \\ &= \int_0^\infty x^t\, \frac{1}{x\,\sigma\,\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} dx. \end{aligned}$$

Substituting $z = \ln(x)$ in the last integral, we get

$$E\left(X^t\right) = \int_{-\infty}^\infty e^{tz}\, \frac{1}{\sigma\,\sqrt{2\,\pi}}\, e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz = M_Z(t),$$

where $M_Z(t)$ denotes the moment generating function of the random variable $Z \sim N(\mu, \sigma^2)$. Therefore,

$$M_Z(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Thus letting $t = 1$, we get

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2}.$$

Similarly, taking $t = 2$, we have

$$E(X^2) = e^{2\mu + 2\sigma^2}.$$

Thus, we have

$$Var(X) = E(X^2) - E(X)^2 = \left[e^{\sigma^2} - 1\right] e^{2\mu+\sigma^2}$$

and now the proof of the theorem is complete.

**Example 6.28.** If $X \sim \bigwedge\!\!\bigwedge(0, 4)$, then what is the probability that $X$ is between 1 and 12.1825?

**Answer:** Since $X \sim \bigwedge\!\!\bigwedge(0, 4)$, the random variable $Y = \ln(X) \sim N(0, 4)$. Hence

$$\begin{aligned} P(1 \le X \le 12.1825) &= P(\ln(1) \le \ln(X) \le \ln(12.1825)) \\ &= P(0 \le Y \le 2.50) \\ &= P(0 \le Z \le 1.25) \\ &= P(Z \le 1.25) - P(Z \le 0) \\ &= 0.8944 - 0.5000 \\ &= 0.4944. \end{aligned}$$

**Example 6.29.** If the amount of time needed to solve a problem by a group of students follows the lognormal distribution with parameters $\mu$ and $\sigma^2$, then what is the value of $\mu$ so that the probability of solving a problem in 10 minutes or less by any randomly picked student is 95% when $\sigma^2 = 4$?

**Answer:** Let the random variable $X$ denote the amount of time needed to a solve a problem. Then $X \sim \bigwedge\!\!\backslash(\mu, 4)$. We want to find $\mu$ so that $P(X \leq 10) = 0.95$. Hence

$$\begin{aligned} 0.95 &= P(X \leq 10) \\ &= P(\ln(X) \leq \ln(10)) \\ &= P(\ln(X) - \mu \leq \ln(10) - \mu) \\ &= P\left(\frac{\ln(X) - \mu}{2} \leq \frac{\ln(10) - \mu}{2}\right) \\ &= P\left(Z \leq \frac{\ln(10) - \mu}{2}\right), \end{aligned}$$

where $Z \sim N(0, 1)$. Using the table for standard normal distribution, we get

$$\frac{\ln(10) - \mu}{2} = 1.65.$$

Hence

$$\mu = \ln(10) - 2(1.65) = 2.3025 - 3.300 = -0.9975.$$

### 6.6. Inverse Gaussian Distribution

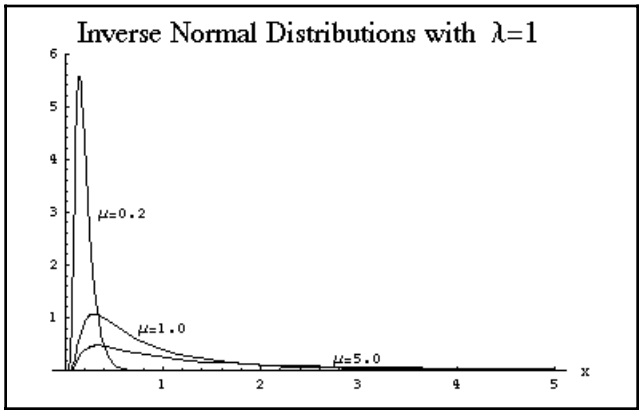
If a sufficiently small macroscopic particle is suspended in a fluid that is in thermal equilibrium, the particle will move about erratically in response to natural collisional bombardments by the individual molecules of the fluid. This erratic motion is called "Brownian motion" after the botanist Robert Brown (1773-1858) who first observed this erratic motion in 1828. Independently, Einstein (1905) and Smoluchowski (1906) gave the mathematical description of Brownian motion. The distribution of the first passage time in Brownian motion is the inverse Gaussian distribution. This distribution was systematically studied by Tweedie in 1945. The interpurchase times of toothpaste of a family, the duration of labor strikes in a geographical region, word frequency in a language, conversion time for convertible bonds, length of employee service, and crop field size follow inverse Gaussian distribution. Inverse Gaussian distribution is very useful for analysis of certain skewed data.

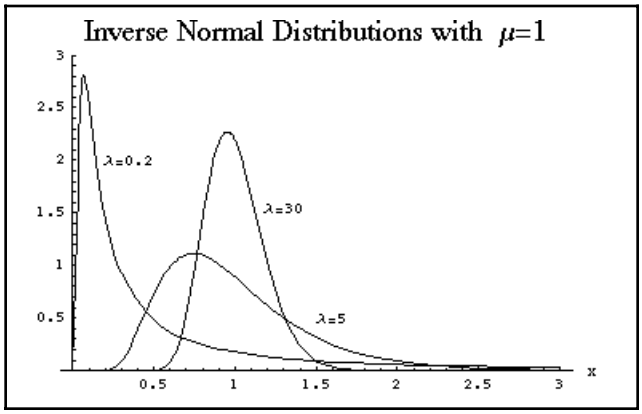
**Definition 6.10.** A random variable $X$ is said to have an inverse Gaussian distribution if its probability density function is given by

$$f(x) = \begin{cases} \sqrt{\frac{\lambda}{2\pi}}\, x^{-\frac{3}{2}}\; e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}, & \text{if } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \mu < \infty$ and $0 < \lambda < \infty$ are arbitrary parameters.

If $X$ has an inverse Gaussian distribution with parameters $\mu$ and $\lambda$, then we write $X \sim IG(\mu, \lambda)$.




The characteristic function $\phi(t)$ of $X \sim IG(\mu, \lambda)$ is

$$\begin{aligned} \phi(t) &= E\left(e^{itX}\right) \\ &= e^{\frac{\lambda}{\mu}\left[1-\sqrt{1-\frac{2i\mu^2 t}{\lambda}}\right]}. \end{aligned}$$

Using this, we have the following theorem.

**Theorem 6.10.** If $X \sim IG(\mu, \lambda)$, then

$$E(X) = \mu$$
$$Var(X) = \frac{\mu^3}{\lambda}.$$

**Proof:** Since $\phi(t) = E\left(e^{itX}\right)$, the derivative $\phi'(t) = i\,E\left(Xe^{itX}\right)$. Therefore $\phi'(0) = i\,E\,(X)$. We know the characteristic function $\phi(t)$ of $X \sim IG(\mu, \lambda)$ is

$$\phi(t) = e^{\frac{\lambda}{\mu}\left[1-\sqrt{1-\frac{2i\mu^2 t}{\lambda}}\right]}.$$

Differentiating $\phi(t)$ with respect to $t$, we have

$$\begin{aligned}
\phi'(t) &= \frac{d}{dt}\left[e^{\frac{\lambda}{\mu}\left[1-\sqrt{1-\frac{2i\mu^2 t}{\lambda}}\right]}\right] \\
&= e^{\frac{\lambda}{\mu}\left[1-\sqrt{1-\frac{2i\mu^2 t}{\lambda}}\right]}\ \frac{d}{dt}\left(\frac{\lambda}{\mu}\left[1-\sqrt{1-\frac{2i\mu^2 t}{\lambda}}\right]\right) \\
&= i\mu\, e^{\frac{\lambda}{\mu}\left[1-\sqrt{1-\frac{2i\mu^2 t}{\lambda}}\right]}\ \left(1-\frac{2i\mu^2 t}{\lambda}\right)^{-\frac{1}{2}}.
\end{aligned}$$

Hence $\phi'(0) = i\,\mu$. Therefore, $E(X) = \mu$. Similarly, one can show that

$$Var(X) = \frac{\mu^3}{\lambda}.$$

This completes the proof of the theorem.

The distribution function $F(x)$ of the inverse Gaussian random variable $X$ with parameters $\mu$ and $\lambda$ was computed by Shuster (1968) as

$$F(x) = \Phi\left(\sqrt{\frac{\lambda}{\mu}}\left[\frac{x}{\mu}-1\right]\right) + e^{\frac{2\lambda}{\mu}}\,\Phi\left(-\sqrt{\frac{\lambda}{\mu}}\left[\frac{x}{\mu}+1\right]\right),$$

where $\Phi$ is the distribution function of the standard normal distribution function.

### 6.7. Logistics Distribution

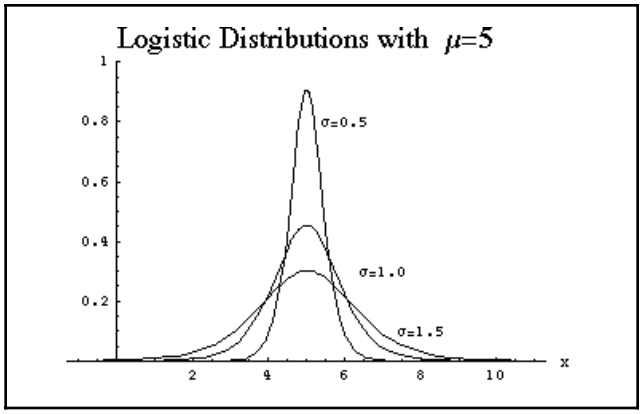
The logistic distribution is often considered as an alternative to the univariate normal distribution. The logistic distribution has a shape very close

to that of a normal distribution but has heavier tails than the normal. The logistic distribution is used in modeling demographic data. It is also used as an alternative to the Weibull distribution in life-testing.

**Definition 6.11.** A random variable $X$ is said to have a logistic distribution if its probability density function is given by

$$f(x) = \frac{\pi}{\sigma\sqrt{3}} \frac{e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}}{\left[1+e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}\right]^2} \qquad -\infty < x < \infty,$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are parameters.




If $X$ has a logistic distribution with parameters $\mu$ and $\sigma$, then we write $X \sim LOG(\mu, \sigma)$.

**Theorem 6.11.** If $X \sim LOG(\mu, \lambda)$, then

$$\begin{aligned} E(X) &= \mu \\ Var(X) &= \sigma^2 \\ M(t) &= e^{\mu t}\,\Gamma\left(1+\frac{\sqrt{3}}{\pi}\,\sigma t\right)\,\Gamma\left(1-\frac{\sqrt{3}}{\pi}\,\sigma t\right), \quad |t| < \frac{\pi}{\sigma\sqrt{3}}. \end{aligned}$$

**Proof:** First, we derive the moment generating function of $X$ and then we

compute the mean and variance of it. The moment generating function is

$$
\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{\pi}{\sigma\sqrt{3}} \frac{e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}}{\left[1+e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}\right]^2} dx \\
&= e^{\mu t} \int_{-\infty}^{\infty} e^{sw} \frac{e^{-w}}{(1+e^{-w})^2}\, dw, \text{ where } w = \frac{\pi(x-\mu)}{\sqrt{3}\,\sigma} \text{ and } s = \frac{\sqrt{3}\,\sigma}{\pi}\, t \\
&= e^{\mu t} \int_{-\infty}^{\infty} \left(e^{-w}\right)^{-s} \frac{e^{-w}}{(1+e^{-w})^2}\, dw \\
&= e^{\mu t} \int_{0}^{1} \left(z^{-1}-1\right)^{-s} dz, \text{ where } z = \frac{1}{1+e^{-w}} \\
&= e^{\mu t} \int_{0}^{1} z^{s}(1-z)^{-s}\, dz \\
&= e^{\mu t}\, B(1+s,\, 1-s) \\
&= e^{\mu t} \frac{\Gamma(1+s)\,\Gamma(1-s)}{\Gamma(1+s+1-s)} \\
&= e^{\mu t} \frac{\Gamma(1+s)\,\Gamma(1-s)}{\Gamma(2)} \\
&= e^{\mu t}\, \Gamma(1+s)\,\Gamma(1-s) \\
&= e^{\mu t}\, \Gamma\left(1+\frac{\sqrt{3}}{\pi}\,\sigma t\right) \Gamma\left(1-\frac{\sqrt{3}}{\pi}\,\sigma t\right) \\
&= e^{\mu t} \left(\frac{\sqrt{3}\,\sigma}{\pi}\, t\right) cosec\left(\frac{\sqrt{3}\,\sigma}{\pi}\, t\right).
\end{aligned}
$$

We leave the rest of the proof to the reader.

### 6.8. Review Exercises

**1.** If $Y \sim UNIF(0,1)$, then what is the probability density function of $X = -\ln Y$?

**2.** Let the probability density function of $X$ be

$$
f(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise .} \end{cases}
$$

Let $Y = 1 - e^{-X}$. Find the distribution of $Y$.

**3.** After a certain time the weight $W$ of crystals formed is given approximately by $W = e^X$ where $X \sim N(\mu, \sigma^2)$. What is the probability density function of $W$ for $0 < w < \infty$ ?

**4.** What is the probability that a normal random variable with mean 6 and standard deviation 3 will fall between 5.7 and 7.5 ?

**5.** Let $X$ have a distribution with the $75^{\text{th}}$ percentile equal to $\frac{1}{3}$ and probability density function equal to

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the value of the parameter $\lambda$ ?

**6.** If a normal distribution with mean $\mu$ and variance $\sigma^2 > 0$ has $46^{\text{th}}$ percentile equal to $20\sigma$, then what is $\mu$ in term of standard deviation?

**7.** Let $X$ be a random variable with cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \\ 1 - e^{-x} & \text{if } x > 0. \end{cases}$$

What is $P\left(0 \leq e^X \leq 4\right)$?

**8.** Let $X$ have the density function

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$. If $\beta = 6$ and $\alpha = 5$, what is the mean of the random variable $(1 - X)^{-1}$ ?

**9.** R.A. Fisher proved that when $n \geq 30$ and $Y$ has a chi-square distribution with $n$ degrees freedom, then $\sqrt{2Y} - \sqrt{2n-1}$ has an approximate standard normal distribution. Under this approximation, what is the $90^{\text{th}}$ percentile of $Y$ when $n = 41$ ?

**10.** Let $Y$ have a chi-square distribution with 32 degrees of freedom so that its variance is 64. If $P\left(Y > c\right) = 0.0668$, then what is the approximate value of the constant $c$?

**11.** If in a certain normal distribution of $X$, the probability is 0.5 that $X$ is less than 500 and 0.0227 that $X$ is greater than 650. What is the standard deviation of $X$?

**12.** If $X \sim N(5, 4)$, then what is the probability that $8 < Y < 13$ where $Y = 2X + 1$?

**13.** Given the probability density function of a random variable $X$ as

$$f(x) = \begin{cases} \theta\, e^{-\theta x} & \text{if } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the $n^{\text{th}}$ moment of $X$ about the origin?

**14.** If the random variable $X$ is normal with mean 1 and standard deviation 2, then what is $P\left(X^2 - 2X \leq 8\right)$?

**15.** Suppose $X$ has a standard normal distribution and $Y = e^X$. What is the $k^{\text{th}}$ moment of $Y$?

**16.** If the random variable $X$ has uniform distribution on the interval $[0, a]$, what is the probability that the random variable greater than its square, that is $P\left(X > X^2\right)$?

**17.** If the random variable $Y$ has a chi-square distribution with 54 degrees of freedom, then what is the approximate $84^{\text{th}}$ percentile of $Y$?

**18.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} \frac{2}{x^2} & \text{for } 1 < x < 2 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

If $Y = \sqrt{X}$, what is the density function for $Y$ where nonzero?

**19.** If $X$ is normal with mean 0 and variance 4, then what is the probability of the event $X - \frac{4}{X} \geq 0$, that is $P\left(X - \frac{4}{X} \geq 0\right)$?

**20.** If the waiting time at Rally's drive-in-window is normally distributed with mean 13 minutes and standard deviation 2 minutes, then what percentage of customers wait longer than 10 minutes but less than 15 minutes?

**21.** If $X$ is uniform on the interval from $-5$ to 5, what is the probability that the quadratic equation $100t^2 + 20tX + 2X + 3 = 0$ has complex solutions?

**22.** If the random variable $X \sim Exp(\theta)$, then what is the probability density function of the random variable $Y = X\sqrt{X}$?

**23.** If the random variable $X \sim N(0, 1)$, then what is the probability density function of the random variable $Y = \sqrt{|X|}$?

**24.** If the random variable $X \sim \bigwedge\!\!\backslash\!\!\backslash(\mu, \sigma^2)$, then what is the probability density function of the random variable $\ln(X)$?

**25.** If the random variable $X \sim \bigwedge\!\!\backslash\!\!\backslash(\mu, \sigma^2)$, then what is the mode of $X$?

**26.** If the random variable $X \sim \bigwedge\!\!\backslash\!\!\backslash(\mu, \sigma^2)$, then what is the median of $X$?

**27.** If the random variable $X \sim \bigwedge\!\!\backslash\!\!\backslash(\mu, \sigma^2)$, then what is the probability that the quadratic equation $4t^2 + 4tX + X + 2 = 0$ has real solutions?

**28.** Consider the Karl Pearson's differential equation $p(x)\,\frac{dy}{dx} + q(x)\,y = 0$ where $p(x) = a + bx + cx^2$ and $q(x) = x - d$. Show that if $a = c = 0$, $b > 0$, $d > -b$, then $y(x)$ is gamma; and if $a = 0$, $b = -c$, $\frac{d-1}{b} < 1$, $\frac{d}{b} > -1$, then $y(x)$ is beta.

**29.** Let $a, b, \alpha, \beta$ be any four real numbers with $a < b$ and $\alpha, \beta$ positive. If $X \sim BETA(\alpha, \beta)$, then what is the probability density function of the random variable $Y = (b - a)X + a$?

**30.** A nonnegative continuous random variable $X$ is said to be memoryless if $P(X > s + t/X > t) = P(X > s)$ for all $s, t \geq 0$. Show that the exponential random variable is memoryless.

**31.** Show that every nonnegative continuous memoryless random variable is an exponential random variable.

**32.** Using gamma function evaluate the following integrals:
(i) $\int_0^\infty e^{-x^2}\,dx$; (ii) $\int_0^\infty x\,e^{-x^2}\,dx$; (iii) $\int_0^\infty x^2\,e^{-x^2}\,dx$; (iv) $\int_0^\infty x^3\,e^{-x^2}\,dx$.

**33.** Using beta function evaluate the following integrals:
(i) $\int_0^1 x^2\,(1-x)^2\,dx$; (ii) $\int_0^{100} x^5\,(100-x)^7\,dx$; (iii) $\int_0^1 x^{11}\,(1-x^3)^7\,dx$.

**34.** If $\Gamma(z)$ denotes the gamma function, then prove that

$$\Gamma(1+t)\,\Gamma(1-t) = t\text{cosec}(t).$$

**35.** Let $\alpha$ and $\beta$ be given positive real numbers, with $\alpha < \beta$. If two points are selected at random from a straight line segment of length $\beta$, what is the probability that the distance between them is at least $\alpha$ ?

**36.** If the random variable $X \sim GAM(\theta, \alpha)$, then what is the $n^{\text{th}}$ moment of $X$ about the origin?

# Chapter 7
# TWO RANDOM VARIABLES

There are many random experiments that involve more than one random variable. For example, an educator may study the joint behavior of grades and time devoted to study; a physician may study the joint behavior of blood pressure and weight. Similarly an economist may study the joint behavior of business volume and profit. In fact, most real problems we come across will have more than one underlying random variable of interest.

## 7.1. Bivariate Discrete Random Variables

In this section, we develop all the necessary terminologies for studying bivariate discrete random variables.

**Definition 7.1.** A discrete *bivariate random variable* $(X, Y)$ is an ordered pair of discrete random variables.

**Definition 7.2.** Let $(X, Y)$ be a bivariate random variable and let $R_X$ and $R_Y$ be the range spaces of $X$ and $Y$, respectively. A real-valued function $f : R_X \times R_Y \to \mathbb{R}$ is called a *joint probability density function* for $X$ and $Y$ if and only if

$$f(x, y) = P(X = x, Y = y)$$

for all $(x, y) \in R_X \times R_Y$. Here, the event $(X = x, Y = y)$ means the intersection of the events $(X = x)$ and $(Y = y)$, that is

$$(X = x) \bigcap (Y = y).$$

**Example 7.1.** Roll a pair of unbiased dice. If $X$ denotes the smaller and $Y$ denotes the larger outcome on the dice, then what is the joint probability density function of $X$ and $Y$ ?

**Answer:** The sample space $S$ of rolling two dice consists of

$$S = \begin{matrix} \{(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\} \end{matrix}$$

The probability density function $f(x,y)$ can be computed for $X = 2$ and $Y = 3$ as follows: There are two outcomes namely $(2,3)$ and $(3,2)$ in the sample $S$ of 36 outcomes which contribute to the joint event $(X = 2, Y = 3)$. Hence

$$f(2,3) = P(X = 2, Y = 3) = \frac{2}{36}.$$

Similarly, we can compute the rest of the probabilities. The following table shows these probabilities:

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |
| 5 | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 0 |
| 4 | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 0 | 0 |
| 3 | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 0 | 0 | 0 |
| 2 | $\frac{2}{36}$ | $\frac{1}{36}$ | 0 | 0 | 0 | 0 |
| 1 | $\frac{1}{36}$ | 0 | 0 | 0 | 0 | 0 |
| | 1 | 2 | 3 | 4 | 5 | 6 |

These tabulated values can be written as

$$f(x,y) = \begin{cases} \frac{1}{36} & \text{if } 1 \le x = y \le 6 \\ \frac{2}{36} & \text{if } 1 \le x < y \le 6 \\ 0 & \text{otherwise.} \end{cases}$$

**Example 7.2.** A group of 9 executives of a certain firm include 4 who are married, 3 who never married, and 2 who are divorced. Three of the

executives are to be selected for promotion. Let $X$ denote the number of married executives and $Y$ the number of never married executives among the 3 selected for promotion. Assuming that the three are randomly selected from the nine available, what is the joint probability density function of the random variables $X$ and $Y$?

**Answer:** The number of ways we can choose 3 out of 9 is $\binom{9}{3}$ which is 84. Thus

$$f(0,0) = P(X = 0, Y = 0) = \frac{0}{84} = 0$$

$$f(1,0) = P(X = 1, Y = 0) = \frac{\binom{4}{1}\binom{3}{0}\binom{2}{2}}{84} = \frac{4}{84}$$

$$f(2,0) = P(X = 2, Y = 0) = \frac{\binom{4}{2}\binom{3}{0}\binom{2}{1}}{84} = \frac{12}{84}$$

$$f(3,0) = P(X = 3, Y = 0) = \frac{\binom{4}{3}\binom{3}{0}\binom{2}{0}}{84} = \frac{4}{84}.$$

Similarly, we can find the rest of the probabilities. The following table gives the complete information about these probabilities.

| | | | | |
|---|---|---|---|---|
| 3 | $\frac{1}{84}$ | 0 | 0 | 0 |
| 2 | $\frac{6}{84}$ | $\frac{12}{84}$ | 0 | 0 |
| 1 | $\frac{3}{84}$ | $\frac{24}{84}$ | $\frac{18}{84}$ | 0 |
| 0 | 0 | $\frac{4}{84}$ | $\frac{12}{84}$ | $\frac{4}{84}$ |
| | 0 | 1 | 2 | 3 |

**Definition 7.3.** Let $(X, Y)$ be a discrete bivariate random variable. Let $R_X$ and $R_Y$ be the range spaces of $X$ and $Y$, respectively. Let $f(x, y)$ be the joint probability density function of $X$ and $Y$. The function

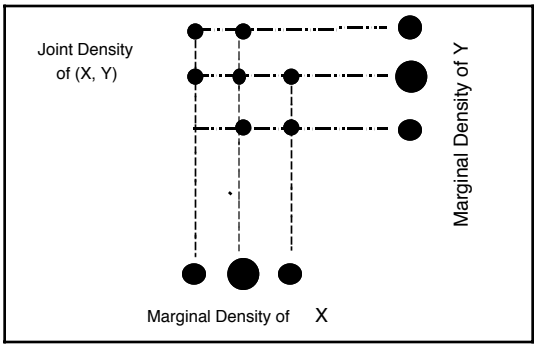
$$f_1(x) = \sum_{y \in R_Y} f(x, y)$$

is called the *marginal probability density function* of $X$. Similarly, the function

$$f_2(y) = \sum_{x \in R_X} f(x, y)$$

is called the *marginal probability density function* of $Y$.

The following diagram illustrates the concept of marginal graphically.




**Example 7.3.** If the joint probability density function of the discrete random variables $X$ and $Y$ is given by

$$f(x,y) = \begin{cases} \frac{1}{36} & \text{if } 1 \leq x = y \leq 6 \\ \frac{2}{36} & \text{if } 1 \leq x < y \leq 6 \\ 0 & \text{otherwise,} \end{cases}$$

then what are marginals of $X$ and $Y$?

**Answer:** The marginal of $X$ can be obtained by summing the joint probability density function $f(x,y)$ for all $y$ values in the range space $R_Y$ of the random variable $Y$. That is

$$\begin{aligned} f_1(x) &= \sum_{y \in R_Y} f(x, y) \\ &= \sum_{y=1}^{6} f(x, y) \\ &= f(x, x) + \sum_{y>x} f(x, y) + \sum_{y<x} f(x, y) \\ &= \frac{1}{36} + (6 - x)\,\frac{2}{36} + 0 \\ &= \frac{1}{36}\,[13 - 2\,x]\,, \qquad x = 1, 2, ..., 6. \end{aligned}$$

Similarly, one can obtain the marginal probability density of $Y$ by summing over for all $x$ values in the range space $R_X$ of the random variable $X$. Hence

$$\begin{aligned} f_2(y) &= \sum_{x \in R_X} f(x,\, y) \\ &= \sum_{x=1}^{6} f(x,\, y) \\ &= f(y,\, y) + \sum_{x<y} f(x,\, y) + \sum_{x>y} f(x,\, y) \\ &= \frac{1}{36} + (y-1)\,\frac{2}{36} + 0 \\ &= \frac{1}{36}\,[2y-1]\,, \qquad y = 1, 2, ..., 6. \end{aligned}$$

**Example 7.4.** Let $X$ and $Y$ be discrete random variables with joint probability density function

$$f(x,y) = \begin{cases} \frac{1}{21}\,(x+y) & \text{if } x = 1, 2;\ y = 1, 2, 3 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What are the marginal probability density functions of $X$ and $Y$?

**Answer:** The marginal of $X$ is given by

$$\begin{aligned} f_1(x) &= \sum_{y=1}^{3} \frac{1}{21}\,(x+y) \\ &= \frac{1}{21}\,3\,x + \frac{1}{21}\,[1+2+3] \\ &= \frac{x+2}{7}, \qquad x = 1, 2. \end{aligned}$$

Similarly, the marginal of $Y$ is given by

$$\begin{aligned} f_2(y) &= \sum_{x=1}^{2} \frac{1}{21}\,(x+y) \\ &= \frac{2y}{21} + \frac{3}{21} \\ &= \frac{3+2\,y}{21}, \qquad y = 1, 2, 3. \end{aligned}$$

From the above examples, note that the marginal $f_1(x)$ is obtained by summing across the columns. Similarly, the marginal $f_2(y)$ is obtained by summing across the rows.

The following theorem follows from the definition of the joint probability density function.

**Theorem 7.1.** A real valued function $f$ of two variables is a joint probability density function of a pair of discrete random variables $X$ and $Y$ (with range spaces $R_X$ and $R_Y$, respectively) if and only if

$$(a) \quad f(x,y) \geq 0 \quad \text{for all } (x,y) \in R_X \times R_Y;$$

$$(b) \quad \sum_{x \in R_X} \sum_{y \in R_Y} f(x,y) = 1.$$

**Example 7.5.** For what value of the constant $k$ the function given by

$$f(x,y) = \begin{cases} k\,xy & \text{if } x = 1,2,3;\ y = 1,2,3 \\ 0 & \text{otherwise} \end{cases}$$

is a joint probability density function of some random variables $X$ and $Y$ ?

**Answer:** Since

$$\begin{aligned} 1 &= \sum_{x=1}^{3} \sum_{y=1}^{3} f(x,y) \\ &= \sum_{x=1}^{3} \sum_{y=1}^{3} k\,x\,y \\ &= k\,[1+2+3+2+4+6+3+6+9] \\ &= 36\,k. \end{aligned}$$

Hence

$$k = \frac{1}{36}$$

and the corresponding density function is given by

$$f(x,y) = \begin{cases} \frac{1}{36}\,xy & \text{if } x = 1,2,3;\ y = 1,2,3 \\ 0 & \text{otherwise .} \end{cases}$$

As in the case of one random variable, there are many situations where one wants to know the probability that the values of two random variables are less than or equal to some real numbers $x$ and $y$.

**Definition 7.4.** Let $X$ and $Y$ be any two discrete random variables. The real valued function $F : \mathbb{R}^2 \to \mathbb{R}$ is called the *joint cumulative probability distribution function* of $X$ and $Y$ if and only if

$$F(x,\, y) = P(X \leq x,\, Y \leq y)$$

for all $(x, y) \in \mathbb{R}^2$. Here, the event $(X \leq x, Y \leq y)$ means $(X \leq x) \bigcap (Y \leq y)$.

From this definition it can be shown that for any real numbers $a$ and $b$

$$F(a \leq X \leq b,\, c \leq Y \leq d) = F(b,\, d) + F(a,\, c) - F(a,\, d) - F(b,\, c).$$

Further, one can also show that

$$F(x,\, y) = \sum_{s \leq x} \sum_{t \leq y} f(s,\, t)$$

where $(s, t)$ is any pair of nonnegative numbers.

### 7.2. Bivariate Continuous Random Variables

In this section, we shall extend the idea of probability density functions of one random variable to that of two random variables.

**Definition 7.5.** The *joint probability density function* of the random variables $X$ and $Y$ is an integrable function $f(x, y)$ such that

(a) $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$; and
(b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,\, y)\, dx\, dy = 1$.

**Example 7.6.** Let the joint density function of $X$ and $Y$ be given by

$$f(x, y) = \begin{cases} k\, xy^2 & \text{if } 0 < x < y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the value of the constant $k$ ?

**Answer:** Since $f$ is a joint probability density function, we have

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,\, y)\, dx\, dy$$

$$= \int_0^1 \int_0^y k\, x\, y^2\, dx\, dy$$

$$= \int_0^1 k\, y^2 \int_0^y x\, dx\, dy$$

$$= \frac{k}{2}\int_0^1 y^4\, dy$$

$$= \frac{k}{10}\,\left[\, y^5 \,\right]_0^1$$

$$= \frac{k}{10}.$$

Hence $k = 10$.

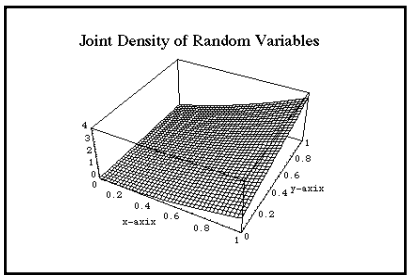
If we know the joint probability density function $f$ of the random variables $X$ and $Y$, then we can compute the probability of the event $A$ from

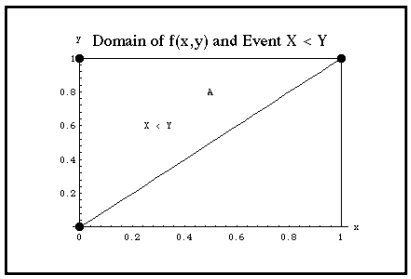
$$P(A) = \int\int_A f(x,\, y)\, dx\, dy.$$

**Example 7.7.** Let the joint density of the continuous random variables $X$ and $Y$ be

$$f(x,\, y) = \begin{cases} \frac{6}{5}\,\left(x^2 + 2\,xy\right) & \text{if } 0 \le x \le 1;\ 0 \le y \le 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the probability of the event $(X \le Y)$ ?



**Answer:** Let $A = (X \leq Y)$. we want to find

$$P(A) = \int\int_A f(x,\, y)\, dx\, dy$$

$$= \int_0^1 \left[ \int_0^y \frac{6}{5} \left( x^2 + 2\, x\, y \right)\, dx \right] dy$$

$$= \frac{6}{5} \int_0^1 \left[ \frac{x^3}{3} + x^2\, y \right]_{x=0}^{x=y} dy$$

$$= \frac{6}{5} \int_0^1 \frac{4}{3}\, y^3\, dy$$

$$= \frac{2}{5} \left[ y^4 \right]_0^1$$

$$= \frac{2}{5}.$$

**Definition 7.6.** Let $(X,\, Y)$ be a continuous bivariate random variable. Let $f(x, y)$ be the joint probability density function of $X$ and $Y$. The function

$$f_1(x) = \int_{-\infty}^{\infty} f(x,\, y)\, dy$$

is called the *marginal probability density function* of $X$. Similarly, the function

$$f_2(y) = \int_{-\infty}^{\infty} f(x,\, y)\, dx$$

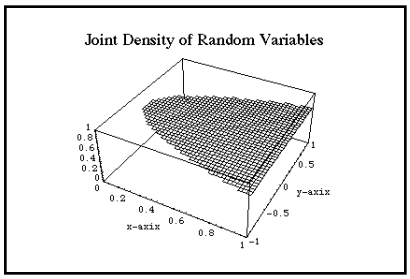
is called the *marginal probability density function* of $Y$.

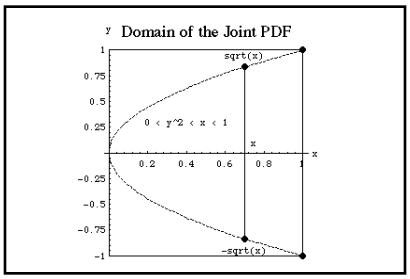
**Example 7.8.** If the joint density function for $X$ and $Y$ is given by

$$f(x, y) = \begin{cases} \frac{3}{4} & \text{for } 0 < y^2 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the marginal density function of $X$, for $0 < x < 1$?

**Answer:** The domain of the $f$ consists of the region bounded by the curve $x = y^2$ and the vertical line $x = 1$. (See the figure on the next page.)



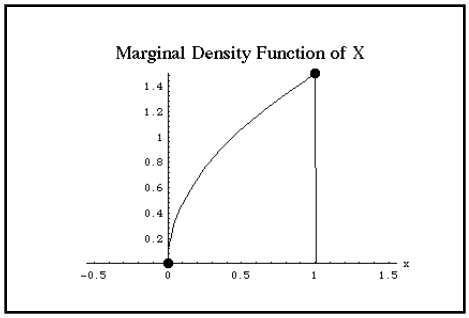




Hence

$$f_1(x) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{3}{4}\, dy$$

$$= \left[\frac{3}{4}\, y\right]_{-\sqrt{x}}^{\sqrt{x}}$$

$$= \frac{3}{2}\sqrt{x}.$$



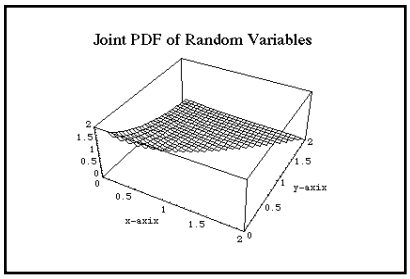

**Example 7.9.** Let $X$ and $Y$ have joint density function

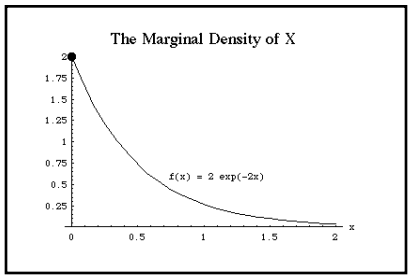
$$f(x,y) = \begin{cases} 2\, e^{-x-y} & \text{for } 0 < x \leq y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the marginal density of $X$ where nonzero?

**Answer:** The marginal density of $X$ is given by

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x,\, y)\, dy \\ &= \int_{x}^{\infty} 2\, e^{-x-y}\, dy \\ &= 2\, e^{-x} \int_{x}^{\infty} e^{-y}\, dy \\ &= 2\, e^{-x} \left[ -e^{-y} \right]_{x}^{\infty} \\ &= 2\, e^{-x}\, e^{-x} \\ &= 2\, e^{-2x} \qquad 0 < x < \infty. \end{aligned}$$




**Example 7.10.** Let $(X, Y)$ be distributed uniformly on the circular disk centered at $(0, 0)$ with radius $\frac{2}{\sqrt{\pi}}$. What is the marginal density function of $X$ where nonzero?

**Answer:** The equation of a circle with radius $\frac{2}{\sqrt{\pi}}$ and center at the origin is

$$x^2 + y^2 = \frac{4}{\pi}.$$

Hence, solving this equation for $y$, we get

$$y = \pm\sqrt{\frac{4}{\pi} - x^2}.$$

Thus, the marginal density of $X$ is given by

$$f_1(x) = \int_{-\sqrt{\frac{4}{\pi}-x^2}}^{\sqrt{\frac{4}{\pi}-x^2}} f(x,y)\,dy$$

$$= \int_{-\sqrt{\frac{4}{\pi}-x^2}}^{\sqrt{\frac{4}{\pi}-x^2}} \frac{1}{\text{area of the circle}}\,dy$$

$$= \int_{-\sqrt{\frac{4}{\pi}-x^2}}^{\sqrt{\frac{4}{\pi}-x^2}} \frac{1}{4}\,dy$$

$$= \left[\frac{1}{4}\ y\right]_{-\sqrt{\frac{4}{\pi}-x^2}}^{\sqrt{\frac{4}{\pi}-x^2}}$$

$$= \frac{1}{2}\sqrt{\frac{4}{\pi}-x^2}.$$

**Definition 7.7.** Let $X$ and $Y$ be the continuous random variables with joint probability density function $f(x,y)$. The *joint cumulative distribution function* $F(x,y)$ of $X$ and $Y$ is defined as

$$F(x,y) = P(X \le x,\, Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,\,v)\,du\,dv$$

for all $(x,y) \in \mathbb{R}^2$.

From the fundamental theorem of calculus, we again obtain

$$f(x,\,y) = \frac{\partial^2 F}{\partial x\,\partial y}.$$

**Example 7.11.** If the joint cumulative distribution function of $X$ and $Y$ is given by

$$F(x,y) = \begin{cases} \frac{1}{5}\left(2\,x^3\,y + 3\,x^2\,y^2\right) & \text{for } 0 < x, y < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

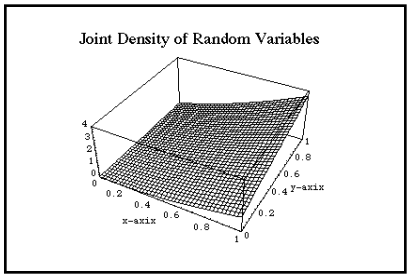
then what is the joint density of $X$ and $Y$ ?

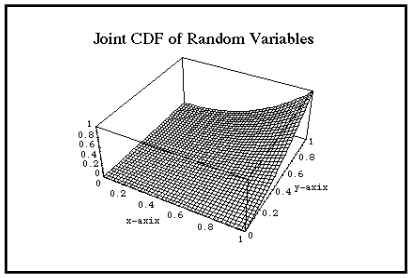
**Answer:**

$$\begin{aligned} f(x,\,y) &= \frac{1}{5}\,\frac{\partial}{\partial x}\,\frac{\partial}{\partial y}\left(2\,x^3\,y + 3\,x^2\,y^2\right) \\ &= \frac{1}{5}\,\frac{\partial}{\partial x}\left(2\,x^3 + 6\,x^2\,y\right) \\ &= \frac{1}{5}\,\left(6\,x^2 + 12\,x\,y\right) \\ &= \frac{6}{5}\,(x^2 + 2\,x\,y). \end{aligned}$$

Hence, the joint density of $X$ and $Y$ is given by

$$f(x,y) = \begin{cases} \frac{6}{5}\,\left(x^2 + 2\,x\,y\right) & \text{for } 0 < x, y < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$



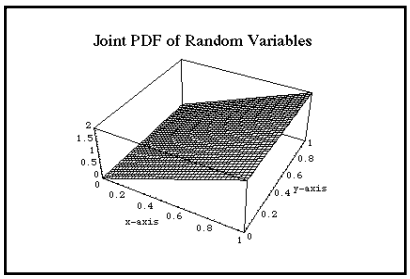



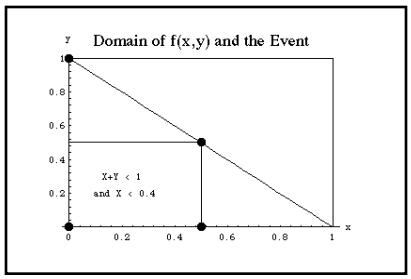

**Example 7.12.** Let $X$ and $Y$ have the joint density function

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

What is $P\left(X + Y \leq 1\,/\,X \leq \frac{1}{2}\right)$ ?

**Answer:** (See the diagram below.)



$$P\left(X+Y\leq 1\,/\,X\leq\frac{1}{2}\right)=\frac{P\left[(X+Y\leq 1)\bigcap\left(X\leq\frac{1}{2}\right)\right]}{P\left(X\leq\frac{1}{2}\right)}$$

$$=\frac{\int_0^{\frac{1}{2}}\left[\int_0^{\frac{1}{2}}2\,x\,dx\right]\,dy+\int_{\frac{1}{2}}^1\left[\int_0^{1-y}2\,x\,dx\right]\,dy}{\int_0^1\left[\int_0^{\frac{1}{2}}2\,x\,dx\right]\,dy}$$

$$=\frac{\frac{1}{6}}{\frac{1}{4}}$$

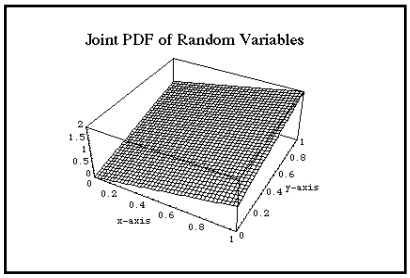
$$=\frac{2}{3}.$$

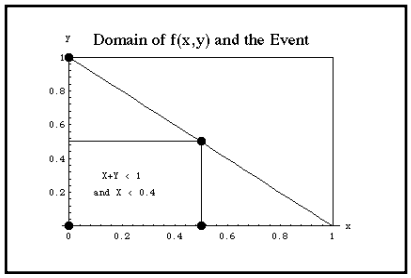
**Example 7.13.** Let $X$ and $Y$ have the joint density function

$$f(x,y)=\begin{cases}x+y & \text{for } 0\leq x\leq 1;\ 0\leq y\leq 1\\ \\ 0 & \text{elsewhere.}\end{cases}$$

What is $P\left(2X\leq 1\,/\,X+Y\leq 1\right)$ ?

**Answer:** We know that




$$P\left(2X\leq 1\,/\,X+Y\leq 1\right)=\frac{P\left[\left(X\leq\frac{1}{2}\right)\bigcap\left(X+Y\leq 1\right)\right]}{P\left(X+Y\leq 1\right)}.$$

$$P\left[X+Y\leq 1\right]=\int_0^1\left[\int_0^{1-x}(x+y)\,dy\right]\,dx$$

$$=\left[\frac{x^2}{2}-\frac{x^3}{3}-\frac{(1-x)^3}{6}\right]_0^1$$

$$=\frac{2}{6}=\frac{1}{3}.$$

Similarly

$$P\left[\left(X \leq \frac{1}{2}\right) \bigcap (X+Y \leq 1)\right] = \int_0^{\frac{1}{2}} \int_0^{1-x} (x+y)\, dy\, dx$$

$$= \left[\frac{x^2}{2} - \frac{x^3}{3} - \frac{(1-x)^3}{6}\right]_0^{\frac{1}{2}}$$

$$= \frac{11}{48}.$$

Thus,

$$P\left(2X \leq 1 \,/\, X+Y \leq 1\right) = \left(\frac{11}{48}\right)\ \left(\frac{3}{1}\right) = \frac{11}{16}.$$

### 7.3. Conditional Distributions

First, we motivate the definition of conditional distribution using discrete random variables and then based on this motivation we give a general definition of the conditional distribution. Let $X$ and $Y$ be two discrete random variables with joint probability density $f(x, y)$. Then by definition of the joint probability density, we have

$$f(x,\, y) = P(X = x,\, Y = y).$$

If $A = \{X = x\}$, $B = \{Y = y\}$ and $f_2(y) = P(Y = y)$, then from the above equation we have

$$\begin{aligned} P\left(\{X = x\} \,/\, \{Y = y\}\right) &= P\left(A \,/\, B\right) \\ &= \frac{P\left(A \bigcap B\right)}{P(B)} \\ &= \frac{P\left(\{X = x\} \text{ and } \{Y = y\}\right)}{P(Y = y)} \\ &= \frac{f(x,\, y)}{f_2(y)}. \end{aligned}$$

If we write the $P\left(\{X = x\} \,/\, \{Y = y\}\right)$ as $g(x \,/\, y)$, then we have

$$g(x \,/\, y) = \frac{f(x,\, y)}{f_2(y)}.$$

For the discrete bivariate random variables, we can write the conditional probability of the event $\{X = x\}$ given the event $\{Y = y\}$ as the ratio of the probability of the event $\{X = x\} \bigcap \{Y = y\}$ to the probability of the event $\{Y = y\}$ which is

$$g(x \,/\, y) = \frac{f(x,\, y)}{f_2(y)}.$$

We use this fact to define the conditional probability density function given two random variables $X$ and $Y$.

**Definition 7.8.** Let $X$ and $Y$ be any two random variables with joint density $f(x, y)$ and marginals $f_1(x)$ and $f_2(y)$. The *conditional probability density function* $g$ of $X$, given (the event) $Y = y$, is defined as

$$g(x \,/\, y) = \frac{f(x,\, y)}{f_2(y)} \qquad\qquad f_2(y) > 0.$$

Similarly, the *conditional probability density function* $h$ of $Y$, given (the event) $X = x$, is defined as

$$h(y \,/\, x) = \frac{f(x,\, y)}{f_1(x)} \qquad\qquad f_1(x) > 0.$$

**Example 7.14.** Let $X$ and $Y$ be discrete random variables with joint probability function

$$f(x,\, y) = \begin{cases} \frac{1}{21}\,(x + y) & \text{for } x = 1, 2, 3;\ y = 1, 2. \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the conditional probability density function of $X$, given $Y = 2$ ?

**Answer:** We want to find $g(x/2)$. Since

$$g(x \,/\, 2) = \frac{f(x,\, 2)}{f_2(2)}$$

we should first compute the marginal of $Y$, that is $f_2(2)$. The marginal of $Y$ is given by

$$\begin{aligned} f_2(y) &= \sum_{x=1}^{3} \frac{1}{21}\,(x + y) \\ &= \frac{1}{21}\,(6 + 3\,y). \end{aligned}$$

Hence $f_2(2) = \frac{12}{21}$. Thus, the conditional probability density function of $X$, given $Y = 2$, is

$$\begin{aligned} g(x/2) &= \frac{f(x,\,2)}{f_2(2)} \\ &= \frac{\frac{1}{21}\,(x+2)}{\frac{12}{21}} \\ &= \frac{1}{12}\,(x+2), \qquad\qquad x = 1, 2, 3. \end{aligned}$$

**Example 7.15.** Let $X$ and $Y$ be discrete random variables with joint probability density function

$$f(x,y) = \begin{cases} \frac{x+y}{32} & \text{for } x = 1, 2;\ y = 1, 2, 3, 4 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional probability of $Y$ given $X = x$ ?

**Answer:**

$$\begin{aligned} f_1(x) &= \sum_{y=1}^{4} f(x,y) \\ &= \frac{1}{32}\sum_{y=1}^{4}(x+y) \\ &= \frac{1}{32}\,(4\,x+10). \end{aligned}$$

Therefore

$$\begin{aligned} h(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{\frac{1}{32}\,(x+y)}{\frac{1}{32}\,(4\,x+10)} \\ &= \frac{x+y}{4x+10}. \end{aligned}$$

Thus, the conditional probability $Y$ given $X = x$ is

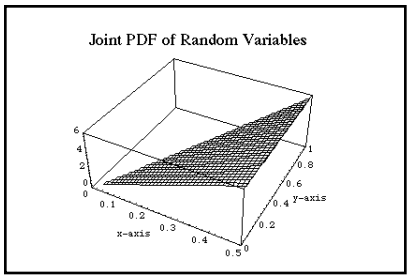
$$h(y/x) = \begin{cases} \frac{x+y}{4x+10} & \text{for } x = 1, 2;\ y = 1, 2, 3, 4 \\ \\ 0 & \text{otherwise.} \end{cases}$$

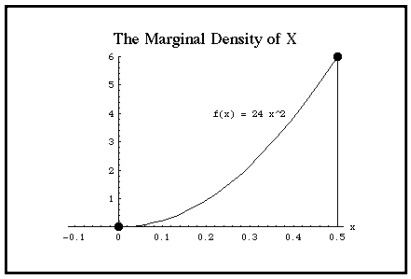
**Example 7.16.** Let $X$ and $Y$ be continuous random variables with joint pdf

$$f(x,y) = \begin{cases} 12\,x & \text{for } 0 < y < 2x < 1 \\ \\ 0 & \text{otherwise .} \end{cases}$$

What is the conditional density function of $Y$ given $X = x$ ?

**Answer:** First, we have to find the marginal of $X$.



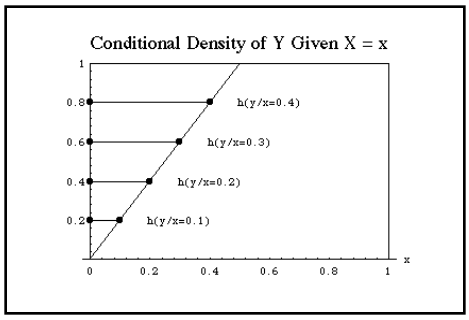




$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \int_0^{2x} 12\,x\,dy \\ &= 24\,x^2. \end{aligned}$$

Thus, the conditional density of $Y$ given $X = x$ is

$$\begin{aligned} h(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{12\,x}{24\,x^2} \\ &= \frac{1}{2x}, \qquad \text{for} \quad 0 < y < 2x < 1 \end{aligned}$$

and zero elsewhere.




**Example 7.17.** Let $X$ and $Y$ be random variables such that $X$ has density function

$$f_1(x) = \begin{cases} 24\,x^2 & \text{for } 0 < x < \frac{1}{2} \\ \\ 0 & \text{elsewhere} \end{cases}$$

and the conditional density of $Y$ given $X = x$ is

$$h(y/x) = \begin{cases} \frac{y}{2\,x^2} & \text{for } 0 < y < 2x \\ 0 & \text{elsewhere .} \end{cases}$$

What is the conditional density of $X$ given $Y = y$ over the appropriate domain?

**Answer:** The joint density $f(x, y)$ of $X$ and $Y$ is given by

$$\begin{aligned} f(x, y) &= h(y/x)\, f_1(x) \\ &= \frac{y}{2\,x^2}\, 24\, x^2 \\ &= 12y \qquad \text{for} \qquad 0 < y < 2x < 1. \end{aligned}$$

The marginal density of $Y$ is given by

$$\begin{aligned} f_2(y) &= \int_{-\infty}^{\infty} f(x,\, y)\, dx \\ &= \int_{\frac{y}{2}}^{\frac{1}{2}} 12\, y\, dx \\ &= 6\, y\, (1 - y), \qquad \text{for} \qquad 0 < y < 1. \end{aligned}$$

Hence, the conditional density of $X$ given $Y = y$ is

$$\begin{aligned} g(x/y) &= \frac{f(x,\, y)}{f_2(y)} \\ &= \frac{12y}{6\, y\, (1 - y)} \\ &= \frac{2}{1 - y}. \end{aligned}$$

Thus, the conditional density of $X$ given $Y = y$ is given by

$$g(x/y) = \begin{cases} \frac{2}{1-y} & \text{for } 0 < y < 2x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note that for a specific $x$, the function $f(x, y)$ is the intersection (profile) of the surface $z = f(x, y)$ by the plane $x =$ constant. The conditional density $f(y/x)$, is the profile of $f(x, y)$ normalized by the factor $\frac{1}{f_1(x)}$.

### 7.4. Independence of Random Variables

In this section, we define the concept of stochastic independence of two random variables $X$ and $Y$. The conditional probability density function $g$ of $X$ given $Y = y$ usually depends on $y$. If $g$ is independent of $y$, then the random variables $X$ and $Y$ are said to be independent. This motivates the following definition.

**Definition 7.8.** Let $X$ and $Y$ be any two random variables with joint density $f(x, y)$ and marginals $f_1(x)$ and $f_2(y)$. The random variables $X$ and $Y$ are (stochastically) independent if and only if

$$f(x,\, y) = f_1(x)\, f_2(y)$$

for all $(x,\, y) \in R_X \times R_Y$.

**Example 7.18.** Let $X$ and $Y$ be discrete random variables with joint density

$$f(x, y) = \begin{cases} \frac{1}{36} & \text{for } 1 \le x = y \le 6 \\ \\ \frac{2}{36} & \text{for } 1 \le x < y \le 6. \end{cases}$$

Are $X$ and $Y$ stochastically independent?

**Answer:** The marginals of $X$ and $Y$ are given by

$$\begin{aligned} f_1(x) &= \sum_{y=1}^{6} f(x,\, y) \\ &= f(x,x) + \sum_{y>x} f(x,y) + \sum_{y<x} f(x,y) \\ &= \frac{1}{36} + (6-x)\,\frac{2}{36} + 0 \\ &= \frac{13-2x}{36}, \qquad \text{for} \quad x = 1, 2, ..., 6 \end{aligned}$$

and

$$\begin{aligned} f_2(y) &= \sum_{x=1}^{6} f(x,\, y) \\ &= f(y,y) + \sum_{x<y} f(x,y) + \sum_{x>y} f(x,y) \\ &= \frac{1}{36} + (y-1)\,\frac{2}{36} + 0 \\ &= \frac{2y-1}{36}, \qquad \text{for} \quad y = 1, 2, ..., 6. \end{aligned}$$

Since

$$f(1,1) = \frac{1}{36} \neq \frac{11}{36}\,\frac{1}{36} = f_1(1)\,f_2(1),$$

we conclude that $f(x,y) \neq f_1(x)\,f_2(y)$, and $X$ and $Y$ are not independent.

This example also illustrates that the marginals of $X$ and $Y$ can be determined if one knows the joint density $f(x,\,y)$. However, if one knows the marginals of $X$ and $Y$, then it is not possible to find the joint density of $X$ and $Y$ unless the random variables are independent.

**Example 7.19.** Let $X$ and $Y$ have the joint density

$$f(x,\,y) = \begin{cases} e^{-(x+y)} & \text{for } 0 < x, y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

Are $X$ and $Y$ stochastically independent?

**Answer:** The marginals of $X$ and $Y$ are given by

$$f_1(x) = \int_0^\infty f(x,y)\,dy = \int_0^\infty e^{-(x+y)}\,dy = e^{-x}$$

and

$$f_2(y) = \int_0^\infty f(x,y)\,dx = \int_0^\infty e^{-(x+y)}\,dx = e^{-y}.$$

Hence

$$f(x,y) = e^{-(x+y)} = e^{-x}\,e^{-y} = f_1(x)\,f_2(y).$$

Thus, $X$ and $Y$ are stochastically independent.

Notice that if the joint density $f(x,y)$ of $X$ and $Y$ can be factored into two nonnegative functions, one solely depending on $x$ and the other solely depending on $y$, then $X$ and $Y$ are independent. We can use this factorization approach to predict when $X$ and $Y$ are not independent.

**Example 7.20.** Let $X$ and $Y$ have the joint density

$$f(x,\,y) = \begin{cases} x + y & \text{for } 0 < x < 1;\ 0 < y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Are $X$ and $Y$ stochastically independent?

**Answer:** Notice that

$$f(x,\,y) = x + y$$

$$= x\left(1 + \frac{y}{x}\right).$$

Thus, the joint density cannot be factored into two nonnegative functions one depending on $x$ and the other depending on $y$; and therefore $X$ and $Y$ are not independent.

If $X$ and $Y$ are independent, then the random variables $U = \phi(X)$ and $V = \psi(Y)$ are also independent. Here $\phi$, $\psi : \mathbb{R} \to \mathbb{R}$ are some real valued functions. From this comment, one can conclude that if $X$ and $Y$ are independent, then the random variables $e^X$ and $Y^3 + Y^2 + 1$ are also independent.

**Definition 7.9.** The random variables $X$ and $Y$ are said to be independent and identically distributed (IID) if and only if they are independent and have the same distribution.

**Example 7.21.** Let $X$ and $Y$ be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the probability density function of $W = \min\{X, Y\}$ ?

**Answer:** Let $G(w)$ be the cumulative distribution function of $W$. Then

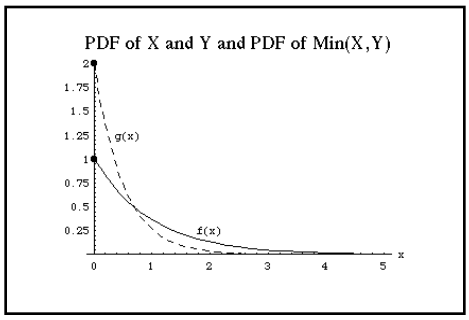
$$\begin{aligned} G(w) &= P(W \leq w) \\ &= 1 - P(W > w) \\ &= 1 - P(\min\{X, Y\} > w) \\ &= 1 - P(X > w \,\text{and}\, Y > w) \\ &= 1 - P(X > w)\, P(Y > w) \qquad \text{(since } X \text{ and } Y \text{ are independent)} \\ &= 1 - \left( \int_w^\infty e^{-x}\, dx \right) \left( \int_w^\infty e^{-y}\, dy \right) \\ &= 1 - \left( e^{-w} \right)^2 \\ &= 1 - e^{-2w}. \end{aligned}$$

Thus, the probability density function of $W$ is

$$g(w) = \frac{d}{dw} G(w) = \frac{d}{dw} \left( 1 - e^{-2w} \right) = 2\, e^{-2w}.$$

Hence

$$g(w) = \begin{cases} 2\, e^{-2w} & \text{for } w > 0 \\ \\ 0 & \text{elsewhere.} \end{cases}$$




### 7.5. Review Exercises

**1.** Let $X$ and $Y$ be discrete random variables with joint probability density function

$$f(x,y) = \begin{cases} \frac{1}{21}(x+y) & \text{for } x = 1,2,3;\ y = 1,2 \\ 0 & \text{otherwise.} \end{cases}$$

What are the marginals of $X$ and $Y$?

**2.** Roll a pair of unbiased dice. Let $X$ be the maximum of the two faces and $Y$ be the sum of the two faces. What is the joint density of $X$ and $Y$ ?

**3.** For what value of $c$ is the real valued function

$$f(x,y) = \begin{cases} c(x+2y) & \text{for } x = 1,2;\ y = 1,2 \\ 0 & \text{otherwise} \end{cases}$$

a joint density for some random variables $X$ and $Y$ ?

**4.** Let $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} e^{-(x+y)} & \text{for } 0 \le x, y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is $P(X \ge Y \ge 2)$ ?

**5.** If the random variable $X$ is uniform on the interval from $-1$ to 1, and the random variable $Y$ is uniform on the interval from 0 to 1, what is the probability that the the quadratic equation $t^2 + 2Xt + Y = 0$ has real solutions? Assume $X$ and $Y$ are independent.

**6.** Let $Y$ have a uniform distribution on the interval $(0,1)$, and let the conditional density of $X$ given $Y = y$ be uniform on the interval from 0 to $\sqrt{y}$. What is the marginal density of $X$ for $0 < x < 1$?

**7.** If the joint cumulative distribution of the random variables $X$ and $Y$ is

$$F(x,y) = \begin{cases} (1-e^{-x})(1-e^{-y}) & \text{for } x>0,\ y>0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the joint probability density function of the random variables $X$ and $Y$, and the $P(1<X<3,\, 1<Y<2)$?

**8.** If the random variables $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} \frac{6}{7}\,x & \text{for } 1 \le x+y \le 2,\ x \ge 0,\ y \ge 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the probability $P(Y \ge X^2)$ ?

**9.** If the random variables $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} \frac{6}{7}\,x & \text{for } 1 \le x+y \le 2,\ x \ge 0,\ y \ge 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the probability $P[\max(X,Y) > 1]$ ?

**10.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} \frac{5}{16}\,xy^2 & \text{for } 0<x<y<2 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the marginal density function of $X$ where it is nonzero?

**11.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} 4x & \text{for } 0<x<\sqrt{y}<1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the marginal density function of $Y$, where nonzero?

**12.** A point $(X,Y)$ is chosen at random from a uniform distribution on the circular disk of radius centered at the point $(1,1)$. For a given value of $X=x$ between 0 and 2 and for $y$ in the appropriate domain, what is the conditional density function for $Y$?

**13.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} \frac{3}{4}\,(2 - x - y) & \text{for } 0 < x, y < 2;\ 0 < x + y < 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional probability $P\,(X < 1 \,|\, Y < 1)$ ?

**14.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} 12x & \text{for } 0 < y < 2x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional density function of $Y$ given $X = x$ ?

**15.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} 24xy & \text{for } x > 0,\ y > 0,\ 0 < x + y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional probability $P\left(X < \frac{1}{2} \,|\, Y = \frac{1}{4}\right)$ ?

**16.** Let $X$ and $Y$ be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the probability density function of $W = \max\{X, Y\}$ ?

**17.** Let $X$ and $Y$ be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} \frac{3\,x^2}{\theta^3} & \text{for } 0 \le x \le \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

for some $\theta > 0$. What is the probability density function of $W = \min\{X, Y\}$?

**18.** Ron and Glenna agree to meet between 5 P.M. and 6 P.M. Suppose that each of them arrive at a time distributed uniformly at random in this time interval, independent of the other. Each will wait for the other at most 10 minutes (and if other does not show up they will leave). What is the probability that they actually go out?

**19.** Let $X$ and $Y$ be two independent random variables distributed uniformly on the interval $[0, 1]$. What is the probability of the event $Y \geq \frac{1}{2}$ given that $Y \geq 1 - 2X$?

**20.** Let $X$ and $Y$ have the joint density

$$f(x, y) = \begin{cases} 8xy & \text{for } 0 < y < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is $P(X + Y > 1)$ ?

**21.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} 2 & \text{for } 0 \leq y \leq x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Are $X$ and $Y$ stochastically independent?

**22.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} 2x & \text{for } 0 < x,\, y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Are $X$ and $Y$ stochastically independent?

**23.** A bus and a passenger arrive at a bus stop at a uniformly distributed time over the interval 0 to 1 hour. Assume the arrival times of the bus and passenger are independent of one another and that the passenger will wait up to 15 minutes for the bus to arrive. What is the probability that the passenger will catch the bus?

**24.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} 4xy & \text{for } 0 \leq x,\, y \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability of the event $X \leq \frac{1}{2}$ given that $Y \geq \frac{3}{4}$?

**25.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,\, y) = \begin{cases} \frac{1}{2} & \text{for } 0 \leq x \leq y \leq 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability of the event $X \leq \frac{1}{2}$ given that $Y = 1$?

**26.** If the joint density of the random variables $X$ and $Y$ is

$$f(x,y) = \begin{cases} 1 & \text{if } 0 \le x \le y \le 1 \\ \frac{1}{2} & \text{if } 1 \le x \le 2,\ \ 0 \le y \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

what is the probability of the event $\left(X \le \frac{3}{2},\ Y \le \frac{1}{2}\right)$?

**27.** If the joint density of the random variables $X$ and $Y$ is

$$f(x,y) = \begin{cases} \left[e^{\min\{x,y\}} - 1\right] e^{-(x+y)} & \text{if } 0 < x, y < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the marginal density function of $X$, where nonzero?

# Chapter 8

# PRODUCT MOMENTS OF BIVARIATE RANDOM VARIABLES

In this chapter, we define various product moments of a bivariate random variable. The main concept we introduce in this chapter is the notion of covariance between two random variables. Using this notion, we study the statistical dependence of two random variables.

## 8.1. Covariance of Bivariate Random Variables

First, we define the notion of product moment of two random variables and then using this product moment, we give the definition of covariance between two random variables.

**Definition 8.1.** Let $X$ and $Y$ be any two random variables with joint density function $f(x, y)$. The product moment of $X$ and $Y$, denoted by $E(XY)$, is defined as

$$E(XY) = \begin{cases} \sum\limits_{x \in R_X} \sum\limits_{y \in R_Y} xy\, f(x,\, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x, y)\, dx\, dy & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases}$$

Here, $R_X$ and $R_Y$ represent the range spaces of $X$ and $Y$ respectively.

**Definition 8.2.** Let $X$ and $Y$ be any two random variables with joint density function $f(x, y)$. The covariance between $X$ and $Y$, denoted by $Cov(X,\, Y)$ (or $\sigma_{XY}$), is defined as

$$Cov(X,\, Y) = E(\,(X - \mu_X)\,(Y - \mu_Y)\,),$$

where $\mu_X$ and $\mu_Y$ are mean of $X$ and $Y$, respectively.

Notice that the covariance of $X$ and $Y$ is really the product moment of $X - \mu_X$ and $Y - \mu_Y$. Further, the mean of $\mu_X$ is given by

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x\, f_1(x)\, dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dx\, dy,$$

and similarly the mean of $Y$ is given by

$$\mu_Y = E(Y) = \int_{-\infty}^{\infty} y\, f_2(y)\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\, dy\, dx.$$

**Theorem 8.1.** Let $X$ and $Y$ be any two random variables. Then

$$Cov(X,\, Y) = E(XY) - E(X)\, E(Y).$$

**Proof:**

$$\begin{aligned} Cov(X,\, Y) &= E((X - \mu_X)\,(Y - \mu_Y)) \\ &= E(XY - \mu_X\, Y - \mu_Y\, X + \mu_X\, \mu_Y) \\ &= E(XY) - \mu_X\, E(Y) - \mu_Y\, E(X) + \mu_X\, \mu_Y \\ &= E(XY) - \mu_X\, \mu_Y - \mu_Y\, \mu_X + \mu_X\, \mu_Y \\ &= E(XY) - \mu_X\, \mu_Y \\ &= E(XY) - E(X)\, E(Y). \end{aligned}$$

**Corollary 8.1.** $Cov(X, X) = \sigma_X^2$.

**Proof:**

$$\begin{aligned} Cov(X, X) &= E(XX) - E(X)\, E(X) \\ &= E(X^2) - \mu_X^2 \\ &= Var(X) \\ &= \sigma_X^2. \end{aligned}$$

**Example 8.1.** Let $X$ and $Y$ be discrete random variables with joint density

$$f(x,y) = \begin{cases} \frac{x+2y}{18} & \text{for } x = 1, 2;\ y = 1, 2 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the covariance $\sigma_{XY}$ between $X$ and $Y$.

**Answer:** The marginal of $X$ is

$$f_1(x) = \sum_{y=1}^{2} \frac{x+2y}{18} = \frac{1}{18}\,(2x+6).$$

Hence the expected value of $X$ is

$$\begin{aligned} E(X) &= \sum_{x=1}^{2} x\, f_1(x) \\ &= 1\, f_1(1) + 2 f_1(2) \\ &= \frac{8}{18} + 2\ \frac{10}{18} \\ &= \frac{28}{18}. \end{aligned}$$

Similarly, the marginal of $Y$ is

$$f_2(y) = \sum_{x=1}^{2} \frac{x+2y}{18} = \frac{1}{18}\,(3+4y).$$

Hence the expected value of $Y$ is

$$\begin{aligned} E(Y) &= \sum_{y=1}^{2} y\, f_2(y) \\ &= 1\, f_2(1) + 2 f_2(2) \\ &= \frac{7}{18} + 2\ \frac{11}{18} \\ &= \frac{29}{18}. \end{aligned}$$

Further, the product moment of $X$ and $Y$ is given by

$$\begin{aligned} E(XY) &= \sum_{x=1}^{2}\sum_{y=1}^{2} x\, y\, f(x,\, y) \\ &= f(1,1) + 2\, f(1,2) + 2\, f(2,1) + 4\, f(2,2) \\ &= \frac{3}{18} + 2\ \frac{5}{18} + 2\ \frac{4}{18} + 4\ \frac{6}{18} \\ &= \frac{3+10+8+24}{18} \\ &= \frac{45}{18}. \end{aligned}$$

Hence, the covariance between $X$ and $Y$ is given by

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)\,E(Y) \\ &= \frac{45}{18} - \left(\frac{28}{18}\right)\left(\frac{29}{18}\right) \\ &= \frac{(45)\,(18) - (28)\,(29)}{(18)\,(18)} \\ &= \frac{810 - 812}{324} \\ &= -\frac{2}{324} = -0.00617. \end{aligned}$$

**Remark 8.1.** For an arbitrary random variable, the product moment and covariance may or may not exist. Further, note that unlike variance, the covariance between two random variables may be negative.

**Example 8.2.** Let $X$ and $Y$ have the joint density function

$$f(x, y) = \begin{cases} x + y & \text{if } 0 < x, y < 1 \\ \\ 0 & \text{elsewhere .} \end{cases}$$

What is the covariance between $X$ and $Y$ ?

**Answer:** The marginal density of $X$ is

$$\begin{aligned} f_1(x) &= \int_0^1 (x + y)\, dy \\ &= \left[x\, y + \frac{y^2}{2}\right]_{y=0}^{y=1} \\ &= x + \frac{1}{2}. \end{aligned}$$

Thus, the expected value of $X$ is given by

$$\begin{aligned} E(X) &= \int_0^1 x\, f_1(x)\, dx \\ &= \int_0^1 x\,(x + \frac{1}{2})\, dx \\ &= \left[\frac{x^3}{3} + \frac{x^2}{4}\right]_0^1 \\ &= \frac{7}{12}. \end{aligned}$$

Similarly (or using the fact that the density is symmetric in $x$ and $y$), we get

$$E(Y) = \frac{7}{12}.$$

Now, we compute the product moment of $X$ and $Y$.

$$\begin{aligned}
E(XY) &= \int_0^1 \int_0^1 x\, y(x+y)\, dx\, dy \\
&= \int_0^1 \int_0^1 (x^2\, y + x\, y^2)\, dx\, dy \\
&= \int_0^1 \left[ \frac{x^3\, y}{3} + \frac{x^2\, y^2}{2} \right]_{x=0}^{x=1} dy \\
&= \int_0^1 \left( \frac{y}{3} + \frac{y^2}{2} \right) dy \\
&= \left[ \frac{y^2}{6} + \frac{y^3}{6} \right]_0^1 dy \\
&= \frac{1}{6} + \frac{1}{6} \\
&= \frac{4}{12}.
\end{aligned}$$

Hence the covariance between $X$ and $Y$ is given by

$$\begin{aligned}
Cov(X,Y) &= E(XY) - E(X)\, E(Y) \\
&= \frac{4}{12} - \left( \frac{7}{12} \right) \left( \frac{7}{12} \right) \\
&= \frac{48-49}{144} \\
&= -\frac{1}{144}.
\end{aligned}$$

**Example 8.3.** Let $X$ and $Y$ be continuous random variables with joint density function

$$f(x,y) = \begin{cases} 2 & \text{if } 0 < y < 1-x;\ 0 < x < 1 \\ \\ 0 & \text{elsewhere .} \end{cases}$$

What is the covariance between $X$ and $Y$ ?

**Answer:** The marginal density of $X$ is given by

$$f_1(x) = \int_0^{1-x} 2\, dy = 2\,(1-x).$$

Hence the expected value of $X$ is

$$\mu_X = E(X) = \int_0^1 x\, f_1(x)\, dx = \int_o^1 2\,(1-x)\, dx = \frac{1}{3}.$$

Similarly, the marginal of $Y$ is

$$f_2(y) = \int_0^{1-y} 2\, dx = 2\,(1-y).$$

Hence the expected value of $Y$ is

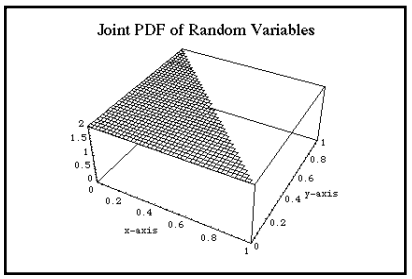
$$\mu_Y = E(Y) = \int_0^1 y\, f_2(y)\, dy = \int_o^1 2\,(1-y)\, dy = \frac{1}{3}.$$

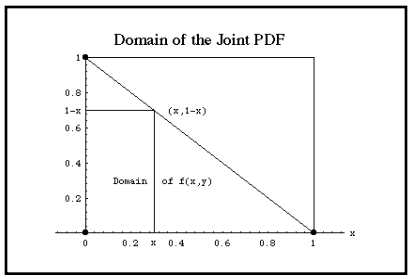
The product moment of $X$ and $Y$ is given by

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^{1-x} x\, y\, f(x,y)\, dy\, dx \\ &= \int_0^1 \int_0^{1-x} x\, y\, 2\, dy\, dx \\ &= 2 \int_0^1 x \left[\frac{y^2}{2}\right]_0^{1-x} dx \\ &= 2\,\frac{1}{2} \int_0^1 x\,(1-x)^2\, dx \\ &= \int_0^1 \left(x - 2x^2 + x^3\right)\, dx \\ &= \left[\frac{1}{2}\,x^2 - \frac{2}{3}\,x^3 + \frac{1}{4}\,x^4\right]_0^1 \\ &= \frac{1}{12}. \end{aligned}$$

Therefore, the covariance between $X$ and $Y$ is given by

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)\,E(Y) \\ &= \frac{1}{12} - \frac{1}{9} \\ &= \frac{3}{36} - \frac{4}{36} = -\frac{1}{36}. \end{aligned}$$



**Theorem 8.2.** If $X$ and $Y$ are any two random variables and $a$, $b$, $c$, and $d$ are real constants, then

$$Cov(a\,X + b,\, c\,Y + d) = a\,c\,Cov(X, Y).$$

**Proof:**

$$\begin{aligned}
&Cov(a\,X + b,\, c\,Y + d)\\
&\quad = E\left((aX + b)(cY + d)\right) - E(aX + b)\,E(cY + d)\\
&\quad = E\left(acXY + adX + bcY + bd\right) - \left(aE(X) + b\right)\,\left(cE(Y) + d\right)\\
&\quad = ac\,E(XY) + ad\,E(X) + bc\,E(Y) + bd\\
&\qquad\quad - \left[ac\,E(X)\,E(Y) + ad\,E(X) + bc\,E(Y) + bd\right]\\
&\quad = ac\,\left[E(XY) - E(X)\,E(Y)\right]\\
&\quad = ac\,Cov(X,\, Y).
\end{aligned}$$

**Example 8.4.** If the product moment of $X$ and $Y$ is 3 and the mean of $X$ and $Y$ are both equal to 2, then what is the covariance of the random variables $2X + 10$ and $-\frac{5}{2}\,Y + 3$ ?

**Answer:** Since $E(XY) = 3$ and $E(X) = 2 = E(Y)$, the covariance of $X$ and $Y$ is given by

$$Cov(X, Y) = E(XY) - E(X)\,E(Y) = 3 - 4 = -1.$$

Then the covariance of $2X + 10$ and $-\frac{5}{2}\,Y + 3$ is given by

$$\begin{aligned}
Cov\left(2X + 10,\, -\frac{5}{2}Y + 3\right) &= 2\,\left(-\frac{5}{2}\right)\,Cov(X, Y)\\
&= (-5)\,(-1)\\
&= 5.
\end{aligned}$$

**Remark 8.2.** Notice that the Theorem 8.2 can be furthered improved. That is, if $X$, $Y$, $Z$ are three random variables, then

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$$

and

$$Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z).$$

The first formula can be established as follows. Consider

$$\begin{aligned} Cov(X+Y,Z) &= E((X+Y)Z) - E(X+Y)\,E(Z) \\ &= E(XZ+YZ) - E(X)E(Z) - E(Y)E(Z) \\ &= E(XZ) - E(X)E(Z) + E(YZ) - E(Y)E(Z) \\ &= Cov(X,Z) + Cov(Y,Z). \end{aligned}$$

## 8.2. Independence of Random Variables

In this section, we study the effect of independence on the product moment (and hence on the covariance). We begin with a simple theorem.

**Theorem 8.3.** If $X$ and $Y$ are independent random variables, then

$$E(XY) = E(X)\,E(Y).$$

**Proof:** Recall that $X$ and $Y$ are independent if and only if

$$f(x,y) = f_1(x)\,f_2(y).$$

Let us assume that $X$ and $Y$ are continuous. Therefore

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_1(x)\,f_2(y)\,dx\,dy \\ &= \left(\int_{-\infty}^{\infty} x\,f_1(x)\,dx\right)\left(\int_{-\infty}^{\infty} y\,f_2(y)\,dy\right) \\ &= E(X)\,E(Y). \end{aligned}$$

If $X$ and $Y$ are discrete, then replace the integrals by appropriate sums to prove the same result.

**Example 8.5.** Let $X$ and $Y$ be two independent random variables with respective density functions:

$$f(x) = \begin{cases} 3\,x^2 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(y) = \begin{cases} 4\,y^3 & \text{if } 0 < y < 1 \\ 0 & \text{otherwise .} \end{cases}$$

What is $E\left(\frac{X}{Y}\right)$ ?

**Answer:** Since $X$ and $Y$ are independent, the joint density of $X$ and $Y$ is given by

$$h(x,y) = f(x)\, g(y).$$

Therefore

$$\begin{aligned} E\left(\frac{X}{Y}\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{x}{y}\, h(x,y)\, dx\, dy \\ &= \int_0^1 \int_0^1 \frac{x}{y}\, f(x)\, g(y)\, dx\, dy \\ &= \int_0^1 \int_0^1 \frac{x}{y}\, 3x^2\, 4y^3\, dx\, dy \\ &= \left(\int_0^1 3x^3\, dx\right)\, \left(\int_0^1 4y^2\, dy\right) \\ &= \left(\frac{3}{4}\right)\, \left(\frac{4}{3}\right) = 1. \end{aligned}$$

**Remark 8.3.** The independence of $X$ and $Y$ does not imply $E\left(\frac{X}{Y}\right) = \frac{E(X)}{E(Y)}$ but only implies $E\left(\frac{X}{Y}\right) = E(X)\, E\left(Y^{-1}\right)$. Further, note that $E\left(Y^{-1}\right)$ is not equal to $\frac{1}{E(Y)}$.

**Theorem 8.4.** If $X$ and $Y$ are independent random variables, then the covariance between $X$ and $Y$ is always zero, that is

$$Cov(X,Y) = 0.$$

**Proof:** Suppose $X$ and $Y$ are independent, then by Theorem 8.3, we have $E(XY) = E(X)\, E(Y)$. Consider

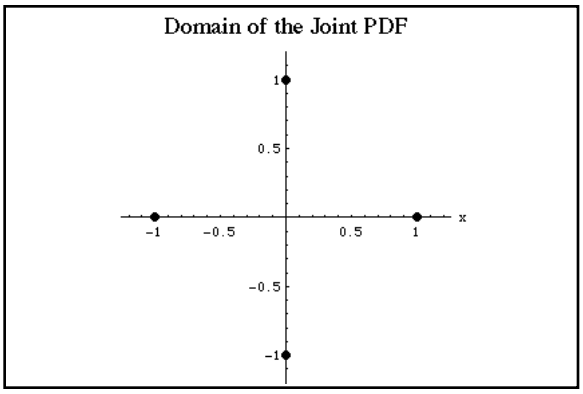
$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)\, E(Y) \\ &= E(X)\, E(Y) - E(X)\, E(Y) \\ &= 0. \end{aligned}$$

**Example 8.6.** Let the random variables $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} \frac{1}{4} & \text{if } (x,y) \in \{\,(0,1),\,(0,-1),\,(1,0),\,(-1,0)\,\} \\ 0 & \text{otherwise.} \end{cases}$$

What is the covariance of $X$ and $Y$ ? Are the random variables $X$ and $Y$ independent?

**Answer:** The joint density of $X$ and $Y$ are shown in the following table with the marginals $f_1(x)$ and $f_2(y)$.




| $(x, y)$ | $-1$ | $0$ | $1$ | $f_2(y)$ |
|---|---|---|---|---|
| $-1$ | $0$ | $\frac{1}{4}$ | $0$ | $\frac{1}{4}$ |
| $0$ | $\frac{1}{4}$ | $0$ | $\frac{1}{4}$ | $\frac{2}{4}$ |
| $1$ | $0$ | $\frac{1}{4}$ | $0$ | $\frac{1}{4}$ |
| $f_1(x)$ | $\frac{1}{4}$ | $\frac{2}{4}$ | $\frac{1}{4}$ | |

From this table, we see that

$$0 = f(0,0) \neq f_1(0)\, f_2(0) = \left(\frac{2}{4}\right)\ \left(\frac{2}{4}\right) = \frac{1}{4}$$

and thus

$$f(x, y) \neq f_1(x)\, f_2(y)$$

for all $(x, y)$ is the range space of the joint variable $(X, Y)$. Therefore $X$ and $Y$ are not independent.

Next, we compute the covariance between $X$ and $Y$. For this we need

$E(X)$, $E(Y)$ and $E(XY)$. The expected value of $X$ is

$$\begin{aligned} E(X) &= \sum_{x=-1}^{1} x f_1(x) \\ &= (-1)\, f_1(-1) + (0) f_1(0) + (1)\, f_1(1) \\ &= -\frac{1}{4} + 0 + \frac{1}{4} \\ &= 0. \end{aligned}$$

Similarly, the expected value of $Y$ is

$$\begin{aligned} E(Y) &= \sum_{y=-1}^{1} y f_2(y) \\ &= (-1)\, f_2(-1) + (0) f_2(0) + (1)\, f_2(1) \\ &= -\frac{1}{4} + 0 + \frac{1}{4} \\ &= 0. \end{aligned}$$

The product moment of $X$ and $Y$ is given by

$$\begin{aligned} E(XY) &= \sum_{x=-1}^{1} \sum_{y=-1}^{1} x\, y\, f(x, y) \\ &= (1)\, f(-1,-1) + (0)\, f(-1,0) + (-1)\, f(-1,1) \\ &+ (0)\, f(0,-1) + (0)\, f(0,0) + (0)\, f(0,1) \\ &+ (-1)\, f(1,-1) + (0)\, f(1,0) + (1)\, f(1,1) \\ &= 0. \end{aligned}$$

Hence, the covariance between $X$ and $Y$ is given by

$$Cov(X, Y) = E(XY) - E(X)\, E(Y) = 0.$$

**Remark 8.4.** This example shows that if the covariance of $X$ and $Y$ is zero that does not mean the random variables are independent. However, we know from Theorem 8.4 that if $X$ and $Y$ are independent, then the $Cov(X, Y)$ is always zero.

### 8.3. Variance of the Linear Combination of Random Variables

Given two random variables, $X$ and $Y$, we determine the variance of their linear combination, that is $aX + bY$.

**Theorem 8.5.** Let $X$ and $Y$ be any two random variables and let $a$ and $b$ be any two real numbers. Then

$$Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) + 2\, a\, b\, Cov(X, Y).$$

**Proof:**

$$\begin{aligned}
&Var(aX + bY)\\
&= E\left(\left[aX + bY - E(aX + bY)\right]^2\right)\\
&= E\left(\left[aX + bY - a\,E(X) - b\,E(Y)\right]^2\right)\\
&= E\left(\left[a\,(X - \mu_X) + b\,(Y - \mu_Y)\right]^2\right)\\
&= E\left(a^2\,(X - \mu_X)^2 + b^2\,(Y - \mu_Y)^2 + 2\,a\,b\,(X - \mu_X)\,(Y - \mu_Y)\right)\\
&= a^2\,E\left((X - \mu_X)^2\right) + b^2\,E\left((X - \mu_X)^2\right) + 2\,a\,b\,E((X - \mu_X)\,(Y - \mu_Y))\\
&= a^2\,Var(X) + b^2\,Var(Y) + 2\,a\,b\,Cov(X, Y).
\end{aligned}$$

**Example 8.7.** If $Var(X + Y) = 3$, $Var(X - Y) = 1$, $E(X) = 1$ and $E(Y) = 2$, then what is $E(XY)$ ?

**Answer:**

$$\begin{aligned}
Var(X + Y) &= \sigma_X^2 + \sigma_Y^2 + 2\,Cov(X, Y),\\
Var(X - Y) &= \sigma_X^2 + \sigma_Y^2 - 2\,Cov(X, Y).
\end{aligned}$$

Hence, we get

$$\begin{aligned}
Cov(X, Y) &= \frac{1}{4}\left[\,Var(X + Y) - Var(X - Y)\,\right]\\
&= \frac{1}{4}\left[3 - 1\right]\\
&= \frac{1}{2}.
\end{aligned}$$

Therefore, the product moment of $X$ and $Y$ is given by

$$\begin{aligned}
E(XY) &= Cov(X, Y) + E(X)\,E(Y)\\
&= \frac{1}{2} + (1)\,(2)\\
&= \frac{5}{2}.
\end{aligned}$$

**Example 8.8.** Let $X$ and $Y$ be random variables with $Var(X) = 4$, $Var(Y) = 9$ and $Var(X - Y) = 16$. What is $Cov(X, Y)$ ?

**Answer:**

$$Var(X - Y) = Var(X) + Var(Y) - 2\,Cov(X, Y)$$
$$16 = 4 + 9 - 2\,Cov(X, Y).$$

Hence

$$Cov(X, Y) = -\frac{3}{2}.$$

**Remark 8.5.** The Theorem 8.5 can be extended to three or more random variables. In case of three random variables $X, Y, Z$, we have

$$\begin{aligned} &Var(X + Y + Z) \\ &= Var(X) + Var(Y) + Var(Z) \\ &\quad + 2Cov(X, Y) + 2Cov(Y, Z) + 2Cov(Z, X). \end{aligned}$$

To see this consider

$$\begin{aligned} &Var(X + Y + Z) \\ &= Var((X + Y) + Z) \\ &= Var(X + Y) + Var(Z) + 2Cov(X + Y, Z) \\ &= Var(X + Y) + Var(Z) + 2Cov(X, Z) + 2Cov(Y, Z) \\ &= Var(X) + Var(Y) + 2Cov(X, Y) \\ &\quad + Var(Z) + 2Cov(X, Z) + 2Cov(Y, Z) \\ &= Var(X) + Var(Y) + Var(Z) \\ &\quad + 2Cov(X, Y) + 2Cov(Y, Z) + 2Cov(Z, X). \end{aligned}$$

**Theorem 8.6.** If $X$ and $Y$ are independent random variables with $E(X) = 0 = E(Y)$, then

$$Var(XY) = Var(X)\,Var(Y).$$

**Proof:**

$$\begin{aligned} Var(XY) &= E\left((XY)^2\right) - \left(E(X)\,E(Y)\right)^2 \\ &= E\left((XY)^2\right) \\ &= E\left(X^2\,Y^2\right) \\ &= E\left(X^2\right)\,E\left(Y^2\right) \qquad \text{(by independence of } X \text{ and } Y) \\ &= Var(X)\,Var(Y). \end{aligned}$$

**Example 8.9.** Let $X$ and $Y$ be independent random variables, each with density

$$f(x) = \begin{cases} \frac{1}{2\theta} & \text{for } -\theta < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

If the $Var(XY) = \frac{64}{9}$, then what is the value of $\theta$ ?

**Answer:**

$$E(X) = \int_{-\theta}^{\theta} \frac{1}{2\theta}\, x\, dx = \frac{1}{2\theta}\left[\frac{x^2}{2}\right]_{-\theta}^{\theta} = 0.$$

Since $Y$ has the same density, we conclude that $E(Y) = 0$. Hence

$$\begin{aligned} \frac{64}{9} &= Var(XY) \\ &= Var(X)\, Var(Y) \\ &= \left(\int_{-\theta}^{\theta} \frac{1}{2\theta}\, x^2\, dx\right)\left(\int_{-\theta}^{\theta} \frac{1}{2\theta}\, y^2\, dy\right) \\ &= \left(\frac{\theta^2}{3}\right)\left(\frac{\theta^2}{3}\right) \\ &= \frac{\theta^4}{9}. \end{aligned}$$

Hence, we obtain

$$\theta^4 = 64 \qquad \text{or} \qquad \theta = 2\sqrt{2}.$$

### 8.4. Correlation and Independence

The functional dependency of the random variable $Y$ on the random variable $X$ can be obtained by examining the correlation coefficient. The definition of the correlation coefficient $\rho$ between $X$ and $Y$ is given below.

**Definition 8.3.** Let $X$ and $Y$ be two random variables with variances $\sigma_X^2$ and $\sigma_Y^2$, respectively. Let the covariance of $X$ and $Y$ be $Cov(X, Y)$. Then the correlation coefficient $\rho$ between $X$ and $Y$ is given by

$$\rho = \frac{Cov(X, Y)}{\sigma_X\, \sigma_Y}.$$

**Theorem 8.7.** If $X$ and $Y$ are independent, the correlation coefficient between $X$ and $Y$ is zero.

**Proof:**

$$\begin{aligned}\rho &= \frac{Cov(X,Y)}{\sigma_X\,\sigma_Y}\\ &= \frac{0}{\sigma_X\,\sigma_Y}\\ &= 0.\end{aligned}$$

**Remark 8.4.** The converse of this theorem is not true. If the correlation coefficient of $X$ and $Y$ is zero, then $X$ and $Y$ are said to be uncorrelated.

**Lemma 8.1.** If $X^\star$ and $Y^\star$ are the standardizations of the random variables $X$ and $Y$, respectively, the correlation coefficient between $X^\star$ and $Y^\star$ is equal to the correlation coefficient between $X$ and $Y$.

**Proof:** Let $\rho^\star$ be the correlation coefficient between $X^\star$ and $Y^\star$. Further, let $\rho$ denote the correlation coefficient between $X$ and $Y$. We will show that $\rho^\star = \rho$. Consider

$$\begin{aligned}\rho^\star &= \frac{Cov\,(X^\star, Y^\star)}{\sigma_{X^\star}\,\sigma_{Y^\star}}\\ &= Cov\,(X^\star, Y^\star)\\ &= Cov\left(\frac{X-\mu_X}{\sigma_X},\,\frac{Y-\mu_Y}{\sigma_Y}\right)\\ &= \frac{1}{\sigma_X\,\sigma_Y}\,Cov\,(X-\mu_X,\,Y-\mu_Y)\\ &= \frac{Cov\,(X,\,Y)}{\sigma_X\,\sigma_Y}\\ &= \rho.\end{aligned}$$

This lemma states that the value of the correlation coefficient between two random variables does not change by standardization of them.

**Theorem 8.8.** For any random variables $X$ and $Y$, the correlation coefficient $\rho$ satisfies

$$-1 \le \rho \le 1,$$

and $\rho = 1$ or $\rho = -1$ implies that the random variable $Y = a\,X + b$, where $a$ and $b$ are arbitrary real constants with $a \neq 0$.

**Proof:** Let $\mu_X$ be the mean of $X$ and $\mu_Y$ be the mean of $Y$, and $\sigma_X^2$ and $\sigma_Y^2$ be the variances of $X$ and $Y$, respectively. Further, let

$$X^* = \frac{X-\mu_X}{\sigma_X} \qquad \text{and} \qquad Y^* = \frac{Y-\mu_Y}{\sigma_Y}$$

be the standardization of $X$ and $Y$, respectively. Then

$$\mu_{X^*} = 0 \quad \text{and} \quad \sigma^2_{X^*} = 1,$$

and

$$\mu_{Y^*} = 0 \quad \text{and} \quad \sigma^2_{Y^*} = 1.$$

Thus

$$\begin{aligned} Var(X^* - Y^*) &= Var(X^*) + Var(Y^*) - 2Cov(X^*, Y^*) \\ &= \sigma^2_{X^*} + \sigma^2_{Y^*} - 2\,\rho^*\,\sigma_{X^*}\,\sigma_{Y^*} \\ &= 1 + 1 - 2\rho^* \\ &= 1 + 1 - 2\rho \qquad \text{(by Lemma 8.1)} \\ &= 2(1 - \rho). \end{aligned}$$

Since the variance of a random variable is always positive, we get

$$2\,(1 - \rho) \geq 0$$

which is

$$\rho \leq 1.$$

By a similar argument, using $Var(X^* + Y^*)$, one can show that $-1 \leq \rho$. Hence, we have $-1 \leq \rho \leq 1$. Now, we show that if $\rho = 1$ or $\rho = -1$, then $Y$ and $X$ are related through an affine transformation. Consider the case $\rho = 1$, then

$$Var(X^* - Y^*) = 0.$$

But if the variance of a random variable is 0, then all the probability mass is concentrated at a point (that is, the distribution of the corresponding random variable is degenerate). Thus $Var(X^* - Y^*) = 0$ implies $X^* - Y^*$ takes only one value. But $E\,[X^* - Y^*] = 0$. Thus, we get

$$X^* - Y^* \equiv 0$$

or

$$X^* \equiv Y^*.$$

Hence

$$\frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y}.$$

Solving this for $Y$ in terms of $X$, we get

$$Y = a\,X + b$$

where

$$a = \frac{\sigma_Y}{\sigma_X} \qquad \text{and} \qquad b = \mu_Y - a\,\mu_X.$$

Thus if $\rho = 1$, then $Y$ is a linear in $X$. Similarly, we can show for the case $\rho = -1$, the random variables $X$ and $Y$ are linearly related. This completes the proof of the theorem.

### 8.5. Moment Generating Functions

Similar to the moment generating function for the univariate case, one can define the moment generating function for the bivariate case to compute the various product moments. The moment generating function for the bivariate case is defined as follows:

**Definition 8.4.** Let $X$ and $Y$ be two random variables with joint density function $f(x, y)$. A real valued function $M : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$M(s,\, t) = E\left(e^{sX+tY}\right)$$

is called the joint moment generating function of $X$ and $Y$ if this expected value exists for all $s$ is some interval $-h < s < h$ and for all $t$ is some interval $-k < t < k$ for some positive $h$ and $k$.

It is easy to see from this definition that

$$M(s, 0) = E\left(e^{sX}\right)$$

and

$$M(0, t) = E\left(e^{tY}\right).$$

From this we see that

$$E(X^k) = \left.\frac{\partial^k M(s,t)}{\partial s^k}\right|_{(0,0)}, \qquad E(Y^k) = \left.\frac{\partial^k M(s,t)}{\partial t^k}\right|_{(0,0)},$$

for $k = 1, 2, 3, 4, ...$; and

$$E(XY) = \left.\frac{\partial^2 M(s,t)}{\partial s\,\partial t}\right|_{(0,0)}.$$

**Example 8.10.** Let the random variables $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the joint moment generating function for $X$ and $Y$ ?

**Answer:** The joint moment generating function of $X$ and $Y$ is given by

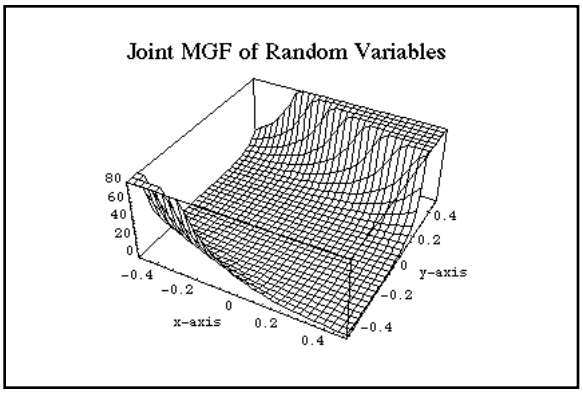
$$\begin{aligned} M(s,t) &= E\left(e^{sX+tY}\right) \\ &= \int_0^\infty \int_0^\infty e^{sx+ty} f(x,y)\, dy\, dx \\ &= \int_0^\infty \int_x^\infty e^{sx+ty}\, e^{-y}\, dy\, dx \\ &= \int_0^\infty \left[\int_x^\infty e^{sx+ty-y}\, dy\right] dx \\ &= \frac{1}{(1-s-t)(1-t)}, \qquad \text{provided} \quad s+t<1 \quad \text{and} \quad t<1. \end{aligned}$$

**Example 8.11.** If the joint moment generating function of the random variables $X$ and $Y$ is

$$M(s,t) = e^{(s+3t+2s^2+18t^2+12st)}$$

what is the covariance of $X$ and $Y$ ?

**Answer:**



$$\begin{aligned} M(s,t) &= e^{(s+3t+2s^2+18t^2+12st)} \\ \frac{\partial M}{\partial s} &= (1+4s+12t)\, M(s,t) \\ \left.\frac{\partial M}{\partial s}\right|_{(0,0)} &= 1\, M(0,0) \\ &= 1. \end{aligned}$$

$$\begin{aligned} \frac{\partial M}{\partial t} &= (3+36t+12s)\, M(s,t) \\ \left.\frac{\partial M}{\partial t}\right|_{(0,0)} &= 3\, M(0,0) \\ &= 3. \end{aligned}$$

Hence

$$\mu_X = 1 \qquad \text{and} \qquad \mu_Y = 3.$$

Now we compute the product moment of $X$ and $Y$.

$$\begin{aligned} \frac{\partial^2 M(s,t)}{\partial s\, \partial t} &= \frac{\partial}{\partial t}\left(\frac{\partial M}{\partial s}\right) \\ &= \frac{\partial}{\partial t}\left(M(s,t)\,(1+4s+12t)\right) \\ &= (1+4s+12t)\,\frac{\partial M}{\partial t} + M(s,t)\,(12). \end{aligned}$$

Therefore

$$\left.\frac{\partial^2 M(s,t)}{\partial s\, \partial t}\right|_{(0,0)} = 1\,(3) + 1\,(12).$$

Thus

$$E(XY) = 15$$

and the covariance of $X$ and $Y$ is given by

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)\,E(Y) \\ &= 15 - (3)\,(1) \\ &= 12. \end{aligned}$$

**Theorem 8.9.** If $X$ and $Y$ are independent then

$$M_{aX+bY}(t) = M_X(at)\, M_Y(bt),$$

where $a$ and $b$ real parameters.

**Proof:** Let $W = aX + bY$. Hence

$$\begin{aligned} M_{aX+bY}(t) &= M_W(t) \\ &= E\left(e^{tW}\right) \\ &= E\left(e^{t(aX+bY)}\right) \\ &= E\left(e^{taX}\, e^{tbY}\right) \\ &= E\left(e^{taX}\right) E\left(e^{tbY}\right) \qquad \text{(by Theorem 8.3)} \\ &= M_X(at)\, M_Y(bt). \end{aligned}$$

This theorem is very powerful. It helps us to find the distribution of a linear combination of independent random variables. The following examples illustrate how one can use this theorem to determine distribution of a linear combination.

**Example 8.12.** Suppose the random variable $X$ is normal with mean 2 and standard deviation 3 and the random variable $Y$ is also normal with mean 0 and standard deviation 4. If $X$ and $Y$ are independent, then what is the probability distribution of the random variable $X + Y$ ?

**Answer:** Since $X \sim N(2, 9)$, the moment generating function of $X$ is given by

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} = e^{2t + \frac{9}{2}t^2}.$$

Similarly, since $Y \sim N(0, 16)$,

$$M_Y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} = e^{\frac{16}{2}t^2}.$$

Since $X$ and $Y$ are independent, the moment generating function of $X + Y$ is given by

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\; M_Y(t) \\ &= e^{2t + \frac{9}{2}t^2}\; e^{\frac{16}{2}t^2} \\ &= e^{2t + \frac{25}{2}t^2}. \end{aligned}$$

Hence $X + Y \sim N(2, 25)$. Thus, $X + Y$ has a normal distribution with mean 2 and variance 25. From this information we can find the probability density function of $W = X + Y$ as

$$f(w) = \frac{1}{\sqrt{50\pi}}\, e^{-\frac{1}{2}\left(\frac{w-2}{5}\right)^2}, \qquad -\infty < w < \infty.$$

**Remark 8.6.** In fact if $X$ and $Y$ are independent normal random variables with means $\mu_X$ and $\mu_Y$ and variances $\sigma_X^2$ and $\sigma_Y^2$, respectively, then $aX+bY$ is also normal with mean $a\mu_X + b\mu_Y$ and variance $a^2\sigma_X^2 + b^2\sigma_Y^2$.

**Example 8.13.** Let $X$ and $Y$ be two independent and identically distributed random variables. If their common distribution is chi-square with one degree of freedom, then what is the distribution of $X + Y$? What is the moment generating function of $X - Y$ ?

**Answer:** Since $X$ and $Y$ are both $\chi^2(1)$, the moment generating functions are

$$M_X(t) = \frac{1}{\sqrt{1-2t}}$$

and

$$M_Y(t) = \frac{1}{\sqrt{1-2t}}.$$

Since, the random variables $X$ and $Y$ are independent, the moment generating function of $X + Y$ is given by

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\ M_Y(t) \\ &= \frac{1}{\sqrt{1-2t}}\ \frac{1}{\sqrt{1-2t}} \\ &= \frac{1}{(1-2t)^{\frac{2}{2}}}. \end{aligned}$$

Hence $X+Y \sim \chi^2(2)$. Thus, if $X$ and $Y$ are independent chi-square random variables, then their sum is also a chi-square random variable.

Next, we show that $X - Y$ is not a chi-square random variable, even if $X$ and $Y$ are both chi-square.

$$\begin{aligned} M_{X-Y}(t) &= M_X(t)\ M_Y(-t) \\ &= \frac{1}{\sqrt{1-2t}}\ \frac{1}{\sqrt{1+2t}} \\ &= \frac{1}{\sqrt{1-4t^2}}. \end{aligned}$$

This moment generating function does not correspond to the moment generating function of a chi-square random variable with any degree of freedoms. Further, it is surprising that this moment generating function does not correspond to that of any known distributions.

**Remark 8.7.** If $X$ and $Y$ are chi-square and independent random variables, then their linear combination is not necessarily a chi-square random variable.

**Example 8.14.** Let $X$ and $Y$ be two independent Bernoulli random variables with parameter $p$. What is the distribution of $X + Y$ ?

**Answer:** Since $X$ and $Y$ are Bernoulli with parameter $p$, their moment generating functions are

$$M_X(t) = (1-p) + pe^t \qquad M_Y(t) = (1-p) + pe^t.$$

Since, $X$ and $Y$ are independent, the moment generating function of their sum is the product of their moment generating functions, that is

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\ M_Y(t) \\ &= \left(1-p+pe^t\right)\ \left(1-p+pe^t\right) \\ &= \left(1-p+pe^t\right)^2. \end{aligned}$$

Hence $X + Y \sim BIN(2,\, p)$. Thus the sum of two independent Bernoulli random variable is a binomial random variable with parameter 2 and $p$.

### 8.6. Review Exercises

**1.** Suppose that $X_1$ and $X_2$ are random variables with zero mean and unit variance. If the correlation coefficient of $X_1$ and $X_2$ is $-0.5$, then what is the variance of $Y = \sum_{k=1}^{2} k^2 X_k$ ?

**2.** If the joint density of the random variables $X$ and $Y$ is

$$f(x,y) = \begin{cases} \frac{1}{8} & \text{if } (x,y) \in \{\,(x,0),\,(0,-y)\,|\,x,y = -2,-1,1,2\,\} \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the covariance of $X$ and $Y$ ? Are $X$ and $Y$ independent?

**3.** Suppose the random variables $X$ and $Y$ are independent and identically distributed. Let $Z = aX + Y$. If the correlation coefficient between $X$ and $Z$ is $\frac{1}{3}$, then what is the value of the constant $a$ ?

**4.** Let $X$ and $Y$ be two independent random variables with chi-square distribution with 2 degrees of freedom. What is the moment generating function of the random variable $2X + 3Y$ ? If possible, what is the distribution of $2X + 3Y$ ?

**5.** Let $X$ and $Y$ be two independent random variables. If $X \sim BIN(n,\, p)$ and $Y \sim BIN(m,\, p)$, then what is the distribution of $X + Y$ ?

**6.** Let $X$ and $Y$ be two independent random variables. If $X$ and $Y$ are both standard normal, then what is the distribution of the random variable $\frac{1}{2}\left(X^2 - Y^2\right)$ ?

**7.** If the joint probability density function of $X$ and $Y$ is

$$f(x,y) = \begin{cases} 1 & \text{if } 0 < x < 1;\ 0 < y < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

then what is the joint moment generating function of $X$ and $Y$ ?

**8.** Let the joint density function of $X$ and $Y$ be

$$f(x,y) = \begin{cases} \frac{1}{36} & \text{if } 1 \le x = y \le 6 \\ \\ \frac{2}{36} & \text{if } 1 \le x < y \le 6. \end{cases}$$

What is the correlation coefficient of $X$ and $Y$ ?

**9.** Suppose that $X$ and $Y$ are random variables with joint moment generating function

$$M(s,t) = \left(\frac{1}{4}\, e^s + \frac{3}{8}\, e^t + \frac{3}{8}\right)^{10},$$

for all real $s$ and $t$. What is the covariance of $X$ and $Y$ ?

**10.** Suppose that $X$ and $Y$ are random variables with joint density function

$$f(x,y) = \begin{cases} \frac{1}{6\pi} & \text{for } \frac{x^2}{4} + \frac{y^2}{9} \le 1 \\ \\ 0 & \text{for } \frac{x^2}{4} + \frac{y^2}{9} > 1. \end{cases}$$

What is the covariance of $X$ and $Y$ ? Are $X$ and $Y$ independent?

**11.** Let $X$ and $Y$ be two random variables. Suppose $E(X) = 1$, $E(Y) = 2$, $Var(X) = 1$, $Var(Y) = 2$, and $Cov(X,Y) = \frac{1}{2}$. For what values of the constants $a$ and $b$, the random variable $aX + bY$, whose expected value is 3, has minimum variance?

**12.** A box contains 5 white balls and 3 black balls. Draw 2 balls without replacement. If $X$ represents the number of white balls and Y represents the number of black balls drawn, what is the covariance of $X$ and $Y$ ?

**13.** If $X$ represents the number of 1's and $Y$ represents the number of 5's in three tosses of a fair six-sided die, what is the correlation between $X$ and $Y$?

**14.** Let $Y$ and $Z$ be two random variables. If $Var(Y) = 4$, $Var(Z) = 16$, and $Cov(Y, Z) = 2$, then what is $Var(3Z - 2Y)$?

**15.** Three random variables $X_1, X_2, X_3$, have equal variances $\sigma^2$ and coefficient of correlation between $X_1$ and $X_2$ of $\rho$ and between $X_1$ and $X_3$ and between $X_2$ and $X_3$ of zero. What is the correlation between $Y$ and $Z$ where $Y = X_1 + X_2$ and $Z = X_2 + X_3$ ?

**16.** If $X$ and $Y$ are two independent Bernoulli random variables with parameter $p$, then what is the joint moment generating function of $X - Y$?

**17.** If $X_1, X_2, ..., X_n$ are normal random variables with variance $\sigma^2$ and covariance between any pair of random variables $\rho\sigma^2$, what is the variance of $\frac{1}{n}\ (X_1 + X_2 + \cdots + X_n)$ ?

**18.** The coefficient of correlation between $X$ and $Y$ is $\frac{1}{3}$ and $\sigma_X^2 = a$, $\sigma_Y^2 = 4a$, and $\sigma_Z^2 = 114$ where $Z = 3X - 4Y$. What is the value of the constant $a$ ?

**19.** Let $X$ and $Y$ be independent random variables with $E(X) = 1$, $E(Y) = 2$, and $Var(X) = Var(Y) = \sigma^2$. For what value of the constant $k$ is the expected value of the random variable $k(X^2 - Y^2) + Y^2$ equals $\sigma^2$?

**20.** Let $X$ be a random variable with finite variance. If $Y = 15 - X$, then what is the coefficient of correlation between the random variables $X$ and $(X + Y)X$ ?

**21.** The mean of a normal random variable $X$ is 10 and the variance is 12. The mean of a normal random variable $Y$ is $-5$ and the variance is 5. If the covariance of $X$ and $Y$ is 4, then what is the probability of the event $X + Y > 5$ ?

# Chapter 9

# CONDITIONAL EXPECTATION OF BIVARIATE RANDOM VARIABLES

This chapter examines the conditional mean and conditional variance associated with two random variables. The conditional mean is very useful in Bayesian estimation of parameters with a square loss function. Further, the notion of conditional mean sets the path for regression analysis in statistics.

## 9.1. Conditional Expected Values

Let $X$ and $Y$ be any two random variables with joint density $f(x, y)$. Recall that the conditional probability density of $X$, given the event $Y = y$, is defined as

$$g(x/y) = \frac{f(x, y)}{f_2(y)}, \qquad f_2(y) > 0$$

where $f_2(y)$ is the marginal probability density of $Y$. Similarly, the conditional probability density of $Y$, given the event $X = x$, is defined as

$$h(y/x) = \frac{f(x, y)}{f_1(x)}, \qquad f_1(x) > 0$$

where $f_1(x)$ is the marginal probability density of $X$.

**Definition 9.1.** The conditional mean of $X$ given $Y = y$ is defined as

$$\mu_{X|y} = E\left(X \,|\, y\right),$$

where

$$E\left(X \mid y\right) = \begin{cases} \displaystyle\sum_{x \in R_X} x\, g(x/y) & \text{if } X \text{ is discrete} \\ \\ \displaystyle\int_{-\infty}^{\infty} x\, g(x/y)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Similarly, the conditional mean of $Y$ given $X = x$ is defined as

$$\mu_{Y|x} = E\left(Y \mid x\right),$$

where

$$E\left(Y \mid x\right) = \begin{cases} \displaystyle\sum_{y \in R_Y} y\, h(y/x) & \text{if } Y \text{ is discrete} \\ \\ \displaystyle\int_{-\infty}^{\infty} y\, h(y/x)\, dy & \text{if } Y \text{ is continuous.} \end{cases}$$

**Example 9.1.** Let $X$ and $Y$ be discrete random variables with joint probability density function

$$f(x,y) = \begin{cases} \frac{1}{21}\left(x+y\right) & \text{for } x = 1,2,3;\ y = 1,2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional mean of $X$ given $Y = y$, that is $E(X|y)$?

**Answer:** To compute the conditional mean of $X$ given $Y = y$, we need the conditional density $g(x/y)$ of $X$ given $Y = y$. However, to find $g(x/y)$, we need to know the marginal of $Y$, that is $f_2(y)$. Thus, we begin with

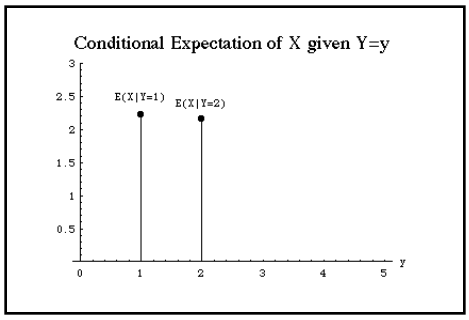
$$\begin{aligned} f_2(y) &= \sum_{x=1}^{3} \frac{1}{21}\left(x+y\right) \\ &= \frac{1}{21}\left(6+3y\right). \end{aligned}$$

Therefore, the conditional density of $X$ given $Y = y$ is given by

$$\begin{aligned} g(x/y) &= \frac{f(x,y)}{f_2(y)} \\ &= \frac{x+y}{6+3y}, \qquad x = 1,2,3. \end{aligned}$$

The conditional expected value of $X$ given the event $Y = y$

$$\begin{aligned} E(X \mid y) &= \sum_{x \in R_X} x\, g(x/y) \\ &= \sum_{x=1}^{3} x\, \frac{x+y}{6+3y} \\ &= \frac{1}{6+3y} \left[ \sum_{x=1}^{3} x^2 + y \sum_{x=1}^{3} x \right] \\ &= \frac{14+6y}{6+3y}, \qquad y = 1, 2. \end{aligned}$$




**Remark 9.1.** Note that the conditional mean of $X$ given $Y = y$ is dependent only on $y$, that is $E(X|y)$ is a function $\phi$ of $y$. In the above example, this function $\phi$ is a rational function, namely $\phi(y) = \frac{14+6y}{6+3y}$.

**Example 9.2.** Let $X$ and $Y$ have the joint density function

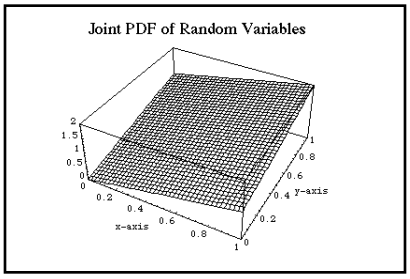
$$f(x,y) = \begin{cases} x+y & \text{for } 0 < x,\, y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

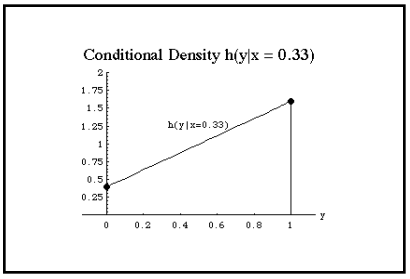
What is the conditional mean $E\left(Y \mid X = \frac{1}{3}\right)$ ?

**Answer:**

$$\begin{aligned} f_1(x) &= \int_0^1 (x+y)\, dy \\ \\ &= \left[ xy + \frac{1}{2}\, y^2 \right]_0^1 \\ \\ &= x + \frac{1}{2}. \end{aligned}$$

$$h(y/x) = \frac{f(x,y)}{f_1(x)} = \frac{x+y}{x+\frac{1}{2}}.$$




$$\begin{aligned}
E\left(Y \mid X = \frac{1}{3}\right) &= \int_0^1 y\, h(y/x)\, dy \\
&= \int_0^1 y\, \frac{x+y}{x+\frac{1}{2}}\, dy \\
&= \int_0^1 y\, \frac{\frac{1}{3}+y}{\frac{5}{6}}\, dy \\
&= \frac{6}{5} \int_0^1 \left(\frac{1}{3}\, y + y^2\right) dy \\
&= \frac{6}{5} \left[\frac{1}{6}\, y^2 + \frac{1}{3}\, y^3\right]_0^1 \\
&= \frac{6}{5} \left[\frac{1}{6} + \frac{2}{6}\right] \\
&= \frac{6}{5} \left(\frac{3}{6}\right) \\
&= \frac{3}{5}.
\end{aligned}$$

The mean of the random variable $Y$ is a deterministic number. The conditional mean of $Y$ given $X = x$, that is $E(Y|x)$, is a function $\phi(x)$ of the variable $x$. Using this function, we form $\phi(X)$. This function $\phi(X)$ is a random variable. Thus starting from the deterministic function $E(Y|x)$, we have formed the random variable $E(Y|X) = \phi(X)$. An important property of conditional expectation is given by the following theorem.

**Theorem 9.1.** The expected value of the random variable $E(Y|X)$ is equal to the expected value of $Y$, that is

$$E_x\left(E_{y|x}\left(Y|X\right)\right) = E_y(Y),$$

where $E_x(X)$ stands for the expectation of $X$ with respect to the distribution of $X$ and $E_{y|x}(Y|X)$ stands for the expected value of $Y$ with respect to the conditional density $h(y/X)$.

**Proof:** We prove this theorem for continuous variables and leave the discrete case to the reader.

$$\begin{aligned}
E_x\left(E_{y|x}(Y|X)\right) &= E_x\left[\int_{-\infty}^{\infty} y\,h(y/X)\,dy\right] \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y\,h(y/x)\,dy\right) f_1(x)\,dx \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y\,h(y/x) f_1(x)\,dy\right) dx \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} h(y/x) f_1(x)\,dx\right) y\,dy \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(x,y)\,dx\right) y\,dy \\
&= \int_{-\infty}^{\infty} y\,f_2(y)\,dy \\
&= E_y(Y).
\end{aligned}$$

**Example 9.3.** An insect lays $Y$ number of eggs, where $Y$ has a Poisson distribution with parameter $\lambda$. If the probability of each egg surviving is $p$, then on the average how many eggs will survive?

**Answer:** Let $X$ denote the number of surviving eggs. Then, given that $Y = y$ (that is given that the insect has laid $y$ eggs) the random variable $X$ has a binomial distribution with parameters $y$ and $p$. Thus

$$\begin{aligned}
X|Y &\sim BIN(Y,p) \\
Y &\sim POI(\lambda).
\end{aligned}$$

Therefore, the expected number of survivors is given by

$$\begin{aligned}
E_x(X) &= E_y\left(E_{x|y}(X|Y)\right) \\
&= E_y(pY) \qquad \text{(since X|Y} \sim \text{ BIN(Y, p))} \\
&= p\,E_y(Y) \\
&= p\,\lambda. \qquad \text{(since Y} \sim \text{ POI}(\lambda))
\end{aligned}$$

**Definition 9.2.** A random variable $X$ is said to have a mixture distribution if the distribution of $X$ depends on a quantity which also has a distribution.

**Example 9.4.** A fair coin is tossed. If a head occurs, 1 die is rolled; if a tail occurs, 2 dice are rolled. Let $Y$ be the total on the die or dice. What is the expected value of $Y$?

**Answer:** Let $X$ denote the outcome of tossing a coin. Then $X \sim BER(p)$, where the probability of success is $p = \frac{1}{2}$.

$$\begin{aligned}
E_y(Y) &= E_x(\, E_{y|x}(Y|X)\,) \\
&= \frac{1}{2}\, E_{y|x}(Y|X=0) + \frac{1}{2}\, E_{y|x}(Y|X=1) \\
&= \frac{1}{2}\left(\frac{1+2+3+4+5+6}{6}\right) \\
&\quad + \frac{1}{2}\left(\frac{2+6+12+20+30+42+40+36+30+22+12}{36}\right) \\
&= \frac{1}{2}\left(\frac{126}{36} + \frac{252}{36}\right) \\
&= \frac{378}{72} \\
&= 5.25.
\end{aligned}$$

Note that the expected number of dots that show when 1 die is rolled is $\frac{126}{36}$, and the expected number of dots that show when 2 dice are rolled is $\frac{252}{36}$.

**Theorem 9.2.** Let $X$ and $Y$ be two random variables with mean $\mu_X$ and $\mu_Y$, and standard deviation $\sigma_X$ and $\sigma_Y$, respectively. If the conditional expectation of $Y$ given $X = x$ is linear in $x$, then

$$E(Y|X=x) = \mu_Y + \rho\,\frac{\sigma_Y}{\sigma_X}\,(x - \mu_X),$$

where $\rho$ denotes the correlation coefficient of $X$ and $Y$.

**Proof:** We assume that the random variables $X$ and $Y$ are continuous. If they are discrete, the proof of the theorem follows exactly the same way by replacing the integrals with summations. We are given that $E(Y|X=x)$ is linear in $x$, that is

$$E(Y|X=x) = a\,x + b, \tag{9.0}$$

where $a$ and $b$ are two constants. Hence, from above we get

$$\int_{-\infty}^{\infty} y\, h(y/x)\, dy = a\,x + b$$

which implies

$$\int_{-\infty}^{\infty} y \, \frac{f(x,y)}{f_1(x)} \, dy = a\,x + b.$$

Multiplying both sides by $f_1(x)$, we get

$$\int_{-\infty}^{\infty} y \, f(x,y) \, dy = (a\,x + b) \, f_1(x) \tag{9.1}$$

Now integrating with respect to $x$, we get

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \, f(x,y) \, dy \, dx = \int_{-\infty}^{\infty} (a\,x + b) \, f_1(x) \, dx$$

This yields

$$\mu_Y = a\,\mu_X + b. \tag{9.2}$$

Now, we multiply (9.1) with $x$ and then integrate the resulting expression with respect to $x$ to get

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \, f(x,y) \, dy \, dx = \int_{-\infty}^{\infty} (a\,x^2 + bx) \, f_1(x) \, dx.$$

From this we get

$$E(XY) = a\,E\left(X^2\right) + b\,\mu_X. \tag{9.3}$$

Solving (9.2) and (9.3) for the unknown $a$ and $b$, we get

$$\begin{aligned} a &= \frac{E(XY) - \mu_X\,\mu_Y}{\sigma_X^2} \\ &= \frac{\sigma_{XY}}{\sigma_X^2} \\ &= \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y} \; \frac{\sigma_Y}{\sigma_X} \\ &= \rho \; \frac{\sigma_Y}{\sigma_X}. \end{aligned}$$

Similarly, we get

$$b = \mu_Y + \rho \; \frac{\sigma_Y}{\sigma_X} \, \mu_X.$$

Letting $a$ and $b$ into (9.0) we obtain the asserted result and the proof of the theorem is now complete.

**Example 9.5.** Suppose $X$ and $Y$ are random variables with $E(Y|X = x) = -x + 3$ and $E(X|Y = y) = -\frac{1}{4}y + 5$. What is the correlation coefficient of $X$ and $Y$ ?

**Answer:** From the Theorem 9.2, we get

$$\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) = -x + 3.$$

Therefore, equating the coefficients of $x$ terms, we get

$$\rho \frac{\sigma_Y}{\sigma_X} = -1. \tag{9.4}$$

Similarly, since

$$\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y) = -\frac{1}{4}y + 5$$

we have

$$\rho \frac{\sigma_X}{\sigma_Y} = -\frac{1}{4}. \tag{9.5}$$

Multiplying (9.4) with (9.5), we get

$$\rho \frac{\sigma_Y}{\sigma_X}\ \rho \frac{\sigma_X}{\sigma_Y} = (-1)\left(-\frac{1}{4}\right)$$

which is

$$\rho^2 = \frac{1}{4}.$$

Solving this, we get

$$\rho = \pm\frac{1}{2}.$$

Since $\rho \frac{\sigma_Y}{\sigma_X} = -1$ and $\frac{\sigma_Y}{\sigma_X} > 0$, we get

$$\rho = -\frac{1}{2}.$$

### 9.2. Conditional Variance

The variance of the probability density function $f(y/x)$ is called the conditional variance of $Y$ given that $X = x$. This conditional variance is defined as follows:

**Definition 9.3.** Let $X$ and $Y$ be two random variables with joint density $f(x, y)$ and $f(y/x)$ be the conditional density of $Y$ given $X = x$. The conditional variance of $Y$ given $X = x$, denoted by $Var(Y|x)$, is defined as

$$Var(Y|x) = E\left(Y^2 \,|\, x\right) - \left(E(Y|x)\right)^2,$$

where $E(Y|x)$ denotes the conditional mean of $Y$ given $X = x$.

**Example 9.6.** Let $X$ and $Y$ be continuous random variables with joint probability density function

$$f(x,y) = \begin{cases} e^{-y} & \text{for } 0 < x < y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional variance of $Y$ given the knowledge that $X = x$?

**Answer:** The marginal density of $f_1(x)$ is given by

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x,y)\, dy \\ &= \int_{x}^{\infty} e^{-y}\, dy \\ &= \left[ -e^{-y} \right]_{x}^{\infty} \\ &= e^{-x}. \end{aligned}$$

Thus, the conditional density of $Y$ given $X = x$ is

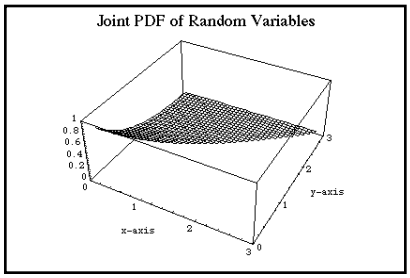
$$\begin{aligned} h(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{e^{-y}}{e^{-x}} \\ &= e^{-(y-x)} \qquad \text{for} \quad y > x. \end{aligned}$$

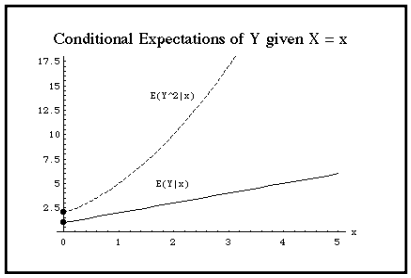
Thus, given $X = x$, $Y$ has an exponential distribution with parameter $\theta = 1$ and location parameter $x$. The conditional mean of $Y$ given $X = x$ is

$$\begin{aligned} E(Y|x) &= \int_{-\infty}^{\infty} y\, h(y/x)\, dy \\ &= \int_{x}^{\infty} y\, e^{-(y-x)}\, dy \\ &= \int_{0}^{\infty} (z+x)\, e^{-z}\, dz \qquad \text{where } z = y - x \\ \\ &= x \int_{0}^{\infty} e^{-z}\, dz + \int_{0}^{\infty} z\, e^{-z}\, dz \\ &= x\, \Gamma(1) + \Gamma(2) \\ &= x + 1. \end{aligned}$$

Similarly, we compute the second moment of the distribution $h(y/x)$.

$$\begin{aligned}
E(Y^2|x) &= \int_{-\infty}^{\infty} y^2\, h(y/x)\, dy \\
&= \int_{x}^{\infty} y^2\, e^{-(y-x)}\, dy \\
&= \int_{0}^{\infty} (z+x)^2\, e^{-z}\, dz \qquad \text{where } z = y - x \\
\\
&= x^2 \int_0^{\infty} e^{-z}\, dz + \int_0^{\infty} z^2\, e^{-z}\, dz + 2\,x \int_0^{\infty} z\, e^{-z}\, dz \\
&= x^2\, \Gamma(1) + \Gamma(3) + 2\,x\,\Gamma(2) \\
&= x^2 + 2 + 2x \\
&= (1+x)^2 + 1.
\end{aligned}$$




Therefore

$$\begin{aligned}
Var(Y|x) &= E\left(Y^2|x\right) - \left[\, E(Y|x)\,\right]^2 \\
&= (1+x)^2 + 1 - (1+x)^2 \\
&= 1.
\end{aligned}$$

**Remark 9.2.** The variance of $Y$ is 2. This can be seen as follows: Since, the marginal of $Y$ is given by $f_2(y) = \int_0^y e^{-y}\, dx = y\, e^{-y}$, the expected value of $Y$ is $E(Y) = \int_0^{\infty} y^2 e^{-y}\, dy = \Gamma(3) = 2$, and $E\left(Y^2\right) = \int_0^{\infty} y^3 e^{-y}\, dy = \Gamma(4) = 6$. Thus, the variance of $Y$ is $Var(Y) = 6 - 4 = 2$. However, given the knowledge $X = x$, the variance of $Y$ is 1. Thus, in a way the prior knowledge reduces the variability (or the variance) of a random variable.

Next, we simply state the following theorem concerning the conditional variance without proof.

**Theorem 9.3.** Let $X$ and $Y$ be two random variables with mean $\mu_X$ and $\mu_Y$, and standard deviation $\sigma_X$ and $\sigma_Y$, respectively. If the conditional expectation of $Y$ given $X = x$ is linear in $x$, then

$$E_x\left(Var(Y|X)\right) = (1-\rho^2)\, Var(Y),$$

where $\rho$ denotes the correlation coefficient of $X$ and $Y$.

**Example 9.7.** Let $E(Y|X=x) = 2x$ and $Var(Y|X=x) = 4x^2$, and let $X$ have a uniform distribution on the interval from 0 to 1. What is the variance of $Y$?

**Answer:** If $E(Y|X=x)$ is linear function of $x$, then

$$E(Y|X=x) = \mu_Y + \rho\,\frac{\sigma_Y}{\sigma_X}\,(x-\mu_X)$$

and

$$E_x\left(\,Var(Y|X)\,\right) = \sigma_Y^2\,(1-\rho^2).$$

We are given that

$$\mu_Y + \rho\,\frac{\sigma_Y}{\sigma_X}\,(x-\mu_X) = 2x.$$

Hence, equating the coefficient of $x$ terms, we get

$$\rho\,\frac{\sigma_Y}{\sigma_X} = 2$$

which is

$$\rho = 2\,\frac{\sigma_X}{\sigma_Y}. \tag{9.6}$$

Further, we are given that

$$Var(Y|X=x) = 4x^2$$

Since $X \sim UNIF(0,1)$, we get the density of $X$ to be $f(x) = 1$ on the interval $(0,1)$ Therefore,

$$\begin{aligned} E_x\left(\,Var(Y|X)\,\right) &= \int_{-\infty}^{\infty} Var(Y|X=x)\, f(x)\, dx \\ &= \int_0^1 4\,x^2\, dx \\ &= 4\left[\,\frac{x^3}{3}\,\right]_0^1 \\ &= \frac{4}{3}. \end{aligned}$$

By Theorem 9.3,

$$\begin{aligned}\frac{4}{3} &= E_x\left(Var(Y|X)\right)\\ &= \sigma_Y^2\left(1-\rho^2\right)\\ &= \sigma_Y^2\left(1-4\,\frac{\sigma_X^2}{\sigma_Y^2}\right)\\ &= \sigma_Y^2-4\,\sigma_X^2\end{aligned}$$

Hence

$$\sigma_Y^2=\frac{4}{3}+4\,\sigma_X^2.$$

Since $X\sim UNIF(0,1)$, the variance of $X$ is given by $\sigma_X^2=\frac{1}{12}$. Therefore, the variance of $Y$ is given by

$$\sigma_Y^2=\frac{4}{3}+\frac{4}{12}=\frac{16}{12}+\frac{4}{12}=\frac{20}{12}=\frac{5}{3}.$$

**Example 9.8.** Let $E(X|Y=y)=3y$ and $Var(X|Y=y)=2$, and let $Y$ have density function

$$f(y)=\begin{cases}e^{-y} & \text{if } y>0\\ \\ 0 & \text{otherwise.}\end{cases}$$

What is the variance of $X$?

**Answer:** By Theorem 9.3, we get

$$Var(X|Y=y)=\sigma_X^2\left(1-\rho^2\right)=2 \tag{9.7}$$

and

$$\mu_X+\rho\,\frac{\sigma_X}{\sigma_Y}\,(y-\mu_Y)=3y.$$

Thus

$$\rho=3\,\frac{\sigma_Y}{\sigma_X}.$$

Hence from (9.7), we get $E_y\left(Var(X|Y)\right)=2$ and thus

$$\sigma_X^2\left(1-9\,\frac{\sigma_Y^2}{\sigma_X^2}\right)=2$$

which is

$$\sigma_X^2=9\,\sigma_Y^2+2.$$

Now, we compute the variance of $Y$. For this, we need $E(Y)$ and $E\left(Y^2\right)$.

$$\begin{aligned} E(Y) &= \int_0^\infty y\, f(y)\, dy \\ &= \int_0^\infty y\, e^{-y}\, dy \\ &= \Gamma(2) \\ &= 1. \end{aligned}$$

Similarly

$$\begin{aligned} E\left(Y^2\right) &= \int_0^\infty y^2\, f(y)\, dy \\ &= \int_0^\infty y^2\, e^{-y}\, dy \\ &= \Gamma(3) \\ &= 2. \end{aligned}$$

Therefore

$$Var(Y) = E\left(Y^2\right) - [\, E(Y)\,]^2 = 2 - 1 = 1.$$

Hence, the variance of $X$ can be calculated as

$$\begin{aligned} \sigma_X^2 &= 9\, \sigma_Y^2 + 2 \\ &= 9\,(1) + 2 \\ &= 11. \end{aligned}$$

**Remark 9.3.** Notice that, in Example 9.8, we calculated the variance of $Y$ directly using the form of $f(y)$. It is easy to note that $f(y)$ has the form of an exponential density with parameter $\theta = 1$, and therefore its variance is the square of the parameter. This straightforward gives $\sigma_Y^2 = 1$.

### 9.3. Regression Curve and Scedastic Curve

One of the major goals in most statistical studies is to establish relationships between two or more random variables. For example, a company would like to know the relationship between the potential sales of a new product in terms of its price. Historically, regression analysis was originated in the works of Sir Francis Galton (1822-1911) but most of the theory of regression analysis was developed by his student Sir Ronald Fisher (1890-1962).

**Definition 9.4.** Let $X$ and $Y$ be two random variables with joint probability density function $f(x,y)$ and let $h(y/x)$ is the conditional density of $Y$ given $X = x$. Then the conditional mean

$$E(Y|X = x) = \int_{-\infty}^{\infty} y\, h(y/x)\, dy$$

is called the regression function of $Y$ on $X$. The graph of this regression function of $Y$ on $X$ is known as the regression curve of $Y$ on $X$.

**Example 9.9.** Let $X$ and $Y$ be two random variables with joint density

$$f(x,y) = \begin{cases} x\, e^{-x(1+y)} & \text{if } x > 0;\ y > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the regression function of $Y$ on $X$?

**Answer:** The marginal density $f_1(x)$ of $X$ is

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x,y)\, dy \\ &= \int_0^{\infty} x\, e^{-x(1+y)}\, dy \\ &= \int_0^{\infty} x\, e^{-x}\, e^{-xy}\, dy \\ &= x\, e^{-x} \int_0^{\infty} e^{-xy}\, dy \\ &= x\, e^{-x} \left[ -\frac{1}{x}\, e^{-xy} \right]_0^{\infty} \\ &= e^{-x}. \end{aligned}$$

The conditional density of $Y$ given $X = x$ is

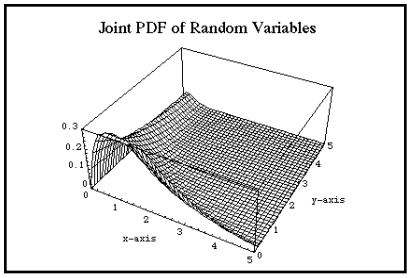
$$\begin{aligned} h(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{x\, e^{-x(1+y)}}{e^{-x}} \\ &= x\, e^{-xy}. \end{aligned}$$

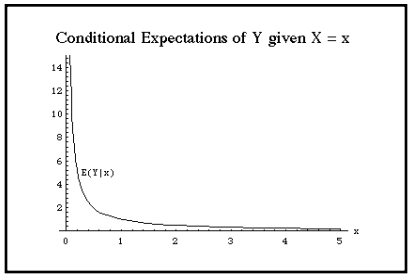
The conditional mean of $Y$ given that $X = x$ is

$$\begin{aligned} E(Y|X=x) &= \int_{-\infty}^{\infty} y\, h(y/x)\, dy \\ &= \int_{0}^{\infty} y\, x\, e^{-xy}\, dy \\ &= \frac{1}{x}\int_{0}^{\infty} ze^{-z}\, dz \qquad (\text{where } z = xy) \\ &= \frac{1}{x}\,\Gamma(2) \\ &= \frac{1}{x}. \end{aligned}$$

Thus, the regression function (or equation) of $Y$ on $X$ is given by

$$E(Y|x) = \frac{1}{x} \qquad \text{for} \quad 0 < x < \infty.$$




**Definition 9.4.** Let $X$ and $Y$ be two random variables with joint probability density function $f(x, y)$ and let $E(Y|X = x)$ be the regression function of $Y$ on $X$. If this regression function is linear, then $E(Y|X = x)$ is called a linear regression of $Y$ on $X$. Otherwise, it is called nonlinear regression of $Y$ on $X$.

**Example 9.10.** Given the regression lines $E(Y|X = x) = x + 2$ and $E(X|Y = y) = 1 + \frac{1}{2}\, y$, what is the expected value of $X$?

**Answer:** Since the conditional expectation $E(Y|X = x)$ is linear in $x$, we get

$$\mu_Y + \rho\, \frac{\sigma_Y}{\sigma_X}\, (x - \mu_X) = x + 2.$$

Hence, equating the coefficients of $x$ and constant terms, we get

$$\rho\, \frac{\sigma_Y}{\sigma_X} = 1 \tag{9.8}$$

and

$$\mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X = 2, \tag{9.9}$$

respectively. Now, using (9.8) in (9.9), we get

$$\mu_Y - \mu_X = 2. \tag{9.10}$$

Similarly, since $E(X|Y = y)$ is linear in $y$, we get

$$\rho \frac{\sigma_X}{\sigma_Y} = \frac{1}{2} \tag{9.11}$$

and

$$\mu_X - \rho \frac{\sigma_X}{\sigma_Y} \mu_Y = 1, \tag{9.12}$$

Hence, letting (9.10) into (9.11) and simplifying, we get

$$2\mu_X - \mu_Y = 2. \tag{9.13}$$

Now adding (9.13) to (9.10), we see that

$$\mu_X = 4.$$

**Remark 9.4.** In statistics, a linear regression usually means the conditional expectation $E\,(Y/x)$ is linear in the parameters, but not in $x$. Therefore, $E\,(Y/x) = \alpha + \theta x^2$ will be a linear model, where as $E\,(Y/x) = \alpha\, x^\theta$ is not a linear regression model.

**Definition 9.5.** Let $X$ and $Y$ be two random variables with joint probability density function $f(x, y)$ and let $h(y/x)$ is the conditional density of $Y$ given $X = x$. Then the conditional variance

$$Var(Y|X = x) = \int_{-\infty}^{\infty} y^2\, h(y/x)\, dy$$

is called the scedastic function of $Y$ on $X$. The graph of this scedastic function of $Y$ on $X$ is known as the scedastic curve of $Y$ on $X$.

Scedastic curves and regression curves are used for constructing families of bivariate probability density functions with specified marginals.

### 9.4. Review Exercises

**1.** Given the regression lines $E(Y|X=x) = x+2$ and $E(X|Y=y) = 1+\frac{1}{2}\,y$, what is expected value of $Y$ ?

**2.** If the joint density of $X$ and $Y$ is

$$f(x,y) = \begin{cases} k & \text{if} \quad -1 < x < 1;\ x^2 < y < 1 \\ \\ 0 & \text{elsewhere ,} \end{cases}$$

where $k$ is a constant, what is $E(Y|X=x)$ ?

**3.** Suppose the joint density of $X$ and $Y$ is defined by

$$f(x,y) = \begin{cases} 10xy^2 & \text{if} \quad 0 < x < y < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is $E\left(X^2|Y=y\right)$ ?

**4.** Let $X$ and $Y$ joint density function

$$f(x,y) = \begin{cases} 2e^{-2(x+y)} & \text{if} \quad 0 < x < y < \infty \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the expected value of $Y$, given $X = x$, for $x > 0$ ?

**5.** Let $X$ and $Y$ joint density function

$$f(x,y) = \begin{cases} 8xy & \text{if} \quad 0 < x < 1;\ 0 < y < x \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the regression curve $y$ on $x$, that is, $E\left(Y/X=x\right)$?

**6.** Suppose $X$ and $Y$ are random variables with means $\mu_X$ and $\mu_Y$, respectively; and $E(Y|X=x) = -\frac{1}{3}x + 10$ and $E(X|Y=y) = -\frac{3}{4}y + 2$. What are the values of $\mu_X$ and $\mu_Y$?

**7.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} \frac{24}{5}\,(x+y) & \text{for } 0 \le 2y \le x \le 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional expectation of $X$ given $Y = y$ ?

**8.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} c\,xy^2 & \text{for } 0 \le y \le 2x;\ 1 \le x \le 5 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional expectation of $Y$ given $X = x$ ?

**9.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} e^{-y} & \text{for } y \ge x \ge 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional expectation of $X$ given $Y = y$ ?

**10.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} 2\,xy & \text{for } 0 \le y \le 2x \le 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional expectation of $Y$ given $X = x$ ?

**11.** Let $E(Y|X = x) = 2 + 5x$, $Var(Y|X = x) = 3$, and let $X$ have the density function

$$f(x) = \begin{cases} \frac{1}{4}\,x\,e^{-\frac{x}{2}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the mean and variance of random variable $Y$?

**12.** Let $E(Y|X = x) = 2x$ and $Var(Y|X = x) = 4x^2 + 3$, and let $X$ have the density function

$$f(x) = \begin{cases} \frac{4}{\sqrt{\pi}}\,x^2\,e^{-x^2} & \text{for } 0 \le x < \infty \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the variance of $Y$?

**13.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} 2 & \text{for } 0 < y < 1 - x;\ \text{and } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional variance of $Y$ given $X = x$ ?

**14.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} 4x & \text{for } 0 < x < \sqrt{y} < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the conditional variance of $Y$ given $X = x$ ?

**15.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} \frac{6}{7}x & \text{for } 1 \leq x+y \leq 2;\ x \geq 0, y \geq 0 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the conditional variance of $X$ given $Y = \frac{3}{2}$ ?

**16.** Let $X$ and $Y$ have joint density

$$f(x,y) = \begin{cases} 12x & \text{for } 0 < y < 2x < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the conditional variance of $Y$ given $X = 0.5$ ?

**17.** Let the random variable $W$ denote the number of students who take business calculus each semester at the University of Louisville. If the random variable $W$ has a Poisson distribution with parameter $\lambda$ equal to 300 and the probability of each student passing the course is $\frac{3}{5}$, then on an average how many students will pass the business calculus?

**18.** If the conditional density of $Y$ given $X = x$ is given by

$$f(y/x) = \begin{cases} \binom{5}{y} x^y (1-x)^{5-y} & \text{if } y = 0, 1, 2, ..., 5 \\ \\ 0 & \text{otherwise,} \end{cases}$$

and the marginal density of $X$ is

$$f_1(x) = \begin{cases} 4x^3 & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the conditional expectation of $Y$ given the event $X = x$?

**19.** If the joint density of the random variables $X$ and $Y$ is

$$f(x,y) = \begin{cases} \frac{2+(2x-1)(2y-1)}{2} & \text{if } 0 < x, y < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the regression function of $Y$ on $X$?

**20.** If the joint density of the random variables $X$ and $Y$ is

$$f(x,y) = \begin{cases} \left[e^{\min\{x,y\}} - 1\right] e^{-(x+y)} & \text{if } 0 < x, y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

then what is the conditional expectation of $Y$ given $X = x$?

# Chapter 10

# TRANSFORMATION OF RANDOM VARIABLES AND THEIR DISTRIBUTIONS

In many statistical applications, given the probability distribution of a univariate random variable $X$, one would like to know the probability distribution of another univariate random variable $Y = \phi(X)$, where $\phi$ is some known function. For example, if we know the probability distribution of the random variable $X$, we would like know the distribution of $Y = \ln(X)$. For univariate random variable $X$, some commonly used transformed random variable $Y$ of $X$ are: $Y = X^2$, $Y = |X|$, $Y = \sqrt{|X|}$, $Y = \ln(X)$, $Y = \frac{X-\mu}{\sigma}$, and $Y = \left(\frac{X-\mu}{\sigma}\right)^2$. Similarly for a bivariate random variable $(X, Y)$, some of the most common transformations of $X$ and $Y$ are $X + Y$, $XY$, $\frac{X}{Y}$, $\min\{X, Y\}$, $\max\{X, Y\}$ or $\sqrt{X^2 + Y^2}$. In this chapter, we examine various methods for finding the distribution of a transformed univariate or bivariate random variable, when transformation and distribution of the variable are known. First, we treat the univariate case. Then we treat the bivariate case.

We begin with an example for univariate discrete random variable.

**Example 10.1.** The probability density function of the random variable $X$ is shown in the table below.

| $x$ | $-2$ | $-1$ | $0$ | $1$ | $2$ | $3$ | $4$ |
|---|---|---|---|---|---|---|---|
| $f(x)$ | $\frac{1}{10}$ | $\frac{2}{10}$ | $\frac{1}{10}$ | $\frac{1}{10}$ | $\frac{1}{10}$ | $\frac{2}{10}$ | $\frac{2}{10}$ |

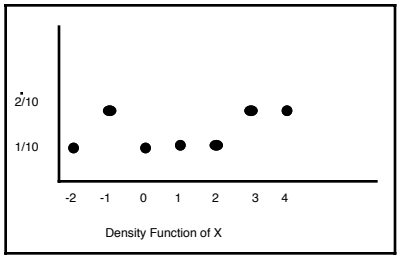
What is the probability density function of the random variable $Y = X^2$?

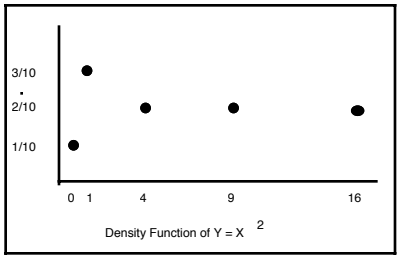
**Answer:** The space of the random variable $X$ is $R_X = \{-2, -1, 0, 1, 2, 3, 4\}$. Then the space of the random variable $Y$ is $R_Y = \{x^2 \mid x \in R_X\}$. Thus, $R_Y = \{0, 1, 4, 9, 16\}$. Now we compute the probability density function $g(y)$ for $y$ in $R_Y$.

$$
\begin{aligned}
g(0) &= P(Y = 0) = P(X^2 = 0) = P(X = 0)) = \frac{1}{10} \\
g(1) &= P(Y = 1) = P(X^2 = 1) = P(X = -1) + P(X = 1) = \frac{3}{10} \\
g(4) &= P(Y = 4) = P(X^2 = 4) = P(X = -2) + P(X = 2) = \frac{2}{10} \\
g(9) &= P(Y = 9) = P(X^2 = 9) = P(X = 3) = \frac{2}{10} \\
g(16) &= P(Y = 16) = P(X^2 = 16) = P(X = 4) = \frac{2}{10}.
\end{aligned}
$$

We summarize the distribution of $Y$ in the following table.

| $y$ | 0 | 1 | 4 | 9 | 16 |
|---|---|---|---|---|---|
| $g(y)$ | $\frac{1}{10}$ | $\frac{3}{10}$ | $\frac{2}{10}$ | $\frac{2}{10}$ | $\frac{2}{10}$ |




**Example 10.2.** The probability density function of the random variable $X$ is shown in the table below.

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $f(x)$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

What is the probability density function of the random variable $Y = 2X + 1$?

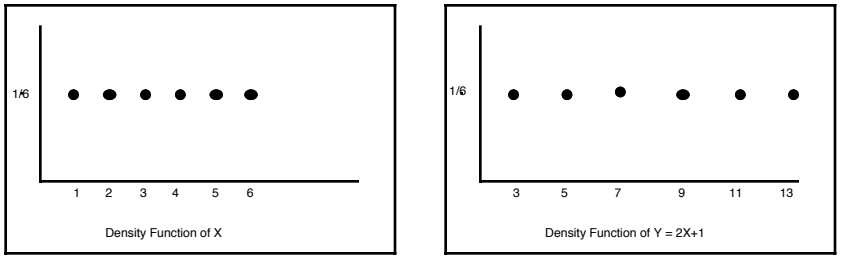
**Answer:** The space of the random variable $X$ is $R_X = \{1, 2, 3, 4, 5, 6\}$. Then the space of the random variable $Y$ is $R_Y = \{2x + 1 \mid x \in R_X\}$. Thus, $R_Y = \{3, 5, 7, 9, 11, 13\}$. Next we compute the probability density function $g(y)$ for $y$ in $R_Y$. The pdf $g(y)$ is given by

$$g(3) = P(Y = 3) = P(2X + 1 = 3) = P(X = 1)) = \frac{1}{6}$$
$$g(5) = P(Y = 5) = P(2X + 1 = 5) = P(X = 2)) = \frac{1}{6}$$
$$g(7) = P(Y = 7) = P(2X + 1 = 7) = P(X = 3)) = \frac{1}{6}$$
$$g(9) = P(Y = 9) = P(2X + 1 = 9) = P(X = 4)) = \frac{1}{6}$$
$$g(11) = P(Y = 11) = P(2X + 1 = 11) = P(X = 5)) = \frac{1}{6}$$
$$g(13) = P(Y = 13) = P(2X + 1 = 13) = P(X = 6)) = \frac{1}{6}.$$

We summarize the distribution of $Y$ in the following table.

| $y$ | 3 | 5 | 7 | 9 | 11 | 13 |
|---|---|---|---|---|---|---|
| $g(y)$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

The distribution of $X$ and $2X + 1$ are illustrated below.




In Example 10.1, we computed the distribution (that is, the probability density function) of transformed random variable $Y = \phi(X)$, where $\phi(x) = x^2$. This transformation is not either increasing or decreasing (that is, monotonic) in the space, $R_X$, of the random variable $X$. Therefore, the distribution of $Y$ turn out to be quite different from that of $X$. In Example 10.2, the form of distribution of the transform random variable $Y = \phi(X)$, where $\phi(x) = 2x + 1$, is essentially same. This is mainly due to the fact that $\phi(x) = 2x + 1$ is monotonic in $R_X$.

In this chapter, we shall examine the probability density function of transformed random variables by knowing the density functions of the original random variables. There are several methods for finding the probability density function of a transformed random variable. Some of these methods are:

(1) distribution function method

(2) transformation method

(3) convolution method, and

(4) moment generating function method.

Among these four methods, the transformation method is the most useful one. The convolution method is a special case of this method. The transformation method is derived using the distribution function method.

## 10.1. Distribution Function Method

We have seen in chapter six that an easy way to find the probability density function of a transformation of continuous random variables is to determine its distribution function and then its density function by differentiation.

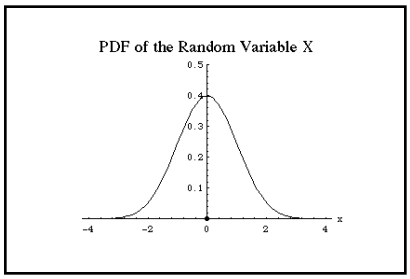
**Example 10.3.** A box is to be constructed so that the height is 4 inches and its base is $X$ inches by $X$ inches. If $X$ has a standard normal distribution, what is the distribution of the volume of the box?

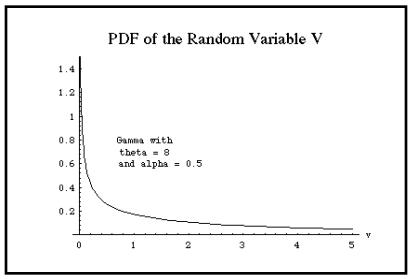
**Answer:** The volume of the box is a random variable, since $X$ is a random variable. This random variable $V$ is given by $V = 4X^2$. To find the density function of $V$, we first determine the form of the distribution function $G(v)$ of $V$ and then we differentiate $G(v)$ to find the density function of $V$. The distribution function of $V$ is given by

$$\begin{aligned} G(v) &= P\left(V \leq v\right) \\ &= P\left(4X^2 \leq v\right) \\ &= P\left(-\frac{1}{2}\sqrt{v} \leq X \leq \frac{1}{2}\sqrt{v}\right) \\ &= \int_{-\frac{1}{2}\sqrt{v}}^{\frac{1}{2}\sqrt{v}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx \\ &= 2\int_{0}^{\frac{1}{2}\sqrt{v}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx \qquad \text{(since the integrand is even).} \end{aligned}$$

Hence, by the Fundamental Theorem of Calculus, we get

$$\begin{aligned} g(v) &= \frac{dG(v)}{dv} \\ &= \frac{d}{dv}\left(2\int_0^{\frac{1}{2}\sqrt{v}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx\right) \\ &= 2\ \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{1}{2}\sqrt{v}\right)^2}\ \left(\frac{1}{2}\right)\ \frac{d\sqrt{v}}{dv} \\ &= \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{8}v}\ \frac{1}{2\sqrt{v}} \\ &= \frac{1}{\Gamma\left(\frac{1}{2}\right)\ \sqrt{8}}\ v^{\frac{1}{2}-1}\ e^{-\frac{v}{8}} \\ &= V \sim GAM\left(8,\ \frac{1}{2}\right). \end{aligned}$$




**Example 10.4.** If the density function of $X$ is

$$f(x) = \begin{cases} \frac{1}{2} & \text{for } -1 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

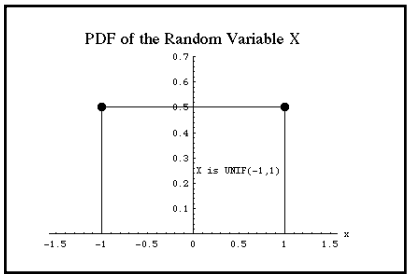
what is the probability density function of $Y = X^2$ ?

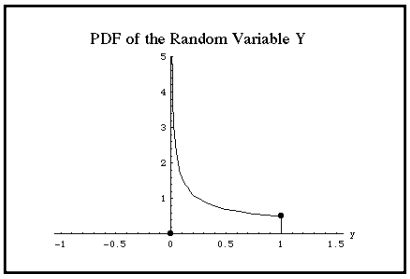
**Answer:** We first find the cumulative distribution function of $Y$ and then by differentiation, we obtain the density of $Y$. The distribution function $G(y)$ of $Y$ is given by

$$\begin{aligned} G(y) &= P\left(Y \leq y\right) \\ &= P\left(X^2 \leq y\right) \\ &= P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}\, dx \\ &= \sqrt{y}. \end{aligned}$$

Hence, the density function of $Y$ is given by

$$\begin{aligned} g(y) &= \frac{dG(y)}{dy} \\ &= \frac{d\sqrt{y}}{dy} \\ &= \frac{1}{2\sqrt{y}} \qquad \text{for} \quad 0 < y < 1. \end{aligned}$$




## 10.2. Transformation Method for Univariate Case

The following theorem is the backbone of the transformation method.

**Theorem 10.1.** Let $X$ be a continuous random variable with probability density function $f(x)$. Let $y = T(x)$ be an increasing (or decreasing) function. Then the density function of the random variable $Y = T(X)$ is given by

$$g(y) = \left|\frac{dx}{dy}\right| f(W(y))$$

where $x = W(y)$ is the inverse function of $T(x)$.

**Proof:** Suppose $y = T(x)$ is an increasing function. The distribution function $G(y)$ of $Y$ is given by

$$\begin{aligned} G(y) &= P\left(Y \leq y\right) \\ &= P\left(T(X) \leq y\right) \\ &= P\left(X \leq W(y)\right) \\ &= \int_{-\infty}^{W(y)} f(x)\, dx. \end{aligned}$$

Then, differentiating we get the density function of $Y$, which is

$$\begin{aligned} g(y) &= \frac{dG(y)}{dy} \\ &= \frac{d}{dy}\left(\int_{-\infty}^{W(y)} f(x)\,dx\right) \\ &= f(W(y))\,\frac{dW(y)}{dy} \\ &= f(W(y))\,\frac{dx}{dy} \qquad\qquad (\text{since} \quad x = W(y)). \end{aligned}$$

On the other hand, if $y = T(x)$ is a decreasing function, then the distribution function of $Y$ is given by

$$\begin{aligned} G(y) &= P\left(Y \leq y\right) \\ &= P\left(T(X) \leq y\right) \\ &= P\left(X \geq W(y)\right) \qquad (\text{since } T(x) \text{ is decreasing}) \\ &= 1 - P\left(X \leq W(y)\right) \\ &= 1 - \int_{-\infty}^{W(y)} f(x)\,dx. \end{aligned}$$

As before, differentiating we get the density function of $Y$, which is

$$\begin{aligned} g(y) &= \frac{dG(y)}{dy} \\ &= \frac{d}{dy}\left(1 - \int_{-\infty}^{W(y)} f(x)\,dx\right) \\ &= -f(W(y))\,\frac{dW(y)}{dy} \\ &= -f(W(y))\,\frac{dx}{dy} \qquad\qquad (\text{since} \quad x = W(y)). \end{aligned}$$

Hence, combining both the cases, we get

$$g(y) = \left|\frac{dx}{dy}\right| f(W(y))$$

and the proof of the theorem is now complete.

**Example 10.5.** Let $Z = \frac{X-\mu}{\sigma}$. If $X \sim N\left(\mu,\, \sigma^2\right)$, what is the probability density function of $Z$?

**Answer:**

$$z = U(x) = \frac{x-\mu}{\sigma}.$$

Hence, the inverse of $U$ is given by

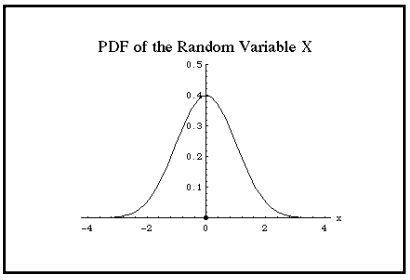
$$\begin{aligned} W(z) &= x \\ &= \sigma\, z + \mu. \end{aligned}$$

Therefore

$$\frac{dx}{dz} = \sigma.$$

Hence, by Theorem 10.1, the density of $Z$ is given by

$$\begin{aligned} g(z) &= \left|\frac{dx}{dz}\right| f(W(y)) \\ &= \sigma\, \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{W(z)-\mu}{\sigma}\right)^2} \\ &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{z\sigma+\mu-\mu}{\sigma}\right)^2} \\ &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}. \end{aligned}$$




**Example 10.6.** Let $Z = \frac{X-\mu}{\sigma}$. If $X \sim N\left(\mu, \sigma^2\right)$, then show that $Z^2$ is chi-square with one degree of freedom, that $Z^2 \sim \chi^2(1)$.

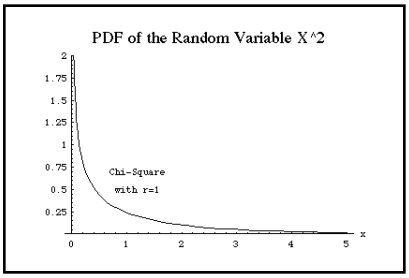
**Answer:**

$$\begin{aligned} y = T(x) &= \left(\frac{x-\mu}{\sigma}\right)^2. \\ x &= \mu + \sigma\sqrt{y}. \\ W(y) &= \mu + \sigma\,\sqrt{y}, \qquad y > 0. \\ \frac{dx}{dy} &= \frac{\sigma}{2\sqrt{y}}. \end{aligned}$$

The density of $Y$ is

$$\begin{aligned} g(y) &= \left|\frac{dx}{dy}\right| f(W(y)) \\ &= \sigma \frac{1}{2\sqrt{y}} f(W(y)) \\ &= \sigma \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{W(y)-\mu}{\sigma}\right)^2} \\ &= \frac{1}{2\sqrt{2\,\pi\,y}} e^{-\frac{1}{2}\left(\frac{\sqrt{y}\sigma+\mu-\mu}{\sigma}\right)^2} \\ &= \frac{1}{2\sqrt{2\,\pi\,y}} e^{-\frac{1}{2}y} \\ &= \frac{1}{2\sqrt{\pi}\,\sqrt{2}} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} \\ &= \frac{1}{2\Gamma\left(\frac{1}{2}\right)\sqrt{2}} y^{-\frac{1}{2}} e^{-\frac{1}{2}y}. \end{aligned}$$

Hence $Y \sim \chi^2(1)$.




**Example 10.7.** Let $Y = -\ln X$. If $X \sim UNIF(0,1)$, then what is the density function of $Y$ where nonzero?

**Answer:** We are given that

$$y = T(x) = -\ln x.$$

Hence, the inverse of $y = T(x)$ is given by

$$\begin{aligned} W(y) &= x \\ &= e^{-y}. \end{aligned}$$

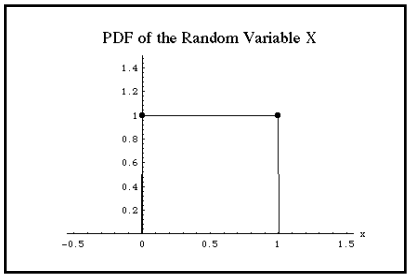
Therefore

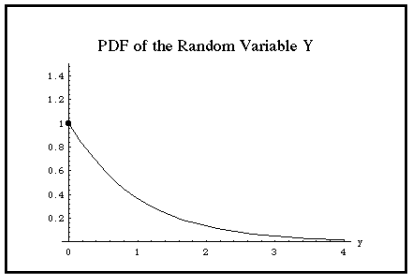
$$\frac{dx}{dy} = -e^{-y}.$$

Hence, by Theorem 10.1, the probability density of $Y$ is given by

$$\begin{aligned} g(y) &= \left|\frac{dx}{dy}\right| f(W(y)) \\ &= e^{-y} f(W(y)) \\ &= e^{-y}. \end{aligned}$$

Thus $Y \sim EXP(1)$. Hence, if $X \sim UNIF(0,1)$, then the random variable $-\ln X \sim EXP(1)$.




Although all the examples we have in this section involve continuous random variables, the transformation method also works for the discrete random variables.

### 10.3. Transformation Method for Bivariate Case

In this section, we extend the Theorem 10.2 to the bivariate case and present some examples to illustrate the importance of this extension. We state this theorem without a proof.

**Theorem 10.2.** Let $X$ and $Y$ be two continuous random variables with joint density $f(x,y)$. Let $U = P(X,Y)$ and $V = Q(X,Y)$ be functions of $X$ and $Y$. If the functions $P(x,y)$ and $Q(x,y)$ have single valued inverses, say $X = R(U,V)$ and $Y = S(U,V)$, then the joint density $g(u,v)$ of $U$ and $V$ is given by

$$g(u,v) = |J| \; f(R(u,v),\, S(u,v)),$$

where $J$ denotes the Jacobian and given by

$$\begin{aligned} J &= \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \\ &= \frac{\partial x}{\partial u}\,\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\,\frac{\partial y}{\partial u}. \end{aligned}$$

**Example 10.8.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} 8\,xy & \text{for } 0 < x < y < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the joint density of $U = \frac{X}{Y}$ and $V = Y$ ?

**Answer:** Since

$$\left.\begin{aligned} U &= \frac{X}{Y} \\ V &= Y \end{aligned}\right\}$$

we get by solving for $X$ and $Y$

$$\left.\begin{aligned} X &= U\,Y = U\,V \\ Y &= V. \end{aligned}\right\}$$

Hence, the Jacobian of the transformation is given by

$$\begin{aligned} J &= \frac{\partial x}{\partial u}\,\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\,\frac{\partial y}{\partial u} \\ &= v \cdot 1 - u \cdot 0 \\ &= v. \end{aligned}$$

The joint density function of $U$ and $V$ is

$$\begin{aligned} g(u,v) &= |J|\ f(R(u,v),\, S(u,v)) \\ &= |v|\, f(uv, v) \\ &= v\, 8\,(uv)\, v \\ &= 8\,uv^3. \end{aligned}$$

Note that, since

$$0 < x < y < 1$$

we have

$$0 < uv < v < 1.$$

The last inequalities yield

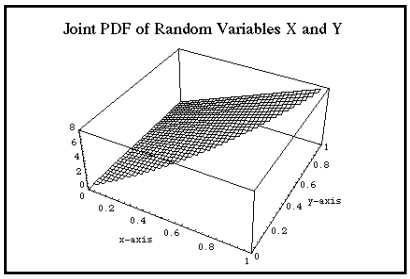
$$\left.\begin{aligned} &0 < uv < v \\ &0 < v < 1. \end{aligned}\right\}$$

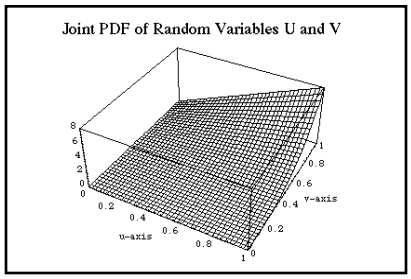
Therefore, we get

$$\left.\begin{array}{l} 0 < u < 1 \\ 0 < v < 1. \end{array}\right\}$$

Thus, the joint density of $U$ and $V$ is given by

$$g(u,v) = \begin{cases} 8\,uv^3 & \text{for } 0 < u < 1;\ 0 < v < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$




**Example 10.9.** Let each of the independent random variables $X$ and $Y$ have the density function

$$f(x) = \begin{cases} e^{-x} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the joint density of $U = X$ and $V = 2X + 3Y$ and the domain on which this density is positive?

**Answer:** Since

$$\left.\begin{array}{l} U = X \\ V = 2X + 3Y, \end{array}\right\}$$

we get by solving for $X$ and $Y$

$$\left.\begin{array}{l} X = U \\ Y = \dfrac{1}{3}\,V - \dfrac{2}{3}\,U. \end{array}\right\}$$

Hence, the Jacobian of the transformation is given by

$$\begin{aligned} J &= \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \\ &= 1 \cdot \left(\frac{1}{3}\right) - 0 \cdot \left(-\frac{2}{3}\right) \\ &= \frac{1}{3}. \end{aligned}$$

The joint density function of $U$ and $V$ is

$$\begin{aligned} g(u,v) &= |J|\ f(R(u,v),\ S(u,v)) \\ &= \left|\frac{1}{3}\right|\ f\left(u,\ \frac{1}{3}v - \frac{2}{3}u\right) \\ &= \frac{1}{3}e^{-u}\ e^{-\frac{1}{3}v+\frac{2}{3}u} \\ &= \frac{1}{3}\ e^{-\left(\frac{u+v}{3}\right)}. \end{aligned}$$

Since

$$\begin{aligned} 0 < x < \infty \\ 0 < y < \infty, \end{aligned}$$

we get

$$\begin{aligned} 0 < u < \infty \\ 0 < v < \infty, \end{aligned}$$

Further, since $v = 2u + 3y$ and $3y > 0$, we have

$$v > 2u.$$

Hence, the domain of $g(u,v)$ where nonzero is given by

$$0 < 2u < v < \infty.$$

The joint density $g(u,v)$ of the random variables $U$ and $V$ is given by

$$g(u,v) = \begin{cases} \frac{1}{3}\ e^{-\left(\frac{u+v}{3}\right)} & \text{for } 0 < 2u < v < \infty \\ 0 & \text{otherwise.} \end{cases}$$

**Example 10.10.** Let $X$ and $Y$ be independent random variables, each with density function

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{for } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda > 0$. Let $U = X + 2Y$ and $V = 2X + Y$. What is the joint density of $U$ and $V$?

**Answer:** Since

$$\left.\begin{aligned} U &= X + 2Y \\ V &= 2X + Y, \end{aligned}\right\}$$

we get by solving for $X$ and $Y$

$$\left.\begin{aligned} X &= -\frac{1}{3}\,U + \frac{2}{3}\,V \\ Y &= \frac{2}{3}\,U - \frac{1}{3}\,V. \end{aligned}\right\}$$

Hence, the Jacobian of the transformation is given by

$$\begin{aligned} J &= \frac{\partial x}{\partial u}\,\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\,\frac{\partial y}{\partial u} \\ &= \left(-\frac{1}{3}\right)\,\left(-\frac{1}{3}\right) - \left(\frac{2}{3}\right)\,\left(\frac{2}{3}\right) \\ &= \frac{1}{9} - \frac{4}{9} \\ &= -\frac{1}{3}. \end{aligned}$$

The joint density function of $U$ and $V$ is

$$\begin{aligned} g(u,v) &= |J|\; f(R(u,v),\, S(u,v)) \\ &= \left|-\frac{1}{3}\right|\; f(R(u,v))\; f(S(u,v)) \\ &= \frac{1}{3}\,\lambda\, e^{\lambda R(u,v)}\,\lambda\, e^{\lambda S(u,v)} \\ &= \frac{1}{3}\,\lambda^2\, e^{\lambda[R(u,v)+S(u,v)]} \\ &= \frac{1}{3}\,\lambda^2\, e^{-\lambda\left(\frac{u+v}{3}\right)}. \end{aligned}$$

Hence, the joint density $g(u,v)$ of the random variables $U$ and $V$ is given by

$$g(u,v) = \begin{cases} \frac{1}{3}\,\lambda^2\, e^{-\lambda\left(\frac{u+v}{3}\right)} & \text{for } 0 < u < \infty;\ 0 < v < \infty \\ 0 & \text{otherwise.} \end{cases}$$

**Example 10.11.** Let $X$ and $Y$ be independent random variables, each with density function

$$f(x) = \frac{1}{\sqrt{2\pi}}\; e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

Let $U = \frac{X}{Y}$ and $V = Y$. What is the joint density of $U$ and $V$? Also, what is the density of $U$?

**Answer:** Since

$$\left.\begin{aligned} U &= \frac{X}{Y} \\ V &= Y, \end{aligned}\right\}$$

we get by solving for $X$ and $Y$

$$\left.\begin{aligned} X &= UV \\ Y &= V. \end{aligned}\right\}$$

Hence, the Jacobian of the transformation is given by

$$\begin{aligned} J &= \frac{\partial x}{\partial u}\,\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\,\frac{\partial y}{\partial u} \\ &= v \cdot (1) - u \cdot (0) \\ &= v. \end{aligned}$$

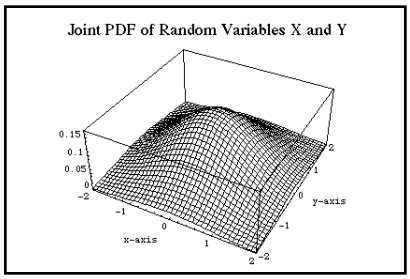
The joint density function of $U$ and $V$ is

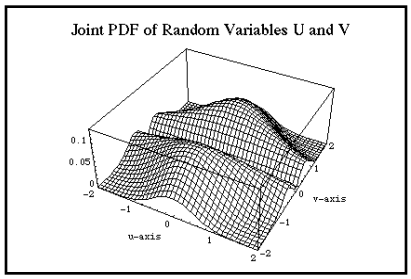
$$\begin{aligned} g(u,v) &= |J|\ f(R(u,v),\, S(u,v)) \\ &= |v|\ f(R(u,v))\ f(S(u,v)) \\ &= |v|\ \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}R^2(u,v)}\ \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}S^2(u,v)} \\ &= |v|\ \frac{1}{2\pi}\, e^{-\frac{1}{2}\left[R^2(u,v)+S^2(u,v)\right]} \\ &= |v|\ \frac{1}{2\pi}\, e^{-\frac{1}{2}\left[u^2v^2+v^2\right]} \\ &= |v|\ \frac{1}{2\pi}\, e^{-\frac{1}{2}v^2\left(u^2+1\right)}. \end{aligned}$$

Hence, the joint density $g(u,v)$ of the random variables $U$ and $V$ is given by

$$g(u,v) = |v|\ \frac{1}{2\pi}\, e^{-\frac{1}{2}v^2\left(u^2+1\right)},$$

where $-\infty < u < \infty$ and $-\infty < v < \infty$.




Next, we want to find the density of $U$. We can obtain this by finding the marginal of $U$ from the joint density of $U$ and $V$. Hence, the marginal $g_1(u)$ of $U$ is given by

$$\begin{aligned}
g_1(u) &= \int_{-\infty}^{\infty} g(u,v)\,dv \\
&= \int_{-\infty}^{\infty} |v|\ \frac{1}{2\pi}\, e^{-\frac{1}{2}v^2\left(u^2+1\right)}\,dv \\
&= \int_{-\infty}^{0} -v\,\frac{1}{2\pi}\, e^{-\frac{1}{2}v^2\left(u^2+1\right)}\,dv + \int_{0}^{\infty} v\,\frac{1}{2\pi}\, e^{-\frac{1}{2}v^2\left(u^2+1\right)}\,dv \\
&= \frac{1}{2\pi}\ \left(\frac{1}{2}\right)\ \left[\frac{2}{u^2+1}\, e^{-\frac{1}{2}\,v^2\left(u^2+1\right)}\right]_{-\infty}^{0} \\
&\quad + \frac{1}{2\pi}\ \left(\frac{1}{2}\right)\ \left[\frac{-2}{u^2+1}\, e^{-\frac{1}{2}\,v^2\left(u^2+1\right)}\right]_{0}^{\infty} \\
&= \frac{1}{2\pi}\ \frac{1}{u^2+1} + \frac{1}{2\pi}\ \frac{1}{u^2+1} \\
&= \frac{1}{\pi\left(u^2+1\right)}.
\end{aligned}$$

Thus $U \sim CAU(1)$.

**Remark 10.1.** If $X$ and $Y$ are independent and standard normal random variables, then the quotient $\frac{X}{Y}$ is always a Cauchy random variable. However, the converse of this is not true. For example, if $X$ and $Y$ are independent and each have the same density function

$$f(x) = \frac{\sqrt{2}}{\pi}\ \frac{x^2}{1+x^4}, \qquad -\infty < x < \infty,$$

then it can be shown that the random variable $\frac{X}{Y}$ is a Cauchy random variable. Laha (1959) and Kotlarski (1960) have given a complete description of the family of all probability density function $f$ such that the quotient $\frac{X}{Y}$

follows the standard Cauchy distribution whenever $X$ and $Y$ are independent and identically distributed random variables with common density $f$.

**Example 10.12.** Let $X$ have a Poisson distribution with mean $\lambda$. Find a transformation $T(x)$ so that $Var\,(\,T(X)\,)$ is free of $\lambda$, for large values of $\lambda$.

**Answer:** We expand the function $T(x)$ by Taylor's series about $\lambda$. Then, neglecting the higher orders terms for large values of $\lambda$, we get

$$T(x) = T(\lambda) + (x-\lambda)\,T'(\lambda) + \cdots\cdots$$

where $T'(\lambda)$ represents derivative of $T(x)$ at $x = \lambda$. Now, we compute the variance of $T(X)$.

$$\begin{aligned} Var\,(\,T(X)\,) &= Var\,(\,T(\lambda) + (X-\lambda)\,T'(\lambda) + \cdots\,) \\ &= Var\,(\,T(\lambda)\,) + Var\,(\,(X-\lambda)\,T'(\lambda)\,) \\ &= 0 + [T'(\lambda)]^2\; Var(X-\lambda) \\ &= [T'(\lambda)]^2\; Var(X) \\ &= [T'(\lambda)]^2\; \lambda. \end{aligned}$$

We want $Var(\,T(X)\,)$ to be free of $\lambda$ for large $\lambda$. Therefore, we have

$$[T'(\lambda)]^2\; \lambda = k,$$

where $k$ is a constant. From this, we get

$$T'(\lambda) = \frac{c}{\sqrt{\lambda}},$$

where $c = \sqrt{k}$. Solving this differential equation, we get

$$\begin{aligned} T(\lambda) &= c \int \frac{1}{\sqrt{\lambda}}\, d\lambda \\ &= 2c\,\sqrt{\lambda}. \end{aligned}$$

Hence, the transformation $T(x) = 2c\sqrt{x}$ will free $Var(\,T(X)\,)$ of $\lambda$ if the random variable $X \sim POI(\lambda)$.

**Example 10.13.** Let $X \sim POI(\lambda_1)$ and $Y \sim POI(\lambda_2)$. What is the probability density function of $X+Y$ if $X$ and $Y$ are independent?

**Answer:** Let us denote $U = X + Y$ and $V = X$. First of all, we find the joint density of $U$ and $V$ and then summing the joint density we determine

the marginal of $U$ which is the density function of $X+Y$? Now writing $X$ and $Y$ in terms of $U$ and $V$, we get

$$\left.\begin{aligned} X &= V \\ Y &= U - X = U - V. \end{aligned}\right\}$$

Hence, the Jacobian of the transformation is given by

$$\begin{aligned} J &= \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \\ &= (0)(-1) - (1)(1) \\ &= -1. \end{aligned}$$

The joint density function of $U$ and $V$ is

$$\begin{aligned} g(u,v) &= |J|\ f(R(u,v),\, S(u,v)) \\ &= |-1|\ f(v, u-v) \\ &= f(v)\, f(u-v) \\ &= \left(\frac{e^{-\lambda_1}\lambda_1^v}{v!}\right)\left(\frac{e^{-\lambda_2}\lambda_2^{u-v}}{(u-v)!}\right) \\ &= \frac{e^{-(\lambda_1+\lambda_2)}\lambda_1^v\lambda_2^{u-v}}{(v)!\,(u-v)!}, \end{aligned}$$

where $v = 0, 1, 2, ..., u$ and $u = 0, 1, 2, ..., \infty$. Hence, the marginal density of $U$ is given by

$$\begin{aligned} g_1(u) &= \sum_{v=0}^{u} \frac{e^{-(\lambda_1+\lambda_2)}\lambda_1^v\lambda_2^{u-v}}{(v)!\,(u-v)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{v=0}^{u} \frac{\lambda_1^v\lambda_2^{u-v}}{(v)!\,(u-v)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{v=0}^{u} \frac{1}{u!}\binom{u}{v}\lambda_1^v\lambda_2^{u-v} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{u!}\,(\lambda_1+\lambda_2)^u. \end{aligned}$$

Thus, the density function of $U = X + Y$ is given by

$$g_1(u) = \begin{cases} \frac{e^{-(\lambda_1+\lambda_2)}}{u!}\,(\lambda_1+\lambda_2)^u & \text{for } u = 0, 1, 2, ..., \infty \\ 0 & \text{otherwise.} \end{cases}$$

This example tells us that if $X \sim POI(\lambda_1)$ and $Y \sim POI(\lambda_2)$ and they are independent, then $X + Y \sim POI(\lambda_1 + \lambda_2)$.

**Theorem 10.3.** Let the joint density of the random variables $X$ and $Y$ be $f(x, y)$. Then probability density functions of $X + Y$, $XY$, and $\frac{Y}{X}$ are given by

$$h_{X+Y}(v) = \int_{-\infty}^{\infty} f(u, v-u)\, du$$
$$h_{XY}(v) = \int_{-\infty}^{\infty} \frac{1}{|u|} f\left(u, \frac{v}{u}\right)\, du$$
$$h_{\frac{X}{Y}}(v) = \int_{-\infty}^{\infty} |u|\, f\left(u, vu\right)\, du,$$

respectively.

**Proof:** Let $U = X$ and $V = X + Y$. So that $X = R(U, V) = U$, and $Y = S(U, V) = V - U$. Hence, the Jacobian of the transformation is given by

$$J = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} = 1.$$

The joint density function of $U$ and $V$ is

$$\begin{aligned} g(u, v) &= |J|\ f(R(u, v),\, S(u, v)) \\ &= f(R(u, v),\, S(u, v)) \\ &= f(u, v-u). \end{aligned}$$

Hence, the marginal density of $V = X + Y$ is given by

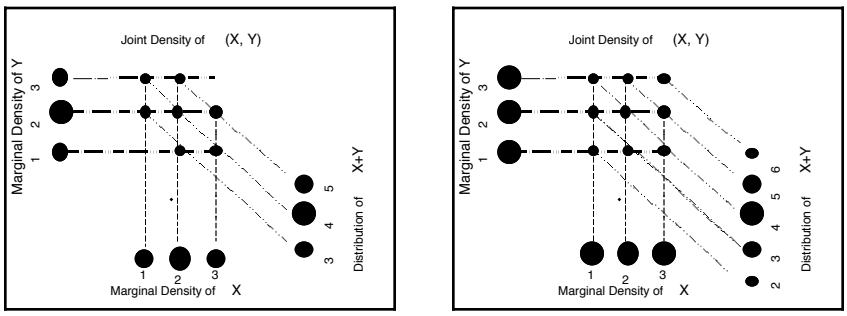
$$h_{X+Y}(v) = \int_{-\infty}^{\infty} f(u, v-u)\, du.$$

Similarly, one can obtain the other two density functions. This completes the proof.

In addition, if the random variables $X$ and $Y$ in Theorem 10.3 are independent and have the probability density functions $f(x)$ and $g(y)$ respectively, then we have

$$h_{X+Y}(z) = \int_{-\infty}^{\infty} g(y)\, f(z-y)\, dy$$
$$h_{XY}(z) = \int_{-\infty}^{\infty} \frac{1}{|y|}\, g(y)\, f\left(\frac{z}{y}\right)\, dy$$
$$h_{\frac{X}{Y}}(z) = \int_{-\infty}^{\infty} |y|\, g(y)\, f\left(zy\right)\, dy.$$

Each of the following figures shows how the distribution of the random variable $X + Y$ is obtained from the joint distribution of $(X, Y)$.




**Example 10.14.** Roll an unbiased die twice. If $X$ denotes the outcome in the first roll and $Y$ denotes the outcome in the second roll, what is the distribution of the random variable $Z = \max\{X, Y\}$?

**Answer:** The space of $X$ is $R_X = \{1, 2, 3, 4, 5, 6\}$. Similarly, the space of $Y$ is $R_Y = \{1, 2, 3, 4, 5, 6\}$. Hence the space of the random variable $(X, Y)$ is $R_X \times R_Y$. The following table shows the distribution of $(X, Y)$.

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 2 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 3 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 4 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 5 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 6 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| | 1 | 2 | 3 | 4 | 5 | 6 |

The space of the random variable $Z = \max\{X, Y\}$ is $R_Z = \{1, 2, 3, 4, 5, 6\}$. Thus $Z = 1$ only if $(X, Y) = (1, 1)$. Hence $P(Z = 1) = \frac{1}{36}$. Similarly, $Z = 2$ only if $(X, Y) = (1, 2), (2, 2)$ or $(2, 1)$. Hence, $P(Z = 2) = \frac{3}{36}$. Proceeding in a similar manner, we get the distribution of $Z$ which is summarized in the table below.

| $z$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $h(z)$ | $\frac{1}{36}$ | $\frac{3}{36}$ | $\frac{5}{36}$ | $\frac{7}{36}$ | $\frac{9}{36}$ | $\frac{11}{36}$ |

In this example, the random variable $Z$ may be described as *the best out of two rolls*. Note that the probability density of $Z$ can also be stated as

$$h(z) = \frac{2z-1}{36}, \qquad \text{for } z \in \{1,2,3,4,5,6\}.$$

### 10.4. Convolution Method for Sums of Random Variables

In this section, we illustrate how convolution technique can be used in finding the distribution of the sum of random variables when they are independent. This convolution technique does not work if the random variables are not independent.

**Definition 10.1.** Let $f$ and $g$ be two real valued functions. The convolution of $f$ and $g$, denoted by $f \star g$, is defined as

$$\begin{aligned}(f \star g)(z) &= \int_{-\infty}^{\infty} f(z-y)\, g(y)\, dy \\ &= \int_{-\infty}^{\infty} g(z-x)\, f(x)\, dx.\end{aligned}$$

Hence from this definition it is clear that $f \star g = g \star f$.

Let $X$ and $Y$ be two independent random variables with probability density functions $f(x)$ and $g(y)$. Then by Theorem 10.3, we get

$$h(z) = \int_{-\infty}^{\infty} f(z-y)\, g(y)\, dy.$$

Thus, this result shows that the density of the random variable $Z = X + Y$ is the convolution of the density of $X$ with the density of $Y$.

**Example 10.15.** What is the probability density of the sum of two independent random variables, each of which is uniformly distributed over the interval [0, 1]?

**Answer:** Let $Z = X + Y$, where $X \sim UNIF(0,1)$ and $Y \sim UNIF(0,1)$. Hence, the density function $f(x)$ of the random variable $X$ is given by

$$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the density function $g(y)$ of $Y$ is given by

$$g(y) = \begin{cases} 1 & \text{for } 0 \leq y \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Since $X$ and $Y$ are independent, the density function of $Z$ can be obtained by the method of convolution. Since, the sum $z = x + y$ is between 0 and 2, we consider two cases. First, suppose $0 \leq z \leq 1$, then

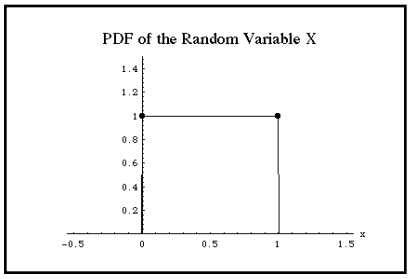
$$\begin{aligned} h(z) &= (f \star g)(z) \\ &= \int_{-\infty}^{\infty} f(z-x)\, g(x)\, dx \\ &= \int_{0}^{1} f(z-x)\, g(x)\, dx \\ &= \int_{o}^{z} f(z-x)\, g(x)\, dx + \int_{z}^{1} f(z-x)\, g(x)\, dx \\ &= \int_{o}^{z} f(z-x)\, g(x)\, dx + 0 \qquad (\text{since } f(z-x) = 0 \text{ between } z \text{ and } 1) \\ &= \int_{0}^{z} dx \\ &= z. \end{aligned}$$

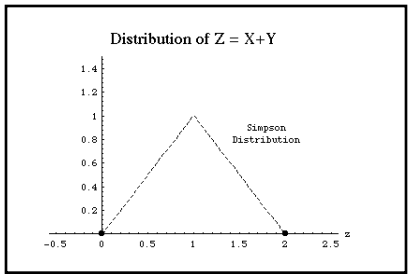
Similarly, if $1 \leq z \leq 2$, then

$$\begin{aligned} h(z) &= (f \star g)(z) \\ &= \int_{-\infty}^{\infty} f(z-x)\, g(x)\, dx \\ &= \int_{0}^{1} f(z-x)\, g(x)\, dx \\ &= \int_{0}^{z-1} f(z-x)\, g(x)\, dx + \int_{z-1}^{1} f(z-x)\, g(x)\, dx \\ &= 0 + \int_{z-1}^{1} f(z-x)\, g(x)\, dx \ (\text{since } f(z-x) = 0 \text{ between } 0 \text{ and } z-1) \\ &= \int_{z-1}^{1} dx \\ &= 2 - z. \end{aligned}$$

Thus, the density function of $Z = X + Y$ is given by

$$h(z) = \begin{cases} 0 & \text{for } -\infty < z \leq 0 \\ z & \text{for } 0 \leq z \leq 1 \\ 2 - z & \text{for } 1 \leq z \leq 2 \\ 0 & \text{for } 2 < z < \infty \ . \end{cases}$$




The graph of this density function looks like a tent and it is called a tent function. However, in literature, this density function is known as the Simpson's distribution.

**Example 10.16.** What is the probability density of the sum of two independent random variables, each of which is gamma with parameter $\alpha = 1$ and $\theta = 1$ ?

**Answer:** Let $Z = X + Y$, where $X \sim GAM(1, 1)$ and $Y \sim GAM(1, 1)$. Hence, the density function $f(x)$ of the random variable $X$ is given by

$$f(x) = \begin{cases} e^{-x} & \text{for } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the density function $g(y)$ of $Y$ is given by

$$g(y) = \begin{cases} e^{-y} & \text{for } 0 < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Since $X$ and $Y$ are independent, the density function of $Z$ can be obtained by the method of convolution. Notice that the sum $z = x + y$ is between 0

and $\infty$, and $0 < x < z$. Hence, the density function of $Z$ is given by

$$\begin{aligned} h(z) &= (f \star g)(z) \\ &= \int_{-\infty}^{\infty} f(z-x)\, g(x)\, dx \\ &= \int_{0}^{\infty} f(z-x)\, g(x)\, dx \\ &= \int_{0}^{z} e^{-(z-x)}\, e^{-x}\, dx \\ &= \int_{0}^{z} e^{-z+x}\, e^{-x}\, dx \\ &= \int_{0}^{z} e^{-z}\, dx \\ &= z\, e^{-z} \\ &= \frac{1}{\Gamma(2)\, 1^2}\; z^{2-1}\; e^{-\frac{z}{1}}. \end{aligned}$$

Hence $Z \sim GAM(1,2)$. Thus, if $X \sim GAM(1,1)$ and $Y \sim GAM(1,1)$, then $X + Y \sim GAM(1,2)$, that $X + Y$ is a gamma with $\alpha = 2$ and $\theta = 1$. Recall that a gamma random variable with $\alpha = 1$ is known as an exponential random variable with parameter $\theta$. Thus, in view of the above example, we see that the sum of two independent exponential random variables is not necessarily an exponential variable.

**Example 10.17.** What is the probability density of the sum of two independent random variables, each of which is standard normal?

**Answer:** Let $Z = X + Y$, where $X \sim N(0,1)$ and $Y \sim N(0,1)$. Hence, the density function $f(x)$ of the random variable $X$ is given by

$$f(x) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{x^2}{2}}$$

Similarly, the density function $g(y)$ of $Y$ is given by

$$g(y) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{y^2}{2}}$$

Since $X$ and $Y$ are independent, the density function of $Z$ can be obtained by the method of convolution. Notice that the sum $z = x + y$ is between $-\infty$

and $\infty$. Hence, the density function of $Z$ is given by

$$\begin{aligned}
h(z) &= (f \star g)(z) \\
&= \int_{-\infty}^{\infty} f(z-x)\, g(x)\, dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{\frac{-(z-x)^2}{2}}\, e^{-\frac{x^2}{2}}\, dx \\
&= \frac{1}{2\pi}\, e^{-\frac{z^2}{4}} \int_{-\infty}^{\infty} e^{-\left(x-\frac{z}{2}\right)^2}\, dx \\
&= \frac{1}{2\pi}\, e^{-\frac{z^2}{4}}\, \sqrt{\pi} \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}}\, e^{-\left(x-\frac{z}{2}\right)^2}\, dx\right] \\
&= \frac{1}{2\pi}\, e^{-\frac{z^2}{4}} \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}}\, e^{-w^2}\, dw\right], \qquad \text{where } w = x - \frac{z}{2} \\
&= \frac{1}{\sqrt{4\pi}}\, e^{-\frac{z^2}{4}} \\
&= \frac{1}{\sqrt{4\pi}}\, e^{-\frac{1}{2}\left(\frac{z-0}{\sqrt{2}}\right)^2}.
\end{aligned}$$

The integral in the brackets equals to one, since the integrand is the normal density function with mean $\mu = 0$ and variance $\sigma^2 = \frac{1}{2}$. Hence sum of two standard normal random variables is again a normal random variable with mean zero and variance 2.

**Example 10.18.** What is the probability density of the sum of two independent random variables, each of which is Cauchy?

**Answer:** Let $Z = X + Y$, where $X \sim N(0,1)$ and $Y \sim N(0,1)$. Hence, the density function $f(x)$ of the random variable $X$ and $Y$ are is given by

$$f(x) = \frac{1}{\pi\,(1+x^2)} \quad \text{and} \quad g(y) = \frac{1}{\pi\,(1+y^2)},$$

respectively. Since $X$ and $Y$ are independent, the density function of $Z$ can be obtained by the method of convolution. Notice that the sum $z = x + y$ is between $-\infty$ and $\infty$. Hence, the density function of $Z$ is given by

$$\begin{aligned}
h(z) &= (f \star g)(z) \\
&= \int_{-\infty}^{\infty} f(z-x)\, g(x)\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi\,(1+(z-x)^2)}\, \frac{1}{\pi\,(1+x^2)}\, dx \\
&= \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{1+(z-x)^2}\, \frac{1}{1+x^2}\, dx.
\end{aligned}$$

To integrate the above integral, we decompose the integrand using partial fraction decomposition. Hence

$$\frac{1}{1+(z-x)^2}\,\frac{1}{1+x^2} = \frac{2\,A\,x+B}{1+x^2} + \frac{2\,C\,(z-x)+D}{1+(z-x)^2}$$

where

$$A = \frac{1}{z\,(4+z^2)} = C \quad \text{and} \quad B = \frac{1}{4+z^2} = D.$$

Now integration yields

$$\begin{aligned}
&\frac{1}{\pi^2}\int_{-\infty}^{\infty} \frac{1}{1+(z-x)^2}\,\frac{1}{1+x^2}\,dx \\
&= \frac{1}{\pi^2\,z^2\,(4+z^2)}\left[z\ln\left(\frac{1+x^2}{1+(z-x)^2}\right) + z^2\,\tan^{-1}x - z^2\tan^{-1}(z-x)\right]_{-\infty}^{\infty} \\
&= \frac{1}{\pi^2\,z^2\,(4+z^2)}\left[\,0+z^2\,\pi+z^2\,\pi\,\right] \\
&= \frac{2}{\pi\,(4+z^2)}.
\end{aligned}$$

Hence the sum of two independent Cauchy random variables is not a Cauchy random variable.

If $X \sim CAU(0)$ and $Y \sim CAU(0)$, then it can be easily shown using Example 10.18 that the random variable $Z = \frac{X+Y}{2}$ is again Cauchy, that is $Z \sim CAU(0)$. This is a remarkable property of the Cauchy distribution.

So far we have considered the convolution of two continuous independent random variables. However, the concept can be modified to the case when the random variables are discrete.

Let $X$ and $Y$ be two discrete random variables both taking on values that are integers. Let $Z = X + Y$ be the sum of the two random variables. Hence $Z$ takes values on the set of integers. Suppose that $X = n$ where $n$ is some integer. Then $Z = z$ if and only if $Y = z - n$. Thus the events $(Z = z)$ is the union of the pair wise disjoint events $(X = n)$ and $(Y = z - n)$ where $n$ runs over the integers. The cdf $H(z)$ of $Z$ can be obtained as follows:

$$P(Z = z) = \sum_{n=-\infty}^{\infty} P(X = n)\,P(Y = z - n)$$

which is

$$h(z) = \sum_{n=-\infty}^{\infty} f(n)\,g(z-n),$$

where $F(x)$ and $G(y)$ are the cdf of $X$ and $Y$, respectively.

**Definition 10.2.** Let $X$ and $Y$ be two independent integer-valued discrete random variables, with pdfs $f(x)$ and $g(y)$ respectively. Then the convolution of $f(x)$ and $g(y)$ is the cdf $h = f \star g$ given by

$$h(m) = \sum_{n=-\infty}^{\infty} f(n)\, g(m-n),$$

for $m = -\infty, ..., -2, -1, 0, 1, 2, ....\infty$. The function $h(z)$ is the pdf of the discrete random variable $Z = X + Y$.

**Example 10.19.** Let each of the random variable $X$ and $Y$ represents the outcomes of a six-sided die. What is the cumulative density function of the sum of $X$ and $Y$?

**Answer:** Since the range of $X$ as well as $Y$ is $\{1, 2, 3, 4, 5, 6\}$, the range of $Z = X + Y$ is $R_Z = \{2, 3, 4, ..., 11, 12\}$. The pdf of $Z$ is given by

$$\begin{aligned}
h(2) &= f(1)\, g(1) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \\
h(3) &= f(1)\, g(2) + f(2)\, g(1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{2}{36} \\
h(4) &= f(1)\, g(3) + h(2)\, g(2) + f(3)\, g(1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{3}{36}.
\end{aligned}$$

Continuing in this manner we obtain $h(5) = \frac{4}{36}$, $h(6) = \frac{5}{36}$, $h(7) = \frac{6}{36}$, $h(8) = \frac{5}{36}$, $h(9) = \frac{4}{36}$, $h(10) = \frac{3}{36}$, $h(11) = \frac{2}{36}$, and $h(12) = \frac{1}{36}$. Putting these into one expression we have

$$\begin{aligned}
h(z) &= \sum_{n=1}^{z-1} f(n) g(z-n) \\
&= \frac{6 - |z-7|}{36}, \qquad z = 2, 3, 4, ..., 12.
\end{aligned}$$

It is easy to note that the convolution operation is commutative as well as associative. Using the associativity of the convolution operation one can compute the pdf of the random variable $S_n = X_1 + X_2 + \cdots + X_n$, where $X_1, X_2, ..., X_n$ are random variables each having the same pdf $f(x)$. Then the pdf of $S_1$ is $f(x)$. Since $S_n = S_{n-1} + X_n$ and the pdf of $X_n$ is $f(x)$, the pdf of $S_n$ can be obtained by induction.

### 10.5. Moment Generating Function Method

We know that if $X$ and $Y$ are independent random variables, then

$$M_{X+Y}(t) = M_X(t)\, M_Y(t).$$

This result can be used to find the distribution of the sum $X + Y$. Like the convolution method, this method can be used in finding the distribution of $X + Y$ if $X$ and $Y$ are independent random variables. We briefly illustrate the method using the following example.

**Example 10.20.** Let $X \sim POI(\lambda_1)$ and $Y \sim POI(\lambda_2)$. What is the probability density function of $X + Y$ if $X$ and $Y$ are independent?

**Answer:** Since, $X \sim POI(\lambda_1)$ and $Y \sim POI(\lambda_2)$, we get

$$M_X(t) = e^{\lambda_1 (e^t - 1)}$$

and

$$M_Y(t) = e^{\lambda_2 (e^t - 1)}.$$

Further, since $X$ and $Y$ are independent, we have

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\, M_Y(t) \\ &= e^{\lambda_1 (e^t - 1)}\, e^{\lambda_2 (e^t - 1)} \\ &= e^{\lambda_1 (e^t - 1) + \lambda_2 (e^t - 1)} \\ &= e^{(\lambda_1 + \lambda_2)(e^t - 1)}, \end{aligned}$$

that is, $X+Y \sim POI(\lambda_1+\lambda_2)$. Hence the density function $h(z)$ of $Z = X+Y$ is given by

$$h(z) = \begin{cases} \frac{e^{-(\lambda_1+\lambda_2)}}{z!}\ (\lambda_1 + \lambda_2)^z & \text{for } z = 0, 1, 2, 3, ... \\ 0 & \text{otherwise.} \end{cases}$$

Compare this example to Example 10.13. You will see that moment method has a definite advantage over the convolution method. However, if you use the moment method in Example 10.15, then you will have problem identifying the form of the density function of the random variable $X + Y$. Thus, it is difficult to say which method always works. Most of the time we pick a particular method based on the type of problem at hand.

**Example 10.21.** What is the probability density function of the sum of two independent random variable, each of which is gamma with parameters $\theta$ and $\alpha$?

**Answer:** Let $X$ and $Y$ be two independent gamma random variables with parameters $\theta$ and $\alpha$, that is $X \sim GAM(\theta, \alpha)$ and $Y \sim GAM(\theta, \alpha)$. From Theorem 6.3, the moment generating functions of $X$ and $Y$ are obtained as $M_X(t) = (1-\theta)^{-\alpha}$ and $M_Y(t) = (1-\theta)^{-\alpha}$, respectively. Since, $X$ and $Y$ are independent, we have

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\, M_Y(t) \\ &= (1-\theta)^{-\alpha}\ (1-\theta)^{-\alpha} \\ &= (1-\theta)^{-2\alpha}. \end{aligned}$$

Thus $X+Y$ has a moment generating function of a gamma random variable with parameters $\theta$ and $2\alpha$. Therefore

$$X + Y \sim GAM(\theta, 2\alpha).$$

### 10.6. Review Exercises

**1.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} e^{-2x} + \frac{1}{2}\, e^{-x} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

If $Y = e^{-2X}$, then what is the density function of $Y$ where nonzero?

**2.** Suppose that $X$ is a random variable with density function

$$f(x) = \begin{cases} \frac{3}{8}\, x^2 & \text{for } 0 < x < 2 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = mX^2$, where $m$ is a fixed positive number. What is the density function of $Y$ where nonzero?

**3.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} 2\, e^{-2x} & \text{for } x > 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

and let $Y = e^{-X}$. What is the density function $g(y)$ of $Y$ where nonzero?

**4.** What is the probability density of the sum of two independent random variables, each of which is uniformly distributed over the interval $[-2, 2]$?

**5.** Let $X$ and $Y$ be random variables with joint density function

$$f(x,y) = \begin{cases} e^{-x} & \text{for } 0 < x < \infty;\ 0 < y < 1 \\ \\ 0 & \text{elsewhere .} \end{cases}$$

If $Z = X + 2Y$, then what is the joint density of $X$ and $Z$ where nonzero?

**6.** Let $X$ be a continuous random variable with density function

$$f(x) = \begin{cases} \frac{2}{x^2} & \text{for } 1 < x < 2 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

If $Y = \sqrt{X}$, then what is the density function of $Y$ for $1 < y < \sqrt{2}$?

**7.** What is the probability density of the sum of two independent random variables, each of which has the density function given by

$$f(x) = \begin{cases} \frac{10-x}{50} & \text{for } 0 < x < 10 \\ \\ 0 & \text{elsewhere?} \end{cases}$$

**8.** What is the probability density of the sum of two independent random variables, each of which has the density function given by

$$f(x) = \begin{cases} \frac{a}{x^2} & \text{for } a \leq x < \infty \\ \\ 0 & \text{elsewhere?} \end{cases}$$

**9.** Roll an unbiased die 3 times. If $U$ denotes the outcome in the first roll, $V$ denotes the outcome in the second roll, and $W$ denotes the outcome of the third roll, what is the distribution of the random variable $Z = \max\{U, V, W\}$?

**10.** The probability density of $V$, the velocity of a gas molecule, by Maxwell-Boltzmann law is given by

$$f(v) = \begin{cases} \frac{4\,h^3}{\sqrt{\pi}}\ v^2\ e^{-h^2 v^2} & \text{for } 0 \leq v < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $h$ is the Plank's constant. If $m$ represents the mass of a gas molecule, then what is the probability density of the kinetic energy $Z = \frac{1}{2}\, mV^2$ ?

**11.** If the random variables $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} \frac{6}{7}\,x & \text{for } 1 \le x+y \le 2,\ x \ge 0,\ y \ge 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the joint density of $U = 2X + 3Y$ and $V = 4X + Y$?

**12.** If the random variables $X$ and $Y$ have the joint density

$$f(x,y) = \begin{cases} \frac{6}{7}\,x & \text{for } 1 \le x+y \le 2,\ x \ge 0,\ y \ge 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the density of $\frac{X}{Y}$?

**13.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} \frac{5}{16}\,xy^2 & \text{for } 0 < x < y < 2 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the joint density function of $U = 3X - 2Y$ and $V = X + 2Y$ where it is nonzero?

**14.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} 4x & \text{for } 0 < x < \sqrt{y} < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the joint density function of $U = 5X - 2Y$ and $V = 3X + 2Y$ where it is nonzero?

**15.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} 4x & \text{for } 0 < x < \sqrt{y} < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the density function of $X - Y$ ?

**16.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} 4x & \text{for } 0 < x < \sqrt{y} < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the density function of $\frac{X}{Y}$ ?

**17.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} 4x & \text{for } 0 < x < \sqrt{y} < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the density function of $XY$ ?

**18.** Let $X$ and $Y$ have the joint probability density function

$$f(x,y) = \begin{cases} \frac{5}{16}\, xy^2 & \text{for } 0 < x < y < 2 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the density function of $\frac{Y}{X}$?

**19.** If $X$ an uniform random variable on the interval $[0,2]$ and $Y$ is an uniform random variable on the interval $[0,3]$, then what is the joint probability density function of $X+Y$ if they are independent?

**20.** What is the probability density function of the sum of two independent random variable, each of which is binomial with parameters $n$ and $p$?

**21.** What is the probability density function of the sum of two independent random variable, each of which is exponential with mean $\theta$?

**22.** What is the probability density function of the average of two independent random variable, each of which is Cauchy with parameter $\theta = 0$?

**23.** What is the probability density function of the average of two independent random variable, each of which is normal with mean $\mu$ and variance $\sigma^2$?

**24.** Both roots of the quadratic equation $x^2 + \alpha x + \beta = 0$ can take all values from $-1$ to $+1$ with equal probabilities. What are the probability density functions of the coefficients $\alpha$ and $\beta$?

**25.** If $A$, $B$, $C$ are independent random variables uniformly distributed on the interval from zero to one, then what is the probability that the quadratic equation $Ax^2 + Bx + C = 0$ has real solutions?

**26.** The price of a stock on a given trading day changes according to the distribution $f(-1) = \frac{1}{4}$, $f(0) = \frac{1}{2}$, $f(1) = \frac{1}{8}$, and $f(2) = \frac{1}{8}$. Find the distribution for the change in stock price after two (independent) trading days.

# Chapter 11

# SOME SPECIAL DISCRETE BIVARIATE DISTRIBUTIONS

In this chapter, we shall examine some bivariate discrete probability density functions. Ever since the first statistical use of the bivariate normal distribution (which will be treated in Chapter 12) by Galton and Dickson in 1886, attempts have been made to develop families of bivariate distributions to describe non-normal variations. In many textbooks, only the bivariate normal distribution is treated. This is partly due to the dominant role the bivariate normal distribution has played in statistical theory. Recently, however, other bivariate distributions have started appearing in probability models and statistical sampling problems. This chapter will focus on some well known bivariate discrete distributions whose marginal distributions are well-known univariate distributions. The book of K.V. Mardia gives an excellent exposition on various bivariate distributions.

## 11.1. Bivariate Bernoulli Distribution

We define a bivariate Bernoulli random variable by specifying the form of the joint probability distribution.

**Definition 11.1.** A discrete bivariate random variable $(X, Y)$ is said to have the bivariate Bernoulli distribution if its joint probability density is of the form

$$f(x,y) = \begin{cases} \frac{1}{x!\,y!\,(1-x-y)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^{1-x-y}, & \text{if } x, y = 0, 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p_1,\, p_2,\, p_1 + p_2 < 1$ and $x + y \leq 1$. We denote a bivariate Bernoulli random variable by writing $(X, Y) \sim BER\,(p_1, p_2)$.

In the following theorem, we present the expected values and the variances of $X$ and $Y$, the covariance between $X$ and $Y$, and their joint moment generating function. Recall that the joint moment generating function of $X$ and $Y$ is defined as $M(s,t) := E\left(e^{sX+tY}\right)$.

**Theorem 11.1.** Let $(X, Y) \sim BER\,(p_1, p_2)$, where $p_1$ and $p_2$ are parameters. Then

$$\begin{aligned} E(X) &= p_1 \\ E(Y) &= p_2 \\ Var(X) &= p_1\,(1 - p_1) \\ Var(Y) &= p_2\,(1 - p_2) \\ Cov(X, Y) &= -p_1\,p_2 \\ M(s,t) &= 1 - p_1 - p_2 + p_1 e^s + p_2 e^t. \end{aligned}$$

**Proof:** First, we derive the joint moment generating function of $X$ and $Y$ and then establish the rest of the results from it. The joint moment generating function of $X$ and $Y$ is given by

$$\begin{aligned} M(s,t) &= E\left(e^{sX+tY}\right) \\ &= \sum_{x=0}^{1}\sum_{y=0}^{1} f(x,y)\, e^{sx+ty} \\ &= f(0,0) + f(1,0)\, e^s + f(0,1)\, e^t + f(1,1)\, e^{t+s} \\ &= 1 - p_1 - p_2 + p_1 e^s + p_2 e^t + 0\, e^{t+s} \\ &= 1 - p_1 - p_2 + p_1 e^s + p_2 e^t. \end{aligned}$$

The expected value of $X$ is given by

$$\begin{aligned} E(X) &= \left.\frac{\partial M}{\partial s}\right|_{(0,0)} \\ &= \left.\frac{\partial}{\partial s}\left(1 - p_1 - p_2 + p_1 e^s + p_2 e^t\right)\right|_{(0,0)} \\ &= p_1\, e^s\,|_{(0,0)} \\ &= p_1. \end{aligned}$$

Similarly, the expected value of $Y$ is given by

$$\begin{aligned} E(Y) &= \left.\frac{\partial M}{\partial t}\right|_{(0,0)} \\ &= \left.\frac{\partial}{\partial t}\left(1 - p_1 - p_2 + p_1 e^s + p_2 e^t\right)\right|_{(0,0)} \\ &= p_2\, e^t\,|_{(0,0)} \\ &= p_2. \end{aligned}$$

The product moment of $X$ and $Y$ is

$$\begin{aligned} E(XY) &= \left.\frac{\partial^2 M}{\partial t\,\partial s}\right|_{(0,0)} \\ &= \left.\frac{\partial^2}{\partial t\,\partial s}\left(1 - p_1 - p_2 + p_1 e^s + p_2 e^t\right)\right|_{(0,0)} \\ &= \left.\frac{\partial}{\partial t}\left(p_1\, e^s\right)\right|_{(0,0)} \\ &= 0. \end{aligned}$$

Therefore the covariance of $X$ and $Y$ is

$$Cov(X,Y) = E(XY) - E(X)\,E(Y) = -p_1 p_2$$

Similarly, it can be shown that

$$E(X^2) = p_1 \qquad \text{and} \qquad E(Y^2) = p_2.$$

Thus, we have

$$Var(X) = E(X^2) - E(X)^2 = p_1 - p_1^2 = p_1\,(1 - p_1)$$

and

$$Var(Y) = E(Y^2) - E(Y)^2 = p_2 - p_2^2 = p_2\,(1 - p_2).$$

This completes the proof of the theorem.

The next theorem presents some information regarding the conditional distributions $f(x/y)$ and $f(y/x)$.

**Theorem 11.2.** Let $(X,Y) \sim BER(p_1,p_2)$, where $p_1$ and $p_2$ are parameters. Then the conditional distributions $f(y/x)$ and $f(x/y)$ are also Bernoulli and

$$E(Y/x) = \frac{p_2\,(1-x)}{1-p_1}$$
$$E(X/y) = \frac{p_1\,(1-y)}{1-p_2}$$
$$Var(Y/x) = \frac{p_2\,(1-p_1-p_2)\,(1-x)}{(1-p_1)^2}$$
$$Var(X/y) = \frac{p_1\,(1-p_1-p_2)\,(1-y)}{(1-p_2)^2}.$$

**Proof:** Notice that

$$\begin{aligned} f(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{f(x,y)}{\sum_{y=0}^{1} f(x,y)} \\ &= \frac{f(x,y)}{f(x,0)+f(x,1)} \qquad x=0,1;\ y=0,1;\ 0\le x+y\le 1. \end{aligned}$$

Hence

$$\begin{aligned} f(1/0) &= \frac{f(0,1)}{f(0,0)+f(0,1)} \\ &= \frac{p_2}{1-p_1-p_2+p_2} \\ &= \frac{p_2}{1-p_1} \end{aligned}$$

and

$$\begin{aligned} f(1/1) &= \frac{f(1,1)}{f(1,0)+f(1,1)} \\ &= \frac{0}{p_1+0} = 0. \end{aligned}$$

Now we compute the conditional expectation $E(Y/x)$ for $x=0,1$. Hence

$$\begin{aligned} E(Y/x=0) &= \sum_{y=0}^{1} y\, f(y/0) \\ &= f(1/0) \\ &= \frac{p_2}{1-p_1} \end{aligned}$$

and

$$E(Y/x=1) = f(1/1) = 0.$$

Merging these together, we have

$$E(Y/x) = \frac{p_2\,(1-x)}{1-p_1} \qquad x = 0, 1.$$

Similarly, we compute

$$\begin{aligned} E(Y^2/x=0) &= \sum_{y=0}^{1} y^2\, f(y/0) \\ &= f(1/0) \\ &= \frac{p_2}{1-p_1} \end{aligned}$$

and

$$E(Y^2/x=1) = f(1/1) = 0.$$

Therefore

$$\begin{aligned} Var(Y/x=0) &= E(Y^2/x=0) - E(Y/x=0)^2 \\ &= \frac{p_2}{1-p_1} - \left(\frac{p_2}{1-p_1}\right)^2 \\ &= \frac{p_2(1-p_1) - p_2^2}{(1-p_1)^2} \\ &= \frac{p_2\,(1-p_1-p_2)}{(1-p_1)^2} \end{aligned}$$

and

$$Var(Y/x=1) = 0.$$

Merging these together, we have

$$Var(Y/x) = \frac{p_2\,(1-p_1-p_2)\,(1-x)}{(1-p_1)^2} \qquad x = 0, 1.$$

The conditional expectation $E(X/y)$ and the conditional variance $Var(X/y)$ can be obtained in a similar manner. We leave their derivations to the reader.

### 11.2. Bivariate Binomial Distribution

The bivariate binomial random variable is defined by specifying the form of the joint probability distribution.

**Definition 11.2.** A discrete bivariate random variable $(X, Y)$ is said to have the bivariate binomial distribution with parameters $n, p_1, p_2$ if its joint probability density is of the form

$$f(x,y) = \begin{cases} \frac{n!}{x!\,y!\,(n-x-y)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^{n-x-y}, & \text{if } x, y = 0, 1, ..., n \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p_1,\ p_2,\ p_1+p_2 < 1$, $x+y \leq n$ and $n$ is a positive integer. We denote a bivariate binomial random variable by writing $(X, Y) \sim BIN\,(n, p_1, p_2)$.

Bivariate binomial distribution is also known as trinomial distribution. It will be shown in the proof of Theorem 11.4 that the marginal distributions of $X$ and $Y$ are $BIN(n, p_1)$ and $BIN(n, p_2)$, respectively.

The following two examples illustrate the applicability of bivariate binomial distribution.

**Example 11.1.** In the city of Louisville on a Friday night, radio station $A$ has 50 percent listeners, radio station $B$ has 30 percent listeners, and radio station $C$ has 20 percent listeners. What is the probability that among 8 listeners in the city of Louisville, randomly chosen on a Friday night, 5 will be listening to station $A$, 2 will be listening to station $B$, and 1 will be listening to station $C$?

**Answer:** Let $X$ denote the number listeners that listen to station $A$, and $Y$ denote the listeners that listen to station $B$. Then the joint distribution of $X$ and $Y$ is bivariate binomial with $n = 8$, $p_1 = \frac{5}{10}$, and $p_2 = \frac{3}{10}$. The probability that among 8 listeners in the city of Louisville, randomly chosen on a Friday night, 5 will be listening to station $A$, 2 will be listening to station $B$, and 1 will be listening to station $C$ is given by

$$\begin{aligned} P(X=5, Y=2) &= f(5,2) \\ &= \frac{n!}{x!\,y!\,(n-x-y)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^{n-x-y} \\ &= \frac{8!}{5!\,2!\,1!} \left(\frac{5}{10}\right)^5 \left(\frac{3}{10}\right)^2 \left(\frac{2}{10}\right) \\ &= 0.0945. \end{aligned}$$

**Example 11.2.** A certain game involves rolling a fair die and watching the numbers of rolls of 4 and 5. What is the probability that in 10 rolls of the die one 4 and three 5 will be observed?

**Answer:** Let $X$ denote the number of 4 and $Y$ denote the number of 5. Then the joint distribution of $X$ and $Y$ is bivariate binomial with $n = 10$, $p_1 = \frac{1}{6}$, $p_2 = \frac{1}{6}$ and $1 - p_1 - p_2 = \frac{4}{6}$. Hence the probability that in 10 rolls of the die one 4 and three 5 will be observed is

$$\begin{aligned}
P(X = 5, Y = 2) &= f(1, 3) \\
&= \frac{n!}{x!\, y!\, (n - x - y)!}\, p_1^x\, p_2^y\, (1 - p_1 - p_2)^{n-x-y} \\
&= \frac{10!}{1!\, 3!\, (10 - 1 - 3)!} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6} - \frac{1}{6}\right)^{10-1-3} \\
&= \frac{10!}{1!\, 3!\, (10 - 1 - 3)!} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^3 \left(\frac{4}{6}\right)^6 \\
&= \frac{573440}{10077696} \\
&= 0.0569.
\end{aligned}$$

Using transformation method discussed in chapter 10, it can be shown that if $X_1$, $X_2$ and $X_3$ are independent binomial random variables, then the joint distribution of the random variables

$$X = X_1 + X_2 \qquad \text{and} \qquad Y = X_1 + X_3$$

is bivariate binomial. This approach is known as trivariate reduction technique for constructing bivariate distribution.

To establish the next theorem, we need a generalization of the binomial theorem which was treated in Chapter 1. The following result generalizes the binomial theorem and can be called trinomial theorem. Similar to the proof of binomial theorem, one can establish

$$(a + b + c)^n = \sum_{x=0}^{n} \sum_{y=0}^{n} \binom{n}{x, y} a^x\, b^y\, c^{n-x-y},$$

where $0 \leq x + y \leq n$ and

$$\binom{n}{x, y} = \frac{n!}{x!\, y!\, (n - x - y)!}.$$

In the following theorem, we present the expected values of $X$ and $Y$, their variances, the covariance between $X$ and $Y$, and the joint moment generating function.

**Theorem 11.3.** Let $(X,Y) \sim BIN(n,p_1,p_2)$, where $n$, $p_1$ and $p_2$ are parameters. Then

$$\begin{aligned} E(X) &= n\,p_1 \\ E(Y) &= n\,p_2 \\ Var(X) &= n\,p_1\,(1-p_1) \\ Var(Y) &= n\,p_2\,(1-p_2) \\ Cov(X,Y) &= -n\,p_1\,p_2 \\ M(s,t) &= \left(1-p_1-p_2+p_1e^s+p_2e^t\right)^n. \end{aligned}$$

**Proof:** First, we find the joint moment generating function of $X$ and $Y$. The moment generating function $M(s,t)$ is given by

$$\begin{aligned} M(s,t) &= E\left(e^{sX+tY}\right) \\ &= \sum_{x=0}^{n}\sum_{y=0}^{n} e^{sx+ty} f(x,y) \\ &= \sum_{x=0}^{n}\sum_{y=0}^{n} e^{sx+ty}\,\frac{n!}{x!\,y!\,(n-x-y)!}\,p_1^x\,p_2^y\,(1-p_1-p_2)^{n-x-y} \\ &= \sum_{x=0}^{n}\sum_{y=0}^{n} \frac{n!}{x!\,y!\,(n-x-y)!}\,(e^s p_1)^x\,\left(e^t p_2\right)^y\,(1-p_1-p_2)^{n-x-y} \\ &= \left(1-p_1-p_2+p_1e^s+p_2e^t\right)^n \qquad \text{(by trinomial theorem).} \end{aligned}$$

The expected value of $X$ is given by

$$\begin{aligned} E(X) &= \left.\frac{\partial M}{\partial s}\right|_{(0,0)} \\ &= \left.\frac{\partial}{\partial s}\left(1-p_1-p_2+p_1e^s+p_2e^t\right)^n\right|_{(0,0)} \\ &= \left.n\left(1-p_1-p_2+p_1e^s+p_2e^t\right)^{n-1}\,p_1\,e^s\right|_{(0,0)} \\ &= n\,p_1. \end{aligned}$$

Similarly, the expected value of $Y$ is given by

$$\begin{aligned} E(Y) &= \left.\frac{\partial M}{\partial t}\right|_{(0,0)} \\ &= \left.\frac{\partial}{\partial t}\left(1-p_1-p_2+p_1e^s+p_2e^t\right)^n\right|_{(0,0)} \\ &= \left.n\left(1-p_1-p_2+p_1e^s+p_2e^t\right)^{n-1}\,p_2\,e^t\right|_{(0,0)} \\ &= n\,p_2. \end{aligned}$$

The product moment of $X$ and $Y$ is

$$\begin{aligned} E(XY) &= \left.\frac{\partial^2 M}{\partial t\,\partial s}\right|_{(0,0)} \\ &= \left.\frac{\partial^2}{\partial t\,\partial s}\left(1-p_1-p_2+p_1e^s+p_2e^t\right)^n\right|_{(0,0)} \\ &= \left.\frac{\partial}{\partial t}\left(n\,\left(1-p_1-p_2+p_1e^s+p_2e^t\right)^{n-1}\,p_1\,e^s\right)\right|_{(0,0)} \\ &= n\,(n-1)p_1p_2. \end{aligned}$$

Therefore the covariance of $X$ and $Y$ is

$$Cov(X,Y) = E(XY) - E(X)\,E(Y) = n(n-1)p_1p_2 - n^2p_1p_2 = -np_1p_2.$$

Similarly, it can be shown that

$$E(X^2) = n(n-1)p_1^2 + np_1 \qquad \text{and} \qquad E(Y^2) = n(n-1)p_2^2 + np_2.$$

Thus, we have

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= n(n-1)p_2^2 + np_2 - n^2p_1^2 \\ &= n\,p_1\,(1-p_1) \end{aligned}$$

and similarly

$$Var(Y) = E(Y^2) - E(Y)^2 = n\,p_2\,(1-p_2).$$

This completes the proof of the theorem.

The following results are needed for the next theorem and they can be established using binomial theorem discussed in chapter 1. For any real numbers $a$ and $b$, we have

$$\sum_{y=0}^{m} y\binom{m}{y}a^y\,b^{m-y} = m\,a\,(a+b)^{m-1} \tag{11.1}$$

and

$$\sum_{y=0}^{m} y^2\binom{m}{y}a^y\,b^{m-y} = m\,a\,\left(ma+b\right)(a+b)^{m-2} \tag{11.2}$$

where $m$ is a positive integer.

**Example 11.3.** If $X$ equals the number of ones and $Y$ equals the number of twos and threes when a pair of fair dice are rolled, then what is the correlation coefficient of $X$ and $Y$?

**Answer:** The joint density of $X$ and $Y$ is bivariate binomial and is given by

$$f(x,y) = \frac{2!}{x!\,y!\,(2-x-y)!}\left(\frac{1}{6}\right)^x\left(\frac{2}{6}\right)^y\left(\frac{3}{6}\right)^{2-x-y}, \qquad 0 \le x+y \le 2,$$

where $x$ and $y$ are nonnegative integers. By Theorem 11.3, we have

$$Var(X) = n\,p_1\,(1-p_1) = 2\;\frac{1}{6}\;\left(1-\frac{1}{6}\right) = \frac{10}{36},$$

$$Var(Y) = n\,p_2\,(1-p_2) = 2\;\frac{2}{6}\;\left(1-\frac{2}{6}\right) = \frac{16}{36},$$

and

$$Cov(X,Y) = -n\,p_1\,p_2 = -2\;\frac{1}{6}\;\frac{2}{6} = -\frac{4}{36}.$$

Therefore

$$\begin{aligned} Corr(X,Y) &= \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} \\ &= -\frac{4}{4\sqrt{10}} \\ &= -0.3162. \end{aligned}$$

The next theorem presents some information regarding the conditional distributions $f(x/y)$ and $f(y/x)$.

**Theorem 11.4.** Let $(X,Y) \sim BIN\,(n,p_1,p_2)$, where $n$, $p_1$ and $p_2$ are parameters. Then the conditional distributions $f(y/x)$ and $f(x/y)$ are also binomial and

$$\begin{aligned} E(Y/x) &= \frac{p_2\,(n-x)}{1-p_1} \\ E(X/y) &= \frac{p_1\,(n-y)}{1-p_2} \\ Var(Y/x) &= \frac{p_2\,(1-p_1-p_2)\,(n-x)}{(1-p_1)^2} \\ Var(X/y) &= \frac{p_1\,(1-p_1-p_2)\,(n-y)}{(1-p_2)^2}. \end{aligned}$$

**Proof:** Since $f(y/x) = \frac{f(x,y)}{f_1(x)}$, first we find the marginal density of $X$. The marginal density $f_1(x)$ of $X$ is given by

$$\begin{aligned} f_1(x) &= \sum_{y=0}^{n-x} \frac{n!}{x!\,y!\,(n-x-y)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^{n-x-y} \\ &= \frac{n!\,p_1^x}{x!\,(n-x)!} \sum_{y=0}^{n-x} \frac{(n-x)!}{y!\,(n-x-y)!}\, p_2^y\, (1-p_1-p_2)^{n-x-y} \\ &= \binom{n}{x} p_1^x\, (1-p_1-p_2+p_2)^{n-x} \qquad \text{(by binomial theorem)} \\ &= \binom{n}{x} p_1^x\, (1-p_1)^{n-x}\,. \end{aligned}$$

In order to compute the conditional expectations, we need the conditional densities of $f(x,y)$. The conditional density of $Y$ given $X = x$ is

$$\begin{aligned} f(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{f(x,y)}{\binom{n}{x} p_1^x\, (1-p_1)^{n-x}} \\ &= \frac{(n-x)!}{(n-x-y)!\,y!}\, p_2^y\, (1-p_1-p_2)^{n-x-y}\, (1-p_1)^{x-n} \\ &= (1-p_1)^{x-n} \binom{n-x}{y} p_2^y\, (1-p_1-p_2)^{n-x-y}\,. \end{aligned}$$

Hence the conditional expectation of $Y$ given the event $X = x$ is

$$\begin{aligned} E\,(Y/x) &= \sum_{y=0}^{n-x} y\,(1-p_1)^{x-n} \binom{n-x}{y} p_2^y\, (1-p_1-p_2)^{n-x-y} \\ &= (1-p_1)^{x-n} \sum_{y=0}^{n-x} y \binom{n-x}{y} p_2^y\, (1-p_1-p_2)^{n-x-y} \\ &= (1-p_1)^{x-n}\, p_2\,(n-x)\,(1-p_1)^{n-x-1} \\ &= \frac{p_2\,(n-x)}{1-p_1}. \end{aligned}$$

Next, we find the conditional variance of $Y$ given event $X = x$. For this

we need the conditional expectation $E\left(Y^2/x\right)$, which is given by

$$\begin{aligned}
E\left(Y^2/x\right) &= \sum_{y=0}^{n-x} y^2 f(x,y) \\
&= \sum_{y=0}^{n-x} y^2 \left(1-p_1\right)^{x-n} \binom{n-x}{y} p_2^y \left(1-p_1-p_2\right)^{n-x-y} \\
&= \left(1-p_1\right)^{x-n} \sum_{y=0}^{n-x} y^2 \binom{n-x}{y} p_2^y \left(1-p_1-p_2\right)^{n-x-y} \\
&= \left(1-p_1\right)^{x-n} p_2 \left(n-x\right) \left(1-p_1\right)^{n-x-2} \left[(n-x)p_2+1-p_1-p_2\right] \\
&= \frac{p_2 \left(n-x\right)[(n-x)p_2+1-p_1-p_2]}{\left(1-p_1\right)^2}.
\end{aligned}$$

Hence, the conditional variance of $Y$ given $X=x$ is

$$\begin{aligned}
Var(Y/x) &= E\left(Y^2/x\right) - E(Y/x)^2 \\
&= \frac{p_2 \left(n-x\right) \left[(n-x)p_2+1-p_1-p_2\right]}{\left(1-p_1\right)^2} - \left(\frac{p_2 \left(n-x\right)}{1-p_1}\right)^2 \\
&= \frac{p_2 \left(1-p_1-p_2\right) \left(n-x\right)}{\left(1-p_1\right)^2}.
\end{aligned}$$

Similarly, one can establish

$$E(X/y) = \frac{p_1(n-y)}{1-p_2} \qquad \text{and} \qquad Var(X/y) = \frac{p_1 \left(1-p_1-p_2\right) \left(n-y\right)}{\left(1-p_2\right)^2}.$$

This completes the proof of the theorem.

Note that $f(y/x)$ in the above theorem is a univariate binomial probability density function. To see this observe that

$$\begin{aligned}
&\left(1-p_1\right)^{x-n} \binom{n-x}{y} p_2^y \left(1-p_1-p_2\right)^{n-x-y} \\
&\qquad = \binom{n-x}{y} \left(\frac{p_2}{1-p_1}\right)^y \left(1-\frac{p_2}{1-p_1}\right)^{n-x-y}.
\end{aligned}$$

Hence, $f(y/x)$ is a probability density function of a binomial random variable with parameters $n-x$ and $\frac{p_2}{1-p_1}$.

The marginal density $f_2(y)$ of $Y$ can be obtained similarly as

$$f_2(y) = \binom{n}{y} p_2^y \left(1-p_2\right)^{n-y},$$

where $y = 0, 1, ..., n$. The form of these densities show that the marginals of bivariate binomial distribution are again binomial.

**Example 11.4.** Let $W$ equal the weight of soap in a 1-kilogram box that is distributed in India. Suppose $P(W < 1) = 0.02$ and $P(W > 1.072) = 0.08$. Call a box of soap light, good, or heavy depending on whether $W < 1$, $1 \leq W \leq 1.072$, or $W > 1.072$, respectively. In a random sample of 50 boxes, let $X$ equal the number of light boxes and $Y$ the number of good boxes. What are the regression and scedastic curves of $Y$ on $X$?

**Answer:** The joint probability density function of $X$ and $Y$ is given by

$$f(x,y) = \frac{50!}{x!\,y!\,(50-x-y)!}\,p_1^x\,p_2^y\,(1-p_1-p_2)^{50-x-y}, \qquad 0 \leq x+y \leq 50,$$

where $x$ and $y$ are nonnegative integers. Hence, $(X,Y) \sim BIN(n, p_1, p_2)$, where $n = 50$, $p_1 = 0.02$ and $p_2 = 0.90$. The regression curve of $Y$ on $X$ is given by

$$\begin{aligned} E(Y/x) &= \frac{p_2\,(n-x)}{1-p_1} \\ &= \frac{0.9\,(50-x)}{1-0.02} \\ &= \frac{45}{49}\,(50-x). \end{aligned}$$

The scedastic curve of $Y$ on $X$ is the conditional variance of $Y$ given $X = x$ and it equal to

$$\begin{aligned} Var(Y/x) &= \frac{p_2\,(1-p_1-p_2)\,(n-x)}{(1-p_1)^2} \\ &= \frac{0.9\,0.08\,(50-x)}{(1-0.02)^2} \\ &= \frac{180}{2401}\,(50-x). \end{aligned}$$

Note that if $n = 1$, then bivariate binomial distribution reduces to bivariate Bernoulli distribution.

### 11.3. Bivariate Geometric Distribution

Recall that if the random variable $X$ denotes the trial number on which first success occurs, then $X$ is univariate geometric. The probability density function of an univariate geometric variable is

$$f(x) = p^{x-1}\,(1-p), \qquad x = 1, 2, 3, ..., \infty,$$

where $p$ is the probability of failure in a single Bernoulli trial. This univariate geometric distribution can be generalized to the bivariate case. Guldberg (1934) introduced the bivariate geometric distribution and Lundberg (1940) first used it in connection with problems of accident proneness. This distribution has found many applications in various statistical methods.

**Definition 11.3.** A discrete bivariate random variable $(X, Y)$ is said to have the bivariate geometric distribution with parameters $p_1$ and $p_2$ if its joint probability density is of the form

$$f(x,y) = \begin{cases} \frac{(x+y)!}{x!\,y!}\, p_1^x\, p_2^y\, (1-p_1-p_2)\,, & \text{if } x, y = 0, 1, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p_1,\, p_2,\, p_1 + p_2 < 1$. We denote a bivariate geometric random variable by writing $(X, Y) \sim GEO\,(p_1, p_2)$.

**Example 11.5.** Motor vehicles arriving at an intersection can turn right or left or continue straight ahead. In a study of traffic patterns at this intersection over a long period of time, engineers have noted that 40 percents of the motor vehicles turn left, 25 percents turn right, and the remainder continue straight ahead. For the next ten cars entering the intersection, what is the probability that 5 cars will turn left, 4 cars will turn right, and the last car will go straight ahead?

**Answer:** Let $X$ denote the number of cars turning left and $Y$ denote the number of cars turning right. Since, the last car will go straight ahead, the joint distribution of $X$ and $Y$ is geometric with parameters $p_1 = 0.4$, $p_2 = 0.25$ and $p_3 = 1 - p_1 - p_2 = 0.35$. For the next ten cars entering the intersection, the probability that 5 cars will turn left, 4 cars will turn right, and the last car will go straight ahead is given by

$$\begin{aligned} P(X = 5, Y = 4) &= f(5,4) \\ &= \frac{(x+y)!}{x!\,y!}\, p_1^x\, p_2^y\, (1-p_1-p_2) \\ &= \frac{(5+4)!}{5!\,4!}\, (0.4)^5\, (0.25)^4\, (1-0.4-0.25) \\ &= \frac{9!}{5!\,4!}\, (0.4)^5\, (0.25)^4\, (0.35) \\ &= 0.00677. \end{aligned}$$

The following technical result is essential for proving the following theorem. If $a$ and $b$ are positive real numbers with $0 < a + b < 1$, then

$$\sum_{x=0}^{\infty}\sum_{y=0}^{\infty}\frac{(x+y)!}{x!\,y!}\,a^x\,b^y = \frac{1}{1-a-b}. \tag{11.3}$$

In the following theorem, we present the expected values and the variances of $X$ and $Y$, the covariance between $X$ and $Y$, and the moment generating function.

**Theorem 11.5.** Let $(X, Y) \sim GEO\,(p_1, p_2)$, where $p_1$ and $p_2$ are parameters. Then

$$\begin{aligned}
E(X) &= \frac{p_1}{1-p_1-p_2}\\
E(Y) &= \frac{p_2}{1-p_1-p_2}\\
Var(X) &= \frac{p_1\,(1-p_2)}{(1-p_1-p_2)^2}\\
Var(Y) &= \frac{p_2\,(1-p_1)}{(1-p_1-p_2)^2}\\
Cov(X,Y) &= \frac{p_1\,p_2}{(1-p_1-p_2)^2}\\
M(s,t) &= \frac{1-p_1-p_2}{1-p_1e^s-p_2e^t}.
\end{aligned}$$

**Proof:** We only find the joint moment generating function $M(s, t)$ of $X$ and $Y$ and leave proof of the rests to the reader of this book. The joint moment generating function $M(s, t)$ is given by

$$\begin{aligned}
M(s,t) &= E\left(e^{sX+tY}\right)\\
&= \sum_{x=0}^{n}\sum_{y=0}^{n} e^{sx+ty} f(x,y)\\
&= \sum_{x=0}^{n}\sum_{y=0}^{n} e^{sx+ty}\,\frac{(x+y)!}{x!\,y!}\,p_1^x\,p_2^y\,(1-p_1-p_2)\\
&= (1-p_1-p_2)\sum_{x=0}^{n}\sum_{y=0}^{n}\frac{(x+y)!}{x!\,y!}\,(p_1e^s)^x\left(p_2e^t\right)^y\\
&= \frac{(1-p_1-p_2)}{1-p_1e^s-p_2e^t} \qquad \text{(by (11.3) ).}
\end{aligned}$$

The following results are needed for the next theorem. Let $a$ be a positive real number less than one. Then

$$\sum_{y=0}^{\infty} \frac{(x+y)!}{x!\,y!}\, a^y = \frac{1}{(1-a)^{x+1}}, \tag{11.4}$$

$$\sum_{y=0}^{\infty} \frac{(x+y)!}{x!\,y!}\, y\, a^y = \frac{a(1+x)}{(1-a)^{x+2}}, \tag{11.5}$$

and

$$\sum_{y=0}^{\infty} \frac{(x+y)!}{x!\,y!}\, y^2\, a^y = \frac{a(1+x)}{(1-a)^{x+3}}\,[a(x+1)+1]. \tag{11.6}$$

The next theorem presents some information regarding the conditional densities $f(x/y)$ and $f(y/x)$.

**Theorem 11.6.** Let $(X,Y) \sim GEO\,(p_1,p_2)$, where $p_1$ and $p_2$ are parameters. Then the conditional distributions $f(y/x)$ and $f(x/y)$ are also geometrical and

$$\begin{aligned} E(Y/x) &= \frac{p_2\,(1+x)}{1-p_2} \\ E(X/y) &= \frac{p_1\,(1+y)}{1-p_1} \\ Var(Y/x) &= \frac{p_2\,(1+x)}{(1-p_2)^2} \\ Var(X/y) &= \frac{p_1\,(1+y)}{(1-p_1)^2}. \end{aligned}$$

**Proof:** Again, as before, we first find the conditional probability density of $Y$ given the event $X = x$. The marginal density $f_1(x)$ is given by

$$\begin{aligned} f_1(x) &= \sum_{y=0}^{\infty} f(x,y) \\ &= \sum_{y=0}^{\infty} \frac{(x+y)!}{x!\,y!}\, p_1^x\, p_2^y\, (1-p_1-p_2) \\ &= (1-p_1-p_2)\, p_1^x \sum_{y=0}^{\infty} \frac{(x+y)!}{x!\,y!}\, p_2^y \\ &= \frac{(1-p_1-p_2)\, p_1^x}{(1-p_2)^{x+1}} \qquad \text{(by (11.4) ).} \end{aligned}$$

Therefore the conditional density of $Y$ given the event $X = x$ is

$$f(y/x) = \frac{f(x,y)}{f_1(x)} = \frac{(x+y)!}{x!\,y!}\; p_2^y\,(1-p_2)^{x+1}.$$

The conditional expectation of $Y$ given $X = x$ is

$$\begin{aligned} E(Y/x) &= \sum_{y=0}^{\infty} y\, f(y/x) \\ &= \sum_{y=0}^{\infty} y\, \frac{(x+y)!}{x!\,y!}\; p_2^y\,(1-p_2)^{x+1} \\ &= \frac{p_2\,(1+x)}{(1-p_2)} \qquad \text{(by (11.5) ).} \end{aligned}$$

Similarly, one can show that

$$E(X/y) = \frac{p_1\,(1+y)}{(1-p_1)}.$$

To compute the conditional variance of $Y$ given the event that $X = x$, first we have to find $E\left(Y^2/x\right)$, which is given by

$$\begin{aligned} E\left(Y^2/x\right) &= \sum_{y=0}^{\infty} y^2\, f(y/x) \\ &= \sum_{y=0}^{\infty} y^2\, \frac{(x+y)!}{x!\,y!}\; p_2^y\,(1-p_2)^{x+1} \\ &= \frac{p_2\,(1+x)}{(1-p_2)^2}\,[p_2\,(1+x)+1] \qquad \text{(by (11.6) ).} \end{aligned}$$

Therefore

$$\begin{aligned} Var\left(Y^2/x\right) &= E\left(Y^2/x\right) - E(Y/x)^2 \\ &= \frac{p_2\,(1+x)}{(1-p_2)^2}\,[(p_2\,(1+x)+1] - \left(\frac{p_2\,(1+x)}{1-p_2}\right)^2 \\ &= \frac{p_2\,(1+x)}{(1-p_2)^2}. \end{aligned}$$

The rest of the moments can be determined in a similar manner. The proof of the theorem is now complete.

### 11.4. Bivariate Negative Binomial Distribution

The univariate negative binomial distribution can be generalized to the bivariate case. Guldberg (1934) introduced this distribution and Lundberg (1940) first used it in connection with problems of accident proneness. Arbous and Kerrich (1951) arrived at this distribution by mixing parameters of the bivariate Poisson distribution.

**Definition 11.4.** A discrete bivariate random variable $(X, Y)$ is said to have the bivariate negative binomial distribution with parameters $k, p_1$ and $p_2$ if its joint probability density is of the form

$$f(x,y) = \begin{cases} \frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^k, & \text{if } x, y = 0, 1, ..., \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p_1, p_2, p_1 + p_2 < 1$ and $k$ is a nonzero positive integer. We denote a bivariate negative binomial random variable by writing $(X, Y) \sim NBIN(k, p_1, p_2)$.

**Example 11.6.** An experiment consists of selecting a marble at random and with replacement from a box containing 10 white marbles, 15 black marbles and 5 green marbles. What is the probability that it takes exactly 11 trials to get 5 white, 3 black and the third green marbles at the $11^{\text{th}}$ trial?

**Answer:** Let $X$ denote the number of white marbles and $Y$ denote the number of black marbles. The joint distribution of $X$ and $Y$ is bivariate negative binomial with parameters $p_1 = \frac{1}{3}$, $p_2 = \frac{1}{2}$, and $k = 3$. Hence the probability that it takes exactly 11 trials to get 5 white, 3 black and the third green marbles at the $11^{\text{th}}$ trial is

$$\begin{aligned} P(X = 5, Y = 3) &= f(5,3) \\ &= \frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\, p_1^x\, p_2^y\, (1-p_1-p_2)^k \\ &= \frac{(5+3+3-1)!}{5!\,3!\,(3-1)!}\, (0.33)^5\, (0.5)^3\, (1-0.33-0.5)^3 \\ &= \frac{10!}{5!\,3!\,2!}\, (0.33)^5\, (0.5)^3\, (0.17)^3 \\ &= 0.0000503. \end{aligned}$$

The negative binomial theorem which was treated in chapter 5 can be generalized to

$$\sum_{x=0}^{\infty} \sum_{y=0}^{\infty} \frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\, p_1^x\, p_2^y = \frac{1}{(1-p_1-p_2)^k}. \tag{11.7}$$

In the following theorem, we present the expected values and the variances of $X$ and $Y$, the covariance between $X$ and $Y$, and the moment generating function.

**Theorem 11.7.** Let $(X, Y) \sim NBIN(k, p_1, p_2)$, where $k$, $p_1$ and $p_2$ are parameters. Then

$$\begin{aligned}
E(X) &= \frac{k\,p_1}{1-p_1-p_2} \\
E(Y) &= \frac{k\,p_2}{1-p_1-p_2} \\
Var(X) &= \frac{k\,p_1\,(1-p_2)}{(1-p_1-p_2)^2} \\
Var(Y) &= \frac{k\,p_2\,(1-p_1)}{(1-p_1-p_2)^2} \\
Cov(X,Y) &= \frac{k\,p_1\,p_2}{(1-p_1-p_2)^2} \\
M(s,t) &= \frac{(1-p_1-p_2)^k}{\left(1-p_1e^s-p_2e^t\right)^k}.
\end{aligned}$$

**Proof:** We only find the joint moment generating function $M(s,t)$ of the random variables $X$ and $Y$ and leave the rests to the reader. The joint moment generating function is given by

$$\begin{aligned}
M(s,t) &= E\left(e^{sX+tY}\right) \\
&= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} e^{sx+ty}\,f(x,y) \\
&= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} e^{sx+ty}\,\frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,p_1^x\,p_2^y\,(1-p_1-p_2)^k \\
&= (1-p_1-p_2)^k \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} \frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,\left(e^s\,p_1\right)^x\,\left(e^t\,p_2\right)^y \\
&= \frac{(1-p_1-p_2)^k}{\left(1-p_1e^s-p_2e^t\right)^k} \qquad \text{(by (11.7))}.
\end{aligned}$$

This completes the proof of the theorem.

To establish the next theorem, we need the following two results. If $a$ is a positive real constant in the interval $(0,1)$, then

$$\sum_{y=0}^{\infty} \frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,a^y = \frac{1\,(x+k)}{(1-a)^{x+k}}, \tag{11.8}$$

$$\sum_{y=0}^{\infty} y \, \frac{(x+y+k-1)!}{x!\, y!\, (k-1)!} \, a^y = \frac{a\,(x+k)}{(1-a)^{x+k+1}}, \tag{11.9}$$

and

$$\sum_{y=0}^{\infty} y^2 \, \frac{(x+y+k-1)!}{x!\, y!\, (k-1)!} \, a^y = \frac{a\,(x+k)}{(1-a)^{x+k+2}} \left[1 + (x+k)a\right]. \tag{11.10}$$

The next theorem presents some information regarding the conditional densities $f(x/y)$ and $f(y/x)$.

**Theorem 11.8.** Let $(X, Y) \sim NBIN\,(k, p_1, p_2)$, where $p_1$ and $p_2$ are parameters. Then the conditional densities $f(y/x)$ and $f(x/y)$ are also negative binomial and

$$\begin{aligned} E(Y/x) &= \frac{p_2\,(k+x)}{1-p_2} \\ E(X/y) &= \frac{p_1\,(k+y)}{1-p_1} \\ Var(Y/x) &= \frac{p_2\,(k+x)}{(1-p_2)^2} \\ Var(X/y) &= \frac{p_1\,(k+y)}{(1-p_1)^2}. \end{aligned}$$

**Proof:** First, we find the marginal density of $X$. The marginal density $f_1(x)$ is given by

$$\begin{aligned} f_1(x) &= \sum_{y=0}^{\infty} f(x,y) \\ &= \sum_{y=0}^{\infty} \frac{(x+y+k-1)!}{x!\, y!\, (k-1)!} \, p_1^x \, p_2^y \\ &= (1-p_1-p_2)^k \; p_1^x \, \frac{(x+y+k-1)!}{x!\, y!\, (k-1)!} \, p_2^y \\ &= (1-p_1-p_2)^k \; p_1^x \, \frac{1}{(1-p_2)^{x+k}} \qquad \text{(by (11.8))}. \end{aligned}$$

The conditional density of $Y$ given the event $X = x$ is

$$\begin{aligned} f(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{(x+y+k-1)!}{x!\, y!\, (k-1)!} \, p_2^y \, (1-p_2)^{x+k}. \end{aligned}$$

The conditional expectation $E(Y/x)$ is given by

$$\begin{aligned} E(Y/x) &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} y\,\frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,p_2^y\,(1-p_2)^{x+k} \\ &= (1-p_2)^{x+k}\sum_{x=0}^{\infty}\sum_{y=0}^{\infty} y\,\frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,p_2^y \\ &= (1-p_2)^{x+k}\,\frac{p_2\,(x+k)}{(1-p_2)^{x+k+1}} \qquad \text{(by (11.9))} \\ &= \frac{p_2\,(x+k)}{(1-p_2)}. \end{aligned}$$

The conditional expectation $E\left(Y^2/x\right)$ can be computed as follows

$$\begin{aligned} E\left(Y^2/x\right) &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} y^2\,\frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,p_2^y\,(1-p_2)^{x+k} \\ &= (1-p_2)^{x+k}\sum_{x=0}^{\infty}\sum_{y=0}^{\infty} y^2\,\frac{(x+y+k-1)!}{x!\,y!\,(k-1)!}\,p_2^y \\ &= (1-p_2)^{x+k}\,\frac{p_2\,(x+k)}{(1-p_2)^{x+k+2}}\,[1+(x+k)\,p_2] \qquad \text{(by (11.10))} \\ &= \frac{p_2\,(x+k)}{(1-p_2)^2}[1+(x+k)\,p_2]. \end{aligned}$$

The conditional variance of $Y$ given $X=x$ is

$$\begin{aligned} Var\,(Y/x) &= E\left(Y^2/x\right) - E\,(Y/x)^2 \\ &= \frac{p_2\,(x+k)}{(1-p_2)^2}[1+(x+k)\,p_2] - \left(\frac{p_2\,(x+k)}{(1-p_2)}\right)^2 \\ &= \frac{p_2\,(x+k)}{(1-p_2)^2}. \end{aligned}$$

The conditional expected value $E(X/y)$ and conditional variance $Var(X/y)$ can be computed in a similar way. This completes the proof.

Note that if $k=1$, then bivariate negative binomial distribution reduces to bivariate geometric distribution.

### 11.5. Bivariate Hypergeometric Distribution

The univariate hypergeometric distribution can be generalized to the bivariate case. Isserlis (1914) introduced this distribution and Pearson (1924)

gave various properties of this distribution. Pearson also fitted this distribution to an observed data of the number of cards of a certain suit in two hands at whist.

**Definition 11.5.** A discrete bivariate random variable $(X, Y)$ is said to have the bivariate hypergeometric distribution with parameters $r, n_1, n_2, n_3$ if its joint probability distribution is of the form

$$f(x,y) = \begin{cases} \frac{\binom{n_1}{x}\binom{n_2}{y}\binom{n_3}{r-x-y}}{\binom{n_1+n_2+n_3}{r}}, & \text{if } x, y = 0, 1, ..., r \\ 0 & \text{otherwise,} \end{cases}$$

where $x \leq n_1$, $y \leq n_2$, $r - x - y \leq n_3$ and $r$ is a positive integer less than or equal to $n_1 + n_2 + n_3$. We denote a bivariate hypergeometric random variable by writing $(X, Y) \sim HYP(r, n_1, n_2, n_3)$.

**Example 11.7.** A panel of prospective jurors includes 6 african american, 4 asian american and 9 white american. If the selection is random, what is the probability that a jury will consists of 4 african american, 3 asian american and 5 white american?

**Answer:** Here $n_1 = 7$, $n_2 = 3$ and $n_3 = 9$ so that $n = 19$. A total of 12 jurors will be selected so that $r = 12$. In this example $x = 4$, $y = 3$ and $r - x - y = 5$. Hence the probability that a jury will consists of 4 african american, 3 asian american and 5 white american is

$$f(4,3) = \frac{\binom{7}{4}\binom{3}{3}\binom{9}{5}}{\binom{19}{12}} = \frac{4410}{50388} = 0.0875.$$

**Example 11.8.** Among 25 silver dollars struck in 1903 there are 15 from the Philadelphia mint, 7 from the New Orleans mint, and 3 from the San Francisco. If 5 of these silver dollars are picked at random, what is the probability of getting 4 from the Philadelphia mint and 1 from the New Orleans?

**Answer:** Here $n = 25$, $r = 5$ and $n_1 = 15$, $n_2 = 7$, $n_3 = 3$. The the probability of getting 4 from the Philadelphia mint and 1 from the New Orleans is

$$f(4,1) = \frac{\binom{15}{4}\binom{7}{1}\binom{3}{0}}{\binom{25}{5}} = \frac{9555}{53130} = 0.1798.$$

In the following theorem, we present the expected values and the variances of $X$ and $Y$, and the covariance between $X$ and $Y$.

**Theorem 11.9.** Let $(X, Y) \sim HYP(r, n_1, n_2, n_3)$, where $r$, $n_1$, $n_2$ and $n_3$ are parameters. Then

$$\begin{aligned}
E(X) &= \frac{r\, n_1}{n_1 + n_2 + n_3} \\
E(Y) &= \frac{r\, n_2}{n_1 + n_2 + n_3} \\
Var(X) &= \frac{r\, n_1\, (n_2 + n_3)}{(n_1 + n_2 + n_3)^2} \left( \frac{n_1 + n_2 + n_3 - r}{n_1 + n_2 + n_3 - 1} \right) \\
Var(Y) &= \frac{r\, n_2\, (n_1 + n_3)}{(n_1 + n_2 + n_3)^2} \left( \frac{n_1 + n_2 + n_3 - r}{n_1 + n_2 + n_3 - 1} \right) \\
Cov(X, Y) &= -\frac{r\, n_1\, n_2}{(n_1 + n_2 + n_3)^2} \left( \frac{n_1 + n_2 + n_3 - r}{n_1 + n_2 + n_3 - 1} \right).
\end{aligned}$$

**Proof:** We find only the mean and variance of $X$. The mean and variance of $Y$ can be found in a similar manner. The covariance of $X$ and $Y$ will be left to the reader as an exercise. To find the expected value of $X$, we need the marginal density $f_1(x)$ of $X$. The marginal of $X$ is given by

$$\begin{aligned}
f_1(x) &= \sum_{y=0}^{r-x} f(x, y) \\
&= \sum_{y=0}^{r-x} \frac{\binom{n_1}{x} \binom{n_2}{y} \binom{n_3}{r-x-y}}{\binom{n_1+n_2+n_3}{r}} \\
&= \frac{\binom{n_1}{x}}{\binom{n_1+n_2+n_3}{r}} \sum_{y=0}^{r-x} \binom{n_2}{y} \binom{n_3}{r-x-y} \\
&= \frac{\binom{n_1}{x}}{\binom{n_1+n_2+n_3}{r}} \binom{n_2+n_3}{r-x} \qquad \text{(by Theorem 1.3)}
\end{aligned}$$

This shows that $X \sim HYP(n_1,\, n_2 + n_3,\, r)$. Hence, by Theorem 5.7, we get

$$E(X) = \frac{r\, n_1}{n_1 + n_2 + n_3},$$

and

$$Var(X) = \frac{r\, n_1\, (n_2 + n_3)}{(n_1 + n_2 + n_3)^2} \left( \frac{n_1 + n_2 + n_3 - r}{n_1 + n_2 + n_3 - 1} \right).$$

Similarly, the random variable $Y \sim HYP(n_2,\, n_1 + n_3,\, r)$. Hence, again by Theorem 5.7, we get

$$E(Y) = \frac{r\, n_2}{n_1 + n_2 + n_3},$$

and

$$Var(Y) = \frac{r\, n_2\, (n_1 + n_3)}{(n_1 + n_2 + n_3)^2} \left( \frac{n_1 + n_2 + n_3 - r}{n_1 + n_2 + n_3 - 1} \right).$$

The next theorem presents some information regarding the conditional densities $f(x/y)$ and $f(y/x)$.

**Theorem 11.10.** Let $(X, Y) \sim HYP\,(r, n_1, n_2, n_3)$, where $r$, $n_1$, $n_2$ and $n_3$ are parameters. Then the conditional distributions $f(y/x)$ and $f(x/y)$ are also hypergeometric and

$$\begin{aligned}
E(Y/x) &= \frac{n_2\,(r-x)}{n_2 + n_3} \\
E(X/y) &= \frac{n_1\,(r-y)}{n_1 + n_3} \\
Var(Y/x) &= \frac{n_2 n_3}{n_2 + n_3 - 1} \left( \frac{n_1 + n_2 + n_3 - x}{n_2 + n_3} \right) \left( \frac{x - n_1}{n_2 + n_3} \right) \\
Var(X/y) &= \frac{n_1 n_3}{n_1 + n_3 - 1} \left( \frac{n_1 + n_2 + n_3 - y}{n_1 + n_3} \right) \left( \frac{y - n_2}{n_1 + n_3} \right).
\end{aligned}$$

**Proof:** To find $E(Y/x)$, we need the conditional density $f(y/x)$ of $Y$ given the event $X = x$. The conditional density $f(y/x)$ is given by

$$\begin{aligned}
f(y/x) &= \frac{f(x,y)}{f_1(x)} \\
&= \frac{\binom{n_2}{y}\binom{n_3}{r-x-y}}{\binom{n_2+n_3}{r-x}}.
\end{aligned}$$

Hence, the random variable $Y$ given $X = x$ is a hypergeometric random variable with parameters $n_2$, $n_3$, and $r - x$, that is

$$Y/x \sim HYP(n_2,\, n_3,\, r - x).$$

Hence, by Theorem 5.7, we get

$$E(Y/x) = \frac{n_2\,(r-x)}{n_2 + n_3}$$

and

$$Var(Y/x) = \frac{n_2 n_3}{n_2 + n_3 - 1} \left( \frac{n_1 + n_2 + n_3 - x}{n_2 + n_3} \right) \left( \frac{x - n_1}{n_2 + n_3} \right).$$

Similarly, one can find $E(X/y)$ and $Var(X/y)$. The proof of the theorem is now complete.

### 11.6. Bivariate Poisson Distribution

The univariate Poisson distribution can be generalized to the bivariate case. In 1934, Campbell, first derived this distribution. However, in 1944, Aitken gave the explicit formula for the bivariate Poisson distribution function. In 1964, Holgate also arrived at the bivariate Poisson distribution by deriving the joint distribution of $X = X_1 + X_3$ and $Y = X_2 + X_3$, where $X_1, X_2, X_3$ are independent Poisson random variables. Unlike the previous bivariate distributions, the conditional distributions of bivariate Poisson distribution are not Poisson. In fact, Seshadri and Patil (1964), indicated that no bivariate distribution exists having both marginal and conditional distributions of Poisson form.

**Definition 11.6.** A discrete bivariate random variable $(X, Y)$ is said to have the bivariate Poisson distribution with parameters $\lambda_1, \lambda_2, \lambda_3$ if its joint probability density is of the form

$$f(x,y) = \begin{cases} \frac{e^{(-\lambda_1-\lambda_2+\lambda_3)}\,(\lambda_1-\lambda_3)^x\,(\lambda_2-\lambda_3)^y}{x!\,y!}\,\psi(x,y) & \text{for } x,y = 0,1,...,\infty \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\psi(x,y) := \sum_{r=0}^{\min(x,y)} \frac{x^{(r)}\,y^{(r)}\,\lambda_3^r}{(\lambda_1-\lambda_3)^r\,(\lambda_2-\lambda_3)^r\,r!}$$

with

$$x^{(r)} := x(x-1)\cdots(x-r+1),$$

and $\lambda_1 > \lambda_3 > 0$, $\lambda_2 > \lambda_3 > 0$ are parameters. We denote a bivariate Poisson random variable by writing $(X, Y) \sim POI\,(\lambda_1, \lambda_2, \lambda_3)$.

In the following theorem, we present the expected values and the variances of $X$ and $Y$, the covariance between $X$ and $Y$ and the joint moment generating function.

**Theorem 11.11.** Let $(X, Y) \sim POI\,(\lambda_1, \lambda_2, \lambda_3)$, where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are

parameters. Then

$$\begin{aligned} E(X) &= \lambda_1 \\ E(Y) &= \lambda_2 \\ Var(X) &= \lambda_1 \\ Var(Y) &= \lambda_2 \\ Cov(X,Y) &= \lambda_3 \\ M(s,t) &= e^{-\lambda_1-\lambda_2-\lambda_3+\lambda_1 e^s+\lambda_2 e^t+\lambda_3 e^{s+t}}. \end{aligned}$$

The next theorem presents some special characteristics of the conditional densities $f(x/y)$ and $f(y/x)$.

**Theorem 11.12.** Let $(X,Y) \sim POI\,(\lambda_1,\lambda_2,\lambda_3)$, where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are parameters. Then

$$\begin{aligned} E(Y/x) &= \lambda_2 - \lambda_3 + \left(\frac{\lambda_3}{\lambda_1}\right) x \\ E(X/y) &= \lambda_1 - \lambda_3 + \left(\frac{\lambda_3}{\lambda_2}\right) y \\ Var(Y/x) &= \lambda_2 - \lambda_3 + \left(\frac{\lambda_3\ (\lambda_1-\lambda_3)}{\lambda_1^2}\right) x \\ Var(X/y) &= \lambda_1 - \lambda_3 + \left(\frac{\lambda_3\ (\lambda_2-\lambda_3)}{\lambda_2^2}\right) y. \end{aligned}$$

### 11.7. Review Exercises

**1.** A box contains 10 white marbles, 15 black marbles and 5 green marbles. If 10 marbles are selected at random and without replacement, what is the probability that 5 are white, 3 are black and 2 are green?

**2.** An urn contains 3 red balls, 2 green balls and 1 yellow ball. Three balls are selected at random and without replacement from the urn. What is the probability that at least 1 color is not drawn?

**3.** An urn contains 4 red balls, 8 green balls and 2 yellow balls. Five balls are randomly selected, without replacement, from the urn. What is the probability that 1 red ball, 2 green balls, and 2 yellow balls will be selected?

**4.** From a group of three Republicans, two Democrats, and one Independent, a committee of two people is to be randomly selected. If $X$ denotes the

number of Republicans and $Y$ the number of Democrats on the committee, then what is the variance of $Y$ given that $X = x$?

**5.** If $X$ equals the number of ones and $Y$ the number of twos and threes when a four fair dice are rolled, then what is the conditional variance of $X$ and $Y = 1$?

**6.** Motor vehicles arriving at an intersection can turn right or left or continue straight ahead. In a study of traffic patterns at this intersection over a long period of time, engineers have noted that 40 percents of the motor vehicles turn left, 25 percents turn right, and the remainder continue straight ahead. For the next five cars entering the intersection, what is the probability that at least one turn right? (Answer: 0.7627)

**7.** Among a large number of applicants for a certain position, 60 percents have only a high school education, 30 percents have some college training, and 10 percents have completed a college degree. If 5 applicants are randomly selected to be interviewed, what is the probability that at least one will have completed a college degree?

**8.** In a population of 200 students who have just completed a first course in calculus, 50 have earned $A$'s, 80 $B$'s and remaining earned $F$'s. A sample of size 25 is taken at random and without replacement from this population. What is the probability that 10 students have $A$'s, 12 students have $B$'s and 3 students have $F$'s ?

**9.** If $X$ equals the number of ones and $Y$ the number of twos and threes when a four fair dice are rolled, then what is the correlation coefficient of $X$ and $Y$?

**10.** If the joint moment generating function of $X$ and $Y$ is $M(s,t) = k\left(\frac{4}{7-e^s-2e^t}\right)^5$, then what is the value of the constant $k$? What is the correlation coefficient between $X$ and $Y$?

**11.** A die with 1 painted on three sides, 2 painted on two sides, and 3 painted on one side is rolled 15 times. What is the probability that we will get eight 1's, six 2's and a 3 on the last roll?

**12.** The output of a machine is graded excellent 80 percents of time, good 15 percents of time, and defective 5 percents of time. What is the probability that a random sample of size 15 has 10 excellent, 3 good, and 2 defective items?

**13.** An industrial product is graded by a machine excellent 80 percents of time, good 15 percents of time, and defective 5 percents of time. A random sample of 15 items is graded. What is the probability that machine will grade 10 excellent, 3 good, and 2 defective of which one being the last one graded?

**14.** If $(X, Y) \sim HYP(n_1, n_2, n_3, r)$, then what is the covariance of the random variables $X$ and $Y$?

# Chapter 12

# SOME SPECIAL CONTINUOUS BIVARIATE DISTRIBUTIONS

In this chapter, we study some well known continuous bivariate probability density functions. First, we present the natural extensions of univariate probability density functions that were treated in chapter 6. Then we present some other bivariate distributions that have been reported in the literature. The bivariate normal distribution has been treated in most textbooks because of its dominant role in the statistical theory. The other continuous bivariate distributions rarely treated in any textbooks. It is in this textbook, well known bivariate distributions have been treated for the first time. The monograph of K.V. Mardia gives an excellent exposition on various bivariate distributions. We begin this chapter with the bivariate uniform distribution.

## 12.1. Bivariate Uniform Distribution

In this section, we study Morgenstern bivariate uniform distribution in detail. The marginals of Morgenstern bivariate uniform distribution are uniform. In this sense, it is an extension of univariate uniform distribution. Other bivariate uniform distributions will be pointed out without any in depth treatment.

In 1956, Morgenstern introduced a one-parameter family of bivariate distributions whose univariate marginal are uniform distributions by the following formula

$$f(x,y) = f_1(x)\, f_2(y)\, (\, 1 + \alpha\, [2F_1(x) - 1]\, [2F_2(y) - 1]\, )\,,$$

where $\alpha \in [-1, 1]$ is a parameter. If one assumes The cdf $F_i(x) = x$ and the pdf $f_i(x) = 1$ $(i = 1, 2)$, then we arrive at the Morgenstern uniform distribution on the unit square. The joint probability density function $f(x, y)$ of the Morgenstern uniform distribution on the unit square is given by

$$f(x, y) = 1 + \alpha\,(2x - 1)\,(2y - 1), \qquad 0 < x, y \leq 1, \qquad -1 \leq \alpha \leq 1.$$

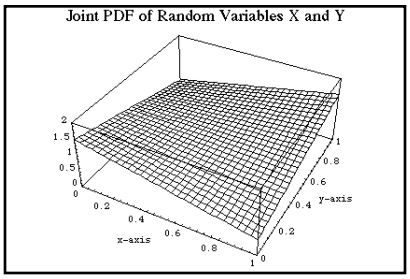
Next, we define the Morgenstern uniform distribution on an arbitrary rectangle $[a, b] \times [c, d]$.

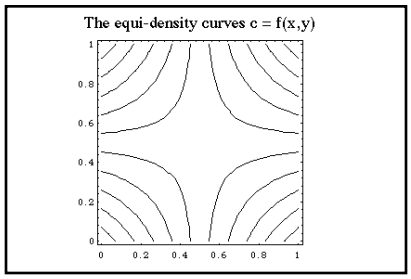
**Definition 12.1.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate uniform distribution on the rectangle $[a, b] \times [c, d]$ if its joint probability density function is of the form

$$f(x, y) = \begin{cases} \frac{1+\alpha\left(\frac{2x-2a}{b-a}-1\right)\left(\frac{2y-2c}{d-c}-1\right)}{(b-a)\,(d-c)} & \text{for } x \in [a, b]\ y \in [c, d] \\ 0 & \text{otherwise ,} \end{cases}$$

where $\alpha$ is an apriori chosen parameter in $[-1, 1]$. We denote a Morgenstern bivariate uniform random variable on a rectangle $[a, b] \times [c, d]$ by writing $(X, Y) \sim UNIF(a, b, c, d, \alpha)$.

The following figures show the graph and the equi-density curves of $f(x, y)$ on unit square with $\alpha = 0.5$.




In the following theorem, we present the expected values, the variances of the random variables $X$ and $Y$, and the covariance between $X$ and $Y$.

**Theorem 12.1.** Let $(X, Y) \sim UNIFM(a, b, c, d, \alpha)$, where $a, b, c, d$ and $\alpha$

are parameters. Then

$$\begin{aligned} E(X) &= \frac{b+a}{2} \\ E(Y) &= \frac{d+c}{2} \\ Var(X) &= \frac{(b-a)^2}{12} \\ Var(Y) &= \frac{(d-c)^2}{12} \\ Cov(X,Y) &= \frac{1}{36}\,\alpha\,(b-a)\,(d-c). \end{aligned}$$

**Proof:** First, we determine the marginal density of $X$ which is given by

$$\begin{aligned} f_1(x) &= \int_c^d f(x,y)\,dy \\ &= \int_c^d \frac{1+\alpha\,\left(\frac{2x-2a}{b-a}-1\right)\left(\frac{2y-2c}{d-c}-1\right)}{(b-a)\,(d-c)}\,dy \\ &= \frac{1}{b-a}. \end{aligned}$$

Thus, the marginal density of $X$ is uniform on the interval from $a$ to $b$. That is $X \sim UNIF(a,b)$. Hence by Theorem 6.1, we have

$$E(X) = \frac{b+a}{2} \qquad \text{and} \qquad Var(X) = \frac{(b-a)^2}{12}.$$

Similarly, one can show that $Y \sim UNIF(c,d)$ and therefore by Theorem 6.1

$$E(Y) = \frac{d+c}{2} \qquad \text{and} \qquad Var(Y) = \frac{(d-c)^2}{12}.$$

The product moment of $X$ and $Y$ is

$$\begin{aligned} E(XY) &= \int_a^b \int_c^d xy\,f(x,y)\,dx\,dy \\ &= \int_a^b \int_c^d xy\,\frac{1+\alpha\,\left(\frac{2x-2a}{b-a}-1\right)\left(\frac{2y-2c}{d-c}-1\right)}{(b-a)\,(d-c)}\,dx\,dy \\ &= \frac{1}{36}\,\alpha\,(b-a)\,(d-c) + \frac{1}{4}\,(b+a)\,(d+c). \end{aligned}$$

Thus, the covariance of $X$ and $Y$ is

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)\,E(Y) \\ &= \frac{1}{36}\,\alpha\,(b-a)\,(d-c) + \frac{1}{4}\,(b+a)\,(d+c) - \frac{1}{4}\,(b+a)\,(d+c) \\ &= \frac{1}{36}\,\alpha\,(b-a)\,(d-c). \end{aligned}$$

This completes the proof of the theorem.

In the next theorem, we states some information related to the conditional densities $f(y/x)$ and $f(x/y)$.

**Theorem 12.2.** Let $(X,Y) \sim UNIF(a,b,c,d,\alpha)$, where $a,b,c,d$ and $\alpha$ are parameters. Then

$$E(Y/x) = \frac{d+c}{2} + \frac{\alpha}{6\,(b-a)}\,\left(c^2+4cd+d^2\right)\,\left(\frac{2x-2a}{b-a}-1\right)$$

$$E(X/y) = \frac{b+a}{2} + \frac{\alpha}{6\,(b-a)}\,\left(a^2+4ab+b^2\right)\,\left(\frac{2y-2c}{d-c}-1\right)$$

$$Var(Y/x) = \frac{1}{36}\,\left(\frac{d-c}{b-a}\right)^2\,\left[\alpha^2(a+b)\,(4x-a-b) + 3(b-a)^2 - 4\alpha^2x^2\right]$$

$$Var(X/y) = \frac{1}{36}\,\left(\frac{b-a}{d-c}\right)^2\,\left[\alpha^2(c+d)\,(4y-c-d) + 3(d-c)^2 - 4\alpha^2y^2\right].$$

**Proof:** First, we determine the conditional density function $f(y/x)$. Recall that $f_1(x) = \frac{1}{b-a}$. Hence,

$$f(y/x) = \frac{1}{d-c}\,\left[1+\alpha\,\left(\frac{2x-2a}{b-a}-1\right)\,\left(\frac{2y-2c}{d-c}-1\right)\right].$$

The conditional expectation $E(Y/x)$ is given by

$$\begin{aligned}
E(Y/x) &= \int_c^d y\, f(y/x)\, dy \\
&= \frac{1}{d-c} \int_c^d y \left[1 + \alpha \left(\frac{2x-2a}{b-a} - 1\right) \left(\frac{2y-2c}{d-c} - 1\right)\right] dy \\
&= \frac{d+c}{2} + \frac{\alpha}{6\,(d-c)^2} \left(\frac{2x-2a}{b-a} - 1\right) \left[d^3 - c^3 + 3dc^2 - 3cd^2\right] \\
&= \frac{d+c}{2} + \frac{\alpha}{6\,(d-c)} \left(\frac{2x-2a}{b-a} - 1\right) \left[d^2 + 4dc + c^2\right].
\end{aligned}$$

Similarly, the conditional expectation $E\left(Y^2/x\right)$ is given by

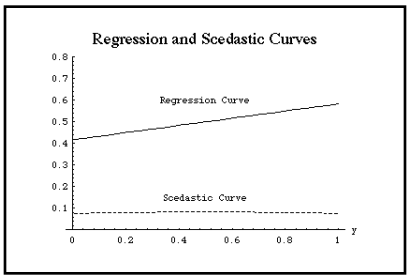
$$\begin{aligned}
E\left(Y^2/x\right) &= \int_c^d y^2\, f(y/x)\, dy \\
&= \frac{1}{d-c} \int_c^d y^2 \left[1 + \alpha \left(\frac{2x-2a}{b-a} - 1\right) \left(\frac{2y-2c}{d-c} - 1\right)\right] dy \\
&= \frac{1}{d-c} \left[\frac{d^2-c^2}{2} + \frac{\alpha}{d-c} \left(\frac{2x-2a}{b-a} - 1\right) \frac{1}{6} \left(d^2-c^2\right) (d-c)^2\right] \\
&= \frac{d+c}{2} + \frac{1}{6}\,\alpha \left(d^2-c^2\right) \left(\frac{2x-2a}{b-a} - 1\right) \\
&= \frac{d+c}{2} \left[1 + \frac{\alpha}{3}\,(d-c) \left(\frac{2x-2a}{b-a} - 1\right)\right].
\end{aligned}$$

Therefore, the conditional variance of $Y$ given the event $X = x$ is

$$\begin{aligned}
Var(Y/x) &= E\left(Y^2/x\right) - E(Y/x)^2 \\
&= \frac{1}{36} \left(\frac{d-c}{b-a}\right)^2 \left[\alpha^2(a+b)(4x-a-b) + 3(b-a)^2 - 4\alpha^2x^2\right].
\end{aligned}$$

The conditional expectation $E(X/y)$ and the conditional variance $Var(X/y)$ can be found in a similar manner. This completes the proof of the theorem.

The following figure illustrate the regression and scedastic curves of the Morgenstern uniform distribution function on unit square with $\alpha = 0.5$.




Next, we give a definition of another generalized bivariate uniform distribution.

**Definition 12.2.** Let $S \subset \mathbb{R}^2$ be a region in the Euclidean plane $\mathbb{R}^2$ with area $A$. The random variables $X$ and $Y$ is said to be bivariate uniform over $S$ if the joint density of $X$ and $Y$ is of the form

$$f(x,y) = \begin{cases} \frac{1}{A} & \text{for } (x,y) \in S \\ 0 & \text{otherwise .} \end{cases}$$

In 1965, Plackett constructed a class of bivariate distribution $F(x,y)$ for given marginals $F_1(x)$ and $F_2(y)$ as the square root of the equation

$$(\alpha - 1)\, F(x,y)^2 - \{1 + (\alpha - 1)\ [F_1(x) + F_2(y)]\ \}\ F(x,y) + \alpha\, F_1(x)\, F_2(y) = 0$$

(where $0 < \alpha < \infty$) which satisfies the Fréchet inequalities

$$\max\{F_1(x) + F_2(y) - 1,\ 0\} \le F(x,y) \le \min\{F_1(x), F_2(y)\}\,.$$

The class of bivariate joint density function constructed by Plackett is the following

$$f(x,y) = \alpha\, f_1(x)\, f_2(y)\, \frac{[(\alpha - 1)\ \{F_1(x) + F_2(y) - 2F_1(x)F_2(y)\} + 1]}{[S(x,y)^2 - 4\,\alpha\,(\alpha - 1)\, F_1(x)\, F_2(y)]^{\frac{3}{2}}},$$

where

$$S(x,y) = 1 + (\alpha - 1)\ (F_1(x) + F_2(y))\,.$$

If one takes $F_i(x) = x$ and $f_i(x) = 1$ (for $i = 1, 2$), then the joint density function constructed by Plackett reduces to

$$f(x,y) = \alpha\, \frac{[(\alpha - 1)\ \{x + y - 2xy\} + 1]}{[\{1 + (\alpha - 1)(x + y)\}^2 - 4\,\alpha\,(\alpha - 1)\, xy]^{\frac{3}{2}}},$$

where $0 \leq x, y \leq 1$, and $\alpha > 0$. But unfortunately this is not a bivariate density function since this bivariate density does not integrate to one. This fact was missed by both Plackett (1965) and Mardia (1967).

### 12.2. Bivariate Cauchy Distribution

Recall that univariate Cauchy probability distribution was defined in Chapter 3 as

$$f(x) = \frac{\theta}{\pi \left[\theta + (x-\alpha)^2\right]}, \qquad -\infty < x < \infty,$$

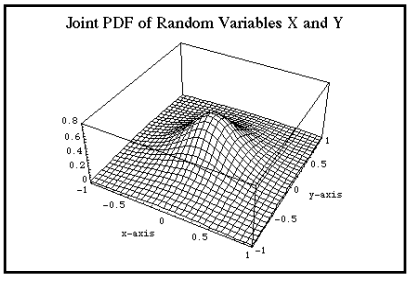
where $\alpha > 0$ and $\theta$ are real parameters. The parameter $\alpha$ is called the location parameter. In Chapter 4, we have pointed out that any random variable whose probability density function is Cauchy has no moments. This random variable is further, has no moment generating function. The Cauchy distribution is widely used for instructional purposes besides its statistical use. The main purpose of this section is to generalize univariate Cauchy distribution to bivariate case and study its various intrinsic properties. We define the bivariate Cauchy random variables by using the form of their joint probability density function.

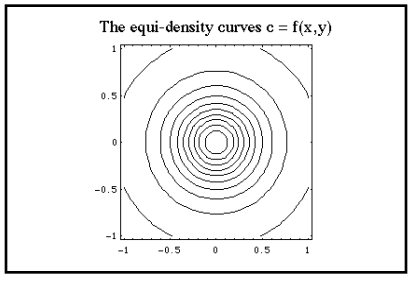
**Definition 12.3.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate Cauchy distribution if its joint probability density function is of the form

$$f(x,y) = \frac{\theta}{2\pi \left[\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,\right]^{\frac{3}{2}}}, \qquad -\infty < x, y < \infty,$$

where $\theta$ is a positive parameter and $\alpha$ and $\beta$ are location parameters. We denote a bivariate Cauchy random variable by writing $(X, Y) \sim CAU(\theta, \alpha, \beta)$.

The following figures show the graph and the equi-density curves of the Cauchy density function $f(x, y)$ with parameters $\alpha = 0 = \beta$ and $\theta = 0.5$.



The bivariate Cauchy distribution can be derived by considering the distribution of radio active particles emanating from a source that hit a two-dimensional screen. This distribution is a special case of the bivariate t-distribution which was first constructed by Karl Pearson in 1923.

The following theorem shows that if a bivariate random variable $(X, Y)$ is Cauchy, then it has no moments like the univariate Cauchy random variable. Further, for a bivariate Cauchy random variable $(X, Y)$, the covariance (and hence the correlation) between $X$ and $Y$ does not exist.

**Theorem 12.3.** Let $(X, Y) \sim CAU(\theta, \alpha, \beta)$, where $\theta > 0$, $\alpha$ and $\beta$ are parameters. Then the moments $E(X)$, $E(Y)$, $Var(X)$, $Var(Y)$, and $Cov(X, Y)$ do not exist.

**Proof:** In order to find the moments of $X$ and $Y$, we need their marginal distributions. First, we find the marginal of $X$ which is given by

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{\theta}{2\pi\ [\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,]^{\frac{3}{2}}}\, dy. \end{aligned}$$

To evaluate the above integral, we make a trigonometric substitution

$$y = \beta + \sqrt{[\theta^2 + (x-\alpha)^2]}\ \tan\psi.$$

Hence

$$dy = \sqrt{[\theta^2 + (x-\alpha)^2]}\, \sec^2\psi\, d\psi$$

and

$$\begin{aligned} &\left[\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,\right]^{\frac{3}{2}} \\ &\qquad = \left[\theta^2 + (x-\alpha)^2\right]^{\frac{3}{2}}\ \left(1 + \tan^2\psi\right)^{\frac{3}{2}} \\ &\qquad = \left[\theta^2 + (x-\alpha)^2\right]^{\frac{3}{2}}\ \sec^3\psi. \end{aligned}$$

Using these in the above integral, we get

$$\int_{-\infty}^{\infty} \frac{\theta}{2\pi\ [\,\theta^2+(x-\alpha)^2+(y-\beta)^2\,]^{\frac{3}{2}}}\,dy$$

$$= \frac{\theta}{2\pi}\ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\sqrt{[\theta^2+(x-\alpha)^2]}\,\sec^2\psi\,d\psi}{[\theta^2+(x-\alpha)^2]^{\frac{3}{2}}\ \sec^3\psi}$$

$$= \frac{\theta}{2\pi\ [\theta^2+(x-\alpha)^2]}\ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos\psi\,d\psi$$

$$= \frac{\theta}{\pi\ [\theta^2+(x-\alpha)^2]}.$$

Hence, the marginal of $X$ is a Cauchy distribution with parameters $\theta$ and $\alpha$. Thus, for the random variable $X$, the expected value $E(X)$ and the variance $Var(X)$ do not exist (see Example 4.2). In a similar manner, it can be shown that the marginal distribution of $Y$ is also Cauchy with parameters $\theta$ and $\beta$ and hence $E(Y)$ and $Var(Y)$ do not exist. Since

$$Cov(X,Y) = E(XY) - E(X)\,E(Y),$$

it easy to note that $Cov(X,Y)$ also does not exist. This completes the proof of the theorem.

The conditional distribution of $Y$ given the event $X=x$ is given by

$$f(y/x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{2}\ \frac{\theta^2+(x-\alpha)^2}{[\,\theta^2+(x-\alpha)^2+(y-\beta)^2\,]^{\frac{3}{2}}}.$$

Similarly, the conditional distribution of $X$ given the event $Y=y$ is

$$f(y/x) = \frac{1}{2}\ \frac{\theta^2+(y-\beta)^2}{[\,\theta^2+(x-\alpha)^2+(y-\beta)^2\,]^{\frac{3}{2}}}.$$

Next theorem states some properties of the conditional densities $f(y/x)$ and $f(x/y)$.

**Theorem 12.4.** Let $(X,Y) \sim CAU(\theta,\alpha,\beta)$, where $\theta > 0$, $\alpha$ and $\beta$ are parameters. Then the conditional expectations

$$E(Y/x) = \beta$$
$$E(X/y) = \alpha,$$

and the conditional variances $Var(Y/x)$ and $Var(X/y)$ do not exist.

**Proof:** First, we show that $E(Y/x)$ is $\beta$. The conditional expectation of $Y$ given the event $X = x$ can be computed as

$$\begin{aligned}
E(Y/x) &= \int_{-\infty}^{\infty} y\, f(y/x)\, dy \\
&= \int_{-\infty}^{\infty} y\, \frac{1}{2}\, \frac{\theta^2 + (x-\alpha)^2}{[\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,]^{\frac{3}{2}}}\, dy \\
&= \frac{1}{4}\, \left[\theta^2 + (x-\alpha)^2\right] \int_{-\infty}^{\infty} \frac{d\left(\theta^2 + (x-\alpha)^2 + (y-\beta)^2\right)}{[\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,]^{\frac{3}{2}}} \\
&\quad + \frac{\beta}{2}\, \left[\theta^2 + (x-\alpha)^2\right] \int_{-\infty}^{\infty} \frac{dy}{[\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,]^{\frac{3}{2}}} \\
&= \frac{1}{4}\, \left[\theta^2 + (x-\alpha)^2\right] \left[-\frac{2}{\sqrt{\theta^2 + (x-\alpha)^2 + (y-\beta)^2}}\right]_{-\infty}^{\infty} \\
&\quad + \frac{\beta}{2}\, \left[\theta^2 + (x-\alpha)^2\right] \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\cos\psi\, d\psi}{[\theta^2 + (x-\alpha)^2]} \\
&= 0 + \beta \\
&= \beta.
\end{aligned}$$

Similarly, it can be shown that $E(X/y) = \alpha$. Next, we show that the conditional variance of $Y$ given $X = x$ does not exist. To show this, we need $E\left(Y^2/x\right)$, which is given by

$$E\left(Y^2/x\right) = \int_{-\infty}^{\infty} y^2\, \frac{1}{2}\, \frac{\theta^2 + (x-\alpha)^2}{[\,\theta^2 + (x-\alpha)^2 + (y-\beta)^2\,]^{\frac{3}{2}}}\, dy.$$

The above integral does not exist and hence the conditional second moment of $Y$ given $X = x$ does not exist. As a consequence, the $Var(Y/x)$ also does not exist. Similarly, the variance of $X$ given the event $Y = y$ also does not exist. This completes the proof of the theorem.

### 12.3. Bivariate Gamma Distributions

In this section, we present three different bivariate gamma probability density functions and study some of their intrinsic properties.

**Definition 12.4.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate gamma distribution if its joint probability density function

is of the form

$$f(x,y) = \begin{cases} \frac{(xy)^{\frac{1}{2}(\alpha-1)}}{(1-\theta)\,\Gamma(\alpha)\,\theta^{\frac{1}{2}(\alpha-1)}}\, e^{-\frac{x+y}{1-\theta}}\, I_{\alpha-1}\left(\frac{2\sqrt{\theta\,xy}}{1-\theta}\right) & \text{if } 0 \le x,y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

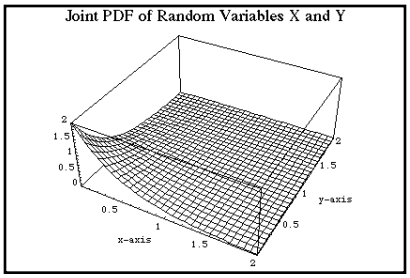
where $\theta \in [0,1)$ and $\alpha > 0$ are parameters, and

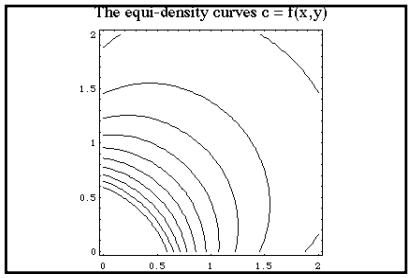
$$I_k(z) := \sum_{r=0}^{\infty} \frac{\left(\frac{1}{2}\,z\right)^{k+2r}}{r!\,\Gamma(k+r+1)}.$$

As usual, we denote this bivariate gamma random variable by writing $(X,Y) \sim GAMK(\alpha,\theta)$. The function $I_k(z)$ is called the modified Bessel function of the first kind of order $k$. In explicit form $f(x,y)$ is given by

$$f(x,y) = \begin{cases} \frac{1}{\theta^{\alpha-1}\,\Gamma(\alpha)}\, e^{-\frac{x+y}{1-\theta}} \sum_{k=0}^{\infty} \frac{(\theta\,x\,y)^{\alpha+k-1}}{k!\,\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}} & \text{for } 0 \le x,y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

The following figures show the graph of the joint density function $f(x,y)$ of a bivariate gamma random variable with parameters $\alpha = 1$ and $\theta = 0.5$ and the equi-density curves of $f(x,y)$.




In 1941, Kibble found this bivariate gamma density function. However, Wicksell in 1933 had constructed the characteristic function of this bivariate gamma density function without knowing the explicit form of this density function. If $\{\,(X_i,\,Y_i\,)\,|\,i = 1,2,...,n\}$ is a random sample from a bivariate normal distribution with zero means, then the bivariate random variable $(X,Y)$, where $X = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ and $Y = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$, has bivariate gamma distribution. This fact was established by Wicksell by finding the characteristic

function of $(X, Y)$. This bivariate gamma distribution has found applications in noise theory (see Rice (1944, 1945)).

The following theorem provides us some important characteristic of the bivariate gamma distribution of Kibble.

**Theorem 12.5.** Let the random variable $(X, Y) \sim GAMK(\alpha, \theta)$, where $0 < \alpha < \infty$ and $0 \leq \theta < 1$ are parameters. Then the marginals of $X$ and $Y$ are univariate gamma and

$$\begin{aligned} E(X) &= \alpha \\ E(Y) &= \alpha \\ Var(X) &= \alpha \\ Var(Y) &= \alpha \\ Cov(X, Y) &= \alpha\,\theta \\ M(s, t) &= \frac{1}{[(1-s)\,(1-t) - \theta\, s\, t]^{\alpha}}. \end{aligned}$$

**Proof:** First, we show that the marginal distribution of $X$ is univariate gamma with parameter $\alpha$ (and $\theta = 1$). The marginal density of $X$ is given by

$$\begin{aligned} f_1(x) &= \int_0^{\infty} f(x, y)\, dy \\ &= \int_0^{\infty} \frac{1}{\theta^{\alpha-1}\,\Gamma(\alpha)}\, e^{-\frac{x+y}{1-\theta}} \sum_{k=0}^{\infty} \frac{(\theta\, x\, y)^{\alpha+k-1}}{k!\,\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, dy \\ &= \sum_{k=0}^{\infty} \frac{1}{\theta^{\alpha-1}\,\Gamma(\alpha)}\, e^{-\frac{x}{1-\theta}}\, \frac{(\theta\, x)^{\alpha+k-1}}{k!\,\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}} \int_0^{\infty} y^{\alpha+k-1}\, e^{-\frac{y}{1-\theta}}\, dy \\ &= \sum_{k=0}^{\infty} \frac{1}{\theta^{\alpha-1}\,\Gamma(\alpha)}\, e^{-\frac{x}{1-\theta}}\, \frac{(\theta\, x)^{\alpha+k-1}}{k!\,\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, (1-\theta)^{\alpha+k}\,\Gamma(\alpha+k) \\ &= \sum_{k=0}^{\infty} \left(\frac{\theta}{1-\theta}\right)^{k} \frac{1}{k!\,\Gamma(\alpha)}\, x^{\alpha+k-1}\, e^{-\frac{x}{1-\theta}} \\ &= \frac{1}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{x\theta}{1-\theta}\right)^{k} \\ &= \frac{1}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\frac{x}{1-\theta}}\, e^{\frac{x\theta}{1-\theta}} \\ &= \frac{1}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-x}. \end{aligned}$$

Thus, the marginal distribution of $X$ is gamma with parameters $\alpha$ and $\theta = 1$. Therefore, by Theorem 6.3, we obtain

$$E(X) = \alpha, \qquad Var(X) = \alpha.$$

Similarly, we can show that the marginal density of $Y$ is gamma with parameters $\alpha$ and $\theta = 1$. Hence, we have

$$E(Y) = \alpha, \qquad Var(Y) = \alpha.$$

The moment generating function can be computed in a similar manner and we leave it to the reader. This completes the proof of the theorem.

The following results are needed for the next theorem. From calculus we know that

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}, \tag{12.1}$$

and the infinite series on the right converges for all $z \in \mathbb{R}$. Differentiating both sides of (12.1) and then multiplying the resulting expression by $z$, one obtains

$$ze^z = \sum_{k=0}^{\infty} k\, \frac{z^k}{k!}. \tag{12.2}$$

If one differentiates (12.2) again with respect to $z$ and multiply the resulting expression by $z$, then he/she will get

$$ze^z + z^2e^z = \sum_{k=0}^{\infty} k^2\, \frac{z^k}{k!}. \tag{12.3}$$

**Theorem 12.6.** Let the random variable $(X, Y) \sim GAMK(\alpha, \theta)$, where $0 < \alpha < \infty$ and $0 \leq \theta < 1$ are parameters. Then

$$\begin{aligned} E(Y/x) &= \theta\, x + (1-\theta)\, \alpha \\ E(X/y) &= \theta\, y + (1-\theta)\, \alpha \\ Var(Y/x) &= (1-\theta)\, [\, 2\theta\, x + (1-\theta)\, \alpha\, ] \\ Var(X/y) &= (1-\theta)\, [\, 2\theta\, y + (1-\theta)\, \alpha\, ]. \end{aligned}$$

**Proof:** First, we will find the conditional probability density function $Y$ given $X = x$, which is given by

$$\begin{aligned}
&f(y/x)\\
&= \frac{f(x,y)}{f_1(x)}\\
&= \frac{1}{\theta^{\alpha-1}\, x^{\alpha-1}\, e^{-x}}\, e^{-\frac{x+y}{1-\theta}} \sum_{k=0}^{\infty} \frac{(\theta\, x\, y)^{\alpha+k-1}}{k!\,\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\\
&= e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!}\, y^{\alpha+k-1}\, e^{-\frac{y}{1-\theta}}.
\end{aligned}$$

Next, we compute the conditional expectation of $Y$ given the event $X = x$. The conditional expectation $E(Y/x)$ is given by

$$\begin{aligned}
&E(Y/x)\\
&= \int_0^{\infty} y\, f(y/x)\, dy\\
&= \int_0^{\infty} y\, e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!}\, y^{\alpha+k-1}\, e^{-\frac{y}{1-\theta}}\, dy\\
&= e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!} \int_0^{\infty} y^{\alpha+k}\, e^{-\frac{y}{1-\theta}}\, dy\\
&= e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!}\, (1-\theta)^{\alpha+k+1}\, \Gamma(\alpha+k)\\
&= (1-\theta)\, e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} (\alpha+k)\, \frac{1}{k!} \left(\frac{\theta\, x}{1-\theta}\right)^k\\
&= (1-\theta)\, e^{x-\frac{x}{1-\theta}} \left[\alpha\, e^{\frac{\theta\, x}{1-\theta}} + \frac{\theta\, x}{1-\theta}\, e^{\frac{\theta\, x}{1-\theta}}\right] \qquad (by\ (12.1)\ and\ (12.2))\\
&= (1-\theta)\,\alpha + \theta\, x.
\end{aligned}$$

In order to determine the conditional variance of $Y$ given the event $X = x$, we need the conditional expectation of $Y^2$ given the event $X = x$. This

conditional expectation can be evaluated as follows:

$$
\begin{aligned}
&E(Y^2/x)\\
&= \int_0^\infty y^2\, f(y/x)\, dy\\
&= \int_0^\infty y^2\, e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!}\, y^{\alpha+k-1}\, e^{-\frac{y}{1-\theta}}\, dy\\
&= e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!} \int_0^\infty y^{\alpha+k+1}\, e^{-\frac{y}{1-\theta}}\, dy\\
&= e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(\alpha+k)\,(1-\theta)^{\alpha+2k}}\, \frac{(\theta\, x)^k}{k!}\, (1-\theta)^{\alpha+k+2}\, \Gamma(\alpha+k+2)\\
&= (1-\theta)^2\, e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} (\alpha+k+1)\,(\alpha+k)\, \frac{1}{k!} \left(\frac{\theta\, x}{1-\theta}\right)^k\\
&= (1-\theta)^2\, e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} (\alpha^2+2\alpha k+k^2+\alpha+k)\, \frac{1}{k!} \left(\frac{\theta\, x}{1-\theta}\right)^k\\
&= (1-\theta)^2 \left[\alpha^2+\alpha+(2\alpha+1)\,\frac{\theta\, x}{1-\theta}+\frac{\theta\, x}{1-\theta}+e^{x-\frac{x}{1-\theta}} \sum_{k=0}^{\infty} \frac{k^2}{k!} \left(\frac{\theta\, x}{1-\theta}\right)^k\right]\\
&= (1-\theta)^2 \left[\alpha^2+\alpha+(2\alpha+1)\,\frac{\theta\, x}{1-\theta}+\frac{\theta\, x}{1-\theta}+\left(\frac{\theta\, x}{1-\theta}\right)^2\right]\\
&= (\alpha^2+\alpha)\,(1-\theta)^2+2(\alpha+1)\,\theta\,(1-\theta)\,x+\theta^2\, x^2.
\end{aligned}
$$

The conditional variance of $Y$ given $X = x$ is

$$
\begin{aligned}
Var(Y/x) &= E(Y^2/x) - E(Y/x)^2\\
&= (\alpha^2+\alpha)\,(1-\theta)^2+2(\alpha+1)\,\theta\,(1-\theta)\,x+\theta^2\, x^2\\
&\quad - \left[(1-\theta)^2\,\alpha^2+\theta^2\, x^2+2\,\alpha\,\theta\,(1-\theta)\,x\right]\\
&= (1-\theta)\,\left[\alpha\,(1-\theta)+2\,\theta\, x\right].
\end{aligned}
$$

Since the density function $f(x, y)$ is symmetric, that is $f(x, y) = f(y, x)$, the conditional expectation $E(X/y)$ and the conditional variance $Var(X/y)$ can be obtained by interchanging $x$ with $y$ in the formulae of $E(Y/x)$ and $Var(Y/x)$. This completes the proof of the theorem.

In 1941, Cherian constructed a bivariate gamma distribution whose probability density function is given by

$$
f(x,y) = \begin{cases} \frac{e^{-(x+y)}}{\prod_{i=1}^{3}\Gamma(\alpha_i)} \int_0^{\min\{x,y\}} \frac{z^{\alpha_3}\,(x-z)^{\alpha_1}\,(y-z)^{\alpha_2}}{z\,(x-z)\,(y-z)}\, e^z\, dz & \text{if } 0 < x, y < \infty \\ 0 & \text{otherwise,} \end{cases}
$$

where $\alpha_1, \alpha_2, \alpha_3 \in (0, \infty)$ are parameters. If a bivariate random variable $(X, Y)$ has a Cherian bivariate gamma probability density function with parameters $\alpha_1, \alpha_2$ and $\alpha_3$, then we denote this by writing $(X, Y) \sim GAMC(\alpha_1, \alpha_2, \alpha_3)$.

It can be shown that the marginals of $f(x, y)$ are given by

$$f_1(x) = \begin{cases} \frac{1}{\Gamma(\alpha_1+\alpha_3)}\, x^{\alpha_1+\alpha_3-1}\, e^{-x} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_2(x) = \begin{cases} \frac{1}{\Gamma(\alpha_2+\alpha_3)}\, x^{\alpha_2+\alpha_3-1}\, e^{-y} & \text{if } 0 < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Hence, we have the following theorem.

**Theorem 12.7.** If $(X, Y) \sim GAMC(\alpha, \beta, \gamma)$, then

$$\begin{aligned} E(X) &= \alpha + \gamma \\ E(Y) &= \beta + \gamma \\ Var(X) &= \alpha + \gamma \\ Var(Y) &= \beta + \gamma \\ E(XY) &= \gamma + (\alpha + \gamma)(\beta + \gamma). \end{aligned}$$

The following theorem can be established by first computing the conditional probability density functions. We leave the proof of the following theorem to the reader.

**Theorem 12.8.** If $(X, Y) \sim GAMC(\alpha, \beta, \gamma)$, then

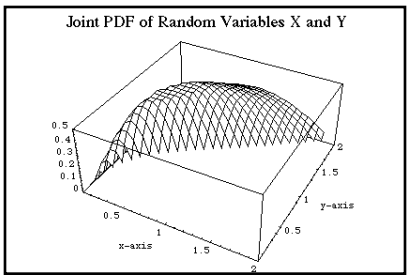
$$E(Y/x) = \beta + \frac{\gamma}{\alpha + \gamma}\, x \qquad \text{and} \qquad E(X/y) = \alpha + \frac{\gamma}{\beta + \gamma}\, y.$$

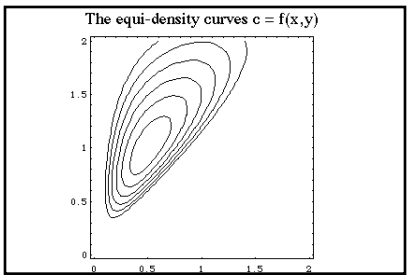
David and Fix (1961) have studied the rank correlation and regression for samples from this distribution. For an account of this bivariate gamma distribution the interested reader should refer to Moran (1967).

In 1934, McKay gave another bivariate gamma distribution whose probability density function is of the form

$$f(x, y) = \begin{cases} \frac{\theta^{\alpha+\beta}}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\, (y-x)^{\beta-1}\, e^{-\theta\, y} & \text{if } 0 < x < y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta, \alpha, \beta \in (0, \infty)$ are parameters. If the form of the joint density of the random variable $(X, Y)$ is similar to the density function of the bivariate gamma distribution of McKay, then we write $(X, Y) \sim GAMM(\theta, \alpha, \beta)$. The graph of probability density function $f(x, y)$ of the bivariate gamma distribution of McKay for $\theta = \alpha = \beta = 2$ is shown below. The other figure illustrates the equi-density curves of this joint density function $f(x, y)$.




It can shown that if $(X, Y) \sim GAMM(\theta, \alpha, \beta)$, then the marginal $f_1(x)$ of $X$ and the marginal $f_2(y)$ of $Y$ are given by

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

and

$$f_2(y) = \begin{cases} \frac{\theta^{\alpha+\beta}}{\Gamma(\alpha+\beta)}\, x^{\alpha+\beta-1}\, e^{-\theta\, x} & \text{if } 0 \le x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Hence $X \sim GAM\left(\alpha, \frac{1}{\theta}\right)$ and $Y \sim GAM\left(\alpha+\beta, \frac{1}{\theta}\right)$. Therefore, we have the following theorem.

**Theorem 12.9.** If $(X, Y) \sim GAMM(\theta, \alpha, \beta)$, then

$$\begin{aligned} E(X) &= \frac{\alpha}{\theta} \\ E(Y) &= \frac{\alpha+\beta}{\theta} \\ Var(X) &= \frac{\alpha}{\theta^2} \\ Var(Y) &= \frac{\alpha+\beta}{\theta^2} \\ M(s,t) &= \left(\frac{\theta}{\theta-s-t}\right)^\alpha \left(\frac{\theta}{\theta-t}\right)^\beta. \end{aligned}$$

We state the various properties of the conditional densities of $f(x,y)$, without proof, in the following theorem.

**Theorem 12.10.** If $(X,Y) \sim GAMM(\theta,\alpha,\beta)$, then

$$\begin{aligned} E(Y/x) &= x + \frac{\beta}{\theta} \\ E(X/y) &= \frac{\alpha\, y}{\alpha+\beta} \\ Var(Y/x) &= \frac{\beta}{\theta^2} \\ Var(X/y) &= \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}\, y^2. \end{aligned}$$

We know that the univariate exponential distribution is a special case of the univariate gamma distribution. Similarly, the bivariate exponential distribution is a special case of bivariate gamma distribution. On taking the index parameters to be unity in the Kibble and Cherian bivariate gamma distribution given above, we obtain the corresponding bivariate exponential distributions. The bivariate exponential probability density function corresponding to bivariate gamma distribution of Kibble is given by

$$f(x,y) = \begin{cases} e^{-\left(\frac{x+y}{1-\theta}\right)} \sum_{k=0}^{\infty} \dfrac{(\theta\, x\, y)^k}{k!\,\Gamma(k+1)\,(1-\theta)^{2k+1}} & \text{if } 0 < x, y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

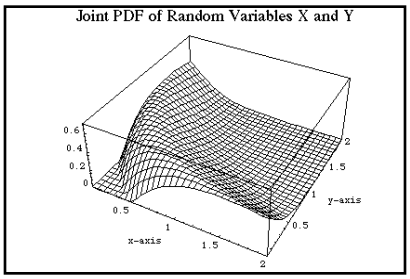
where $\theta \in (0,1)$ is a parameter. The bivariate exponential distribution corresponding to the Cherian bivariate distribution is the following:

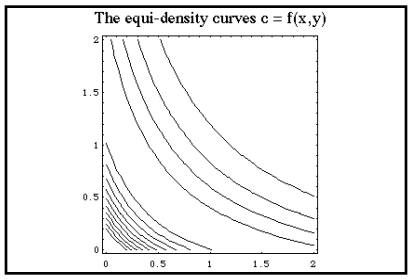
$$f(x,y) = \begin{cases} \left[e^{\min\{x,y\}} - 1\right]\, e^{-(x+y)} & \text{if } 0 < x,\, y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

In 1960, Gumble has studied the following bivariate exponential distribution whose density function is given by

$$f(x,y) = \begin{cases} [(1+\theta x)\,(1+\theta y) - \theta]\, e^{-(x+y+\theta\, x\, y)} & \text{if } 0 < x,\, y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter.




In 1967, Marshall and Olkin introduced the following bivariate exponential distribution

$$F(x,y) = \begin{cases} 1 - e^{-(\alpha+\gamma)x} - e^{-(\beta+\gamma)y} + e^{-(\alpha x+\beta y+\gamma \max\{x,y\})} & \text{if } x,y > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha, \beta, \gamma > 0$ are parameters. The exponential distribution function of Marshall and Olkin satisfies the lack of memory property

$$P(X > x + t,\, Y > y + t \;/\; X > t,\, Y > t) = P(X > x,\, Y > y).$$

### 12.4. Bivariate Beta Distribution

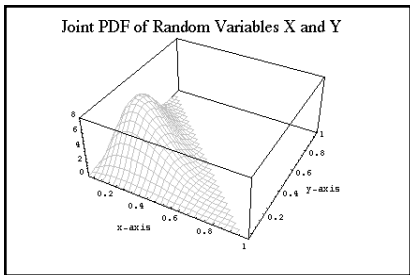
The bivariate beta distribution (also known as Dirichlet distribution ) is one of the basic distributions in statistics. The bivariate beta distribution is used in geology, biology, and chemistry for handling compositional data which are subject to nonnegativity and constant-sum constraints. It is also used nowadays with increasing frequency in statistical modeling, distribution theory and Bayesian statistics. For example, it is used to model the distribution of brand shares of certain consumer products, and in describing the joint distribution of two soil strength parameters. Further, it is used in modeling the proportions of the electorates who vote for a candidates in a two-candidate election. In Bayesian statistics, the beta distribution is very popular as a prior since it yields a beta distribution as posterior. In this section, we give some basic facts about the bivariate beta distribution.

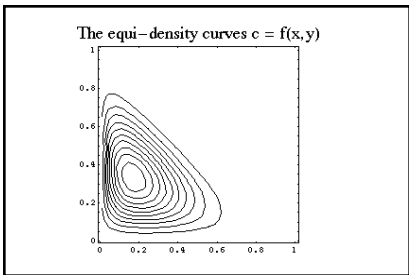
**Definition 12.5.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate beta distribution if its joint probability density function is of the form

$$f(x,y) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+\theta_3)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} x^{\theta_1-1} y^{\theta_2-1} (1-x-y)^{\theta_3-1} & \text{if } 0 < x, y, x+y < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta_1, \theta_2, \theta_3$ are positive parameters. We will denote a bivariate beta random variable $(X, Y)$ with positive parameters $\theta_1, \theta_2$ and $\theta_3$ by writing $(X, Y) \sim Beta(\theta_1, \theta_2, \theta_3)$.

The following figures show the graph and the equi-density curves of $f(x, y)$ on the domain of its definition.




In the following theorem, we present the expected values, the variances of the random variables $X$ and $Y$, and the correlation between $X$ and $Y$.

**Theorem 12.11.** Let $(X, Y) \sim Beta(\theta_1, \theta_2, \theta_3)$, where $\theta_1, \theta_2$ and $\theta_3$ are positive apriori chosen parameters. Then $X \sim Beta(\theta_1, \theta_2 + \theta_3)$ and $Y \sim Beta(\theta_2, \theta_1 + \theta_3)$ and

$$E(X) = \frac{\theta_1}{\theta}, \qquad Var(X) = \frac{\theta_1 (\theta - \theta_1)}{\theta^2 (\theta + 1)}$$

$$E(Y) = \frac{\theta_2}{\theta}, \qquad Var(Y) = \frac{\theta_2 (\theta - \theta_2)}{\theta^2 (\theta + 1)}$$

$$Cov(X, Y) = -\frac{\theta_1 \theta_2}{\theta^2 (\theta + 1)}$$

where $\theta = \theta_1 + \theta_2 + \theta_3$.

**Proof:** First, we show that $X \sim Beta(\theta_1, \theta_2 + \theta_3)$ and $Y \sim Beta(\theta_2, \theta_1 + \theta_3)$. Since $(X, Y) \sim Beta(\theta_2, \theta_1, \theta_3)$, the joint density of $(X, Y)$ is given by

$$f(x, y) = \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} x^{\theta_1 - 1} y^{\theta_2 - 1} (1 - x - y)^{\theta_3 - 1},$$

where $\theta = \theta_1 + \theta_2 + \theta_3$. Thus the marginal density of $X$ is given by

$$\begin{aligned} f_1(x) &= \int_0^1 f(x,y)\,dy \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} x^{\theta_1-1} \int_0^{1-x} y^{\theta_2-1}(1-x-y)^{\theta_3-1}\,dy \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} x^{\theta_1-1}(1-x)^{\theta_3-1} \int_0^{1-x} y^{\theta_2-1}\left(1-\frac{y}{1-x}\right)^{\theta_3-1} dy \end{aligned}$$

Now we substitute $u = 1 - \frac{y}{1-x}$ in the above integral. Then we have

$$\begin{aligned} f_1(x) &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} x^{\theta_1-1}(1-x)^{\theta_2+\theta_3-1} \int_0^1 u^{\theta_2-1}(1-u)^{\theta_3-1}\,du \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} x^{\theta_1-1}(1-x)^{\theta_2+\theta_3-1}\, B(\theta_2,\theta_3) \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2+\theta_3)} x^{\theta_1-1}(1-x)^{\theta_2+\theta_3-1} \end{aligned}$$

since

$$\int_0^1 u^{\theta_2-1}(1-u)^{\theta_3-1}\,du = B(\theta_2,\theta_3) = \frac{\Gamma(\theta_2)\Gamma(\theta_3)}{\Gamma(\theta_2+\theta_3)}.$$

This proves that the random variable $X \sim Beta(\theta_1, \theta_2+\theta_3)$. Similarly, one can shows that the random variable $Y \sim Beta(\theta_2, \theta_1+\theta_3)$. Now using Theorem 6.5, we see that

$$E(X) = \frac{\theta_1}{\theta}, \qquad Var(X) = \frac{\theta_1\,(\theta-\theta_1)}{\theta^2\,(\theta+1)}$$

$$E(Y) = \frac{\theta_2}{\theta}, \qquad Var(X) = \frac{\theta_2\,(\theta-\theta_2)}{\theta^2\,(\theta+1)},$$

where $\theta = \theta_1 + \theta_2 + \theta_3$.

Next, we compute the product moment of $X$ and $Y$. Consider

$$\begin{aligned} &E(XY) \\ &= \int_0^1 \left[\int_0^{1-x} xy\, f(x,y)\,dy\right] dx \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^1 \left[\int_0^{1-x} xy\, x^{\theta_1-1} y^{\theta_2-1}(1-x-y)^{\theta_3-1} dy\right] dx \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^1 \left[\int_0^{1-x} x^{\theta_1} y^{\theta_2}(1-x-y)^{\theta_3-1} dy\right] dx \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^1 x^{\theta_1}(1-x)^{\theta_3-1} \left[\int_0^{1-x} y^{\theta_2}\left(1-\frac{y}{1-x}\right)^{\theta_3-1} dy\right] dx. \end{aligned}$$

Now we substitute $u = \frac{y}{1-x}$ in the above integral to obtain

$$E(XY) = \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^1 x^{\theta_1}(1-x)^{\theta_2+\theta_3} \left[\int_0^1 u^{\theta_2}(1-u)^{\theta_3-1}du\right] dx$$

Since

$$\int_0^1 u^{\theta_2}(1-u)^{\theta_3-1}du = B(\theta_2+1, \theta_3)$$

and

$$\int_0^1 x^{\theta_1}(1-x)^{\theta_2+\theta_3}dx = B(\theta_1+1, \theta_2+\theta_3+1)$$

we have

$$\begin{aligned} E(XY) &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} B(\theta_2+1,\theta_3)\, B(\theta_1+1,\theta_2+\theta_3+1) \\ &= \frac{\Gamma(\theta)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} \frac{\theta_1\Gamma(\theta_1)(\theta_2+\theta_3)\Gamma(\theta_2+\theta_3)}{(\theta)(\theta+1)\Gamma(\theta)} \frac{\theta_2\Gamma(\theta_2)\Gamma(\theta_3)}{(\theta_2+\theta_3)\Gamma(\theta_2+\theta_3)} \\ &= \frac{\theta_1\theta_2}{\theta\,(\theta+1)} \quad \text{where } \theta = \theta_1+\theta_2+\theta_3. \end{aligned}$$

Now it is easy to compute the covariance of $X$ and $Y$ since

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= \frac{\theta_1\theta_2}{\theta\,(\theta+1)} - \frac{\theta_1}{\theta}\,\frac{\theta_2}{\theta} \\ &= -\frac{\theta_1\,\theta_2}{\theta^2\,(\theta+1)}. \end{aligned}$$

The proof of the theorem is now complete.

The correlation coefficient of $X$ and $Y$ can be computed using the covariance as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = -\sqrt{\frac{\theta_1\theta_2}{(\theta_1+\theta_3)(\theta_2+\theta_3)}}.$$

Next theorem states some properties of the conditional density functions $f(x/y)$ and $f(y/x)$.

**Theorem 12.12.** Let $(X,Y) \sim Beta(\theta_1,\theta_2,\theta_3)$ where $\theta_1,\theta_2$ and $\theta_3$ are positive parameters. Then

$$\begin{aligned} E(Y/x) &= \frac{\theta_2(1-x)}{\theta_2+\theta_3}, \qquad Var(Y/x) = \frac{\theta_2\,\theta_3(1-x)^2}{(\theta_2+\theta_3)^2(\theta_2+\theta_3+1)} \\ E(X/y) &= \frac{\theta_1(1-y)}{\theta_1+\theta_3}, \qquad Var(X/y) = \frac{\theta_1\,\theta_3(1-y)^2}{(\theta_1+\theta_3)^2(\theta_1+\theta_3+1)}. \end{aligned}$$

**Proof:** We know that if $(X, Y) \sim Beta(\theta_1, \theta_2, \theta_3)$, the random variable $X \sim Beta(\theta_1, \theta_2 + \theta_3)$. Therefore

$$\begin{aligned} f(y/x) &= \frac{f(x,y)}{f_1(x)} \\ &= \frac{1}{1-x}\ \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)} \left(\frac{y}{1-x}\right)^{\theta_2-1} \left(1-\frac{y}{1-x}\right)^{\theta_3-1} \end{aligned}$$

for all $0 < y < 1 - x$. Thus the random variable $\frac{Y}{1-x}\Big|_{X=x}$ is a beta random variable with parameters $\theta_2$ and $\theta_3$.

Now we compute the conditional expectation of $Y/x$. Consider

$$\begin{aligned} E(Y/x) &= \int_0^{1-x} y\, f(y/x)\, dy \\ &= \frac{1}{1-x}\ \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^{1-x} y \left(\frac{y}{1-x}\right)^{\theta_2-1} \left(1-\frac{y}{1-x}\right)^{\theta_3-1} dy. \end{aligned}$$

Now we substitute $u = \frac{y}{1-x}$ in the above integral to obtain

$$\begin{aligned} E(Y/x) &= \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)}\,(1-x) \int_0^1 u^{\theta_2}(1-u)^{\theta_3-1}\, du \\ &= \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)}\,(1-x)\, B(\theta_2+1, \theta_3) \\ &= \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)}\,(1-x)\, \frac{\theta_2\,\Gamma(\theta_2)\Gamma(\theta_3)}{(\theta_2+\theta_3)\,\Gamma(\theta_2+\theta_3)} \\ &= \frac{\theta_2}{\theta_2+\theta_3}\,(1-x). \end{aligned}$$

Next, we compute $E(Y^2/x)$. Consider

$$\begin{aligned} E(Y^2/x) &= \int_0^{1-x} y^2\, f(y/x)\, dy \\ &= \frac{1}{1-x}\ \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)} \int_0^{1-x} y^2 \left(\frac{y}{1-x}\right)^{\theta_2-1} \left(1-\frac{y}{1-x}\right)^{\theta_3-1} dy \\ &= \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)}\,(1-x)^2 \int_0^1 u^{\theta_2+1}\,(1-u)^{\theta_3-1}\, du \\ &= \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)}\,(1-x)^2\, B(\theta_2+2, \theta_3) \\ &= \frac{\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2)\Gamma(\theta_3)}\,(1-x)^2\, \frac{(\theta_2+1)\,\theta_2\,\Gamma(\theta_2)\Gamma(\theta_3)}{(\theta_2+\theta_3+1)\,(\theta_2+\theta_3)\,\Gamma(\theta_2+\theta_3)} \\ &= \frac{(\theta_2+1)\,\theta_2}{(\theta_2+\theta_3+1)\,(\theta_2+\theta_3}(1-x)^2. \end{aligned}$$

Therefore

$$Var(Y/x) = E(Y^2/x) - E(Y/x)^2 = \frac{\theta_2\theta_3\,(1-x)^2}{(\theta_2+\theta_3)^2(\theta_2+\theta_3+1)}.$$

Similarly, one can compute $E(X/y)$ and $Var(X/y)$. We leave this computation to the reader. Now the proof of the theorem is now complete.

The Dirichlet distribution can be extended from the unit square $(0,1)^2$ to an arbitrary rectangle $(a_1,b_1)\times(a_2,b_2)$.

**Definition 12.6.** A continuous bivariate random variable $(X_1,X_2)$ is said to have the generalized bivariate beta distribution if its joint probability density function is of the form

$$f(x_1,x_2) = \frac{\Gamma(\theta_1+\theta_2+\theta_3)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)}\prod_{k=1}^{2}\left(\frac{x_k-a_k}{b_k-a_k}\right)^{\theta_k-1}\left(1-\frac{x_k-a_k}{b_k-a_k}\right)^{\theta_3-1}$$

where $0<x_1,x_2,x_1+x_2<1$ and $\theta_1,\theta_2,\theta_3,a_1,b_1,a_2,b_2$ are parameters. We will denote a bivariate generalized beta random variable $(X,Y)$ with positive parameters $\theta_1,\theta_2$ and $\theta_3$ by writing $(X,Y)\sim GBeta(\theta_1,\theta_2,\theta_3,a_1,b_1,a_2,b_2)$.

It can be shown that if $X_k=(b_k-a_k)Y_k+a_k$ (for $k=1,2$) and each $(Y_1,Y_2)\sim Beta(\theta_1,\theta_2,\theta_3)$, then $(X_1,X_2)\sim GBeta(\theta_1,\theta_2,\theta_3,a_1,b_1,a_2,b_2)$. Therefore, by Theorem 12.11

**Theorem 12.13.** Let $(X,Y)\sim GBeta(\theta_1,\theta_2,\theta_3,a_1,b_1,a_2,b_2)$, where $\theta_1,\theta_2$ and $\theta_3$ are positive apriori chosen parameters. Then $X\sim Beta(\theta_1,\theta_2+\theta_3)$ and $Y\sim Beta(\theta_2,\theta_1+\theta_3)$ and

$$E(X) = (b_1-a_1)\frac{\theta_1}{\theta}+a_1, \qquad Var(X) = (b_1-a_1)^2\,\frac{\theta_1\,(\theta-\theta_1)}{\theta^2\,(\theta+1)}$$

$$E(Y) = (b_2-a_2)\frac{\theta_2}{\theta}+a_2, \qquad Var(Y) = (b_2-a_2)^2\,\frac{\theta_2\,(\theta-\theta_2)}{\theta^2\,(\theta+1)}$$

$$Cov(X,Y) = -(b_1-a_1)(b_2-a_2)\,\frac{\theta_1\,\theta_2}{\theta^2\,(\theta+1)}$$

where $\theta=\theta_1+\theta_2+\theta_3$.

Another generalization of the bivariate beta distribution is the following:

**Definition 12.7.** A continuous bivariate random variable $(X_1,X_2)$ is said to have the generalized bivariate beta distribution if its joint probability density function is of the form

$$f(x_1,x_2) = \frac{1}{B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)}x^{\alpha_1-1}(1-x_1)^{\beta_1-\alpha_2-\beta_2}\,x_2^{\alpha_2-1}\,(1-x_1x_2)^{\beta_2-1}$$

where $0 < x_1, x_2, x_1 + x_2 < 1$ and $\alpha_1, \alpha_2, \beta_1, \beta_2$ are parameters.

It is not difficult to see that $X \sim Beta(\alpha_1, \beta_1)$ and $Y \sim Beta(\alpha_2, \beta_2)$.

### 12.5. Bivariate Normal Distribution

The bivariate normal distribution is a generalization of the univariate normal distribution. The first statistical treatment of the bivariate normal distribution was given by Galton and Dickson in 1886. Although there are several other bivariate distributions as discussed above, the bivariate normal distribution still plays a dominant role. The development of normal theory has been intensive and most thinking has centered upon bivariate normal distribution because of the relative simplicity of mathematical treatment of it. In this section, we give an in depth treatment of the bivariate normal distribution.

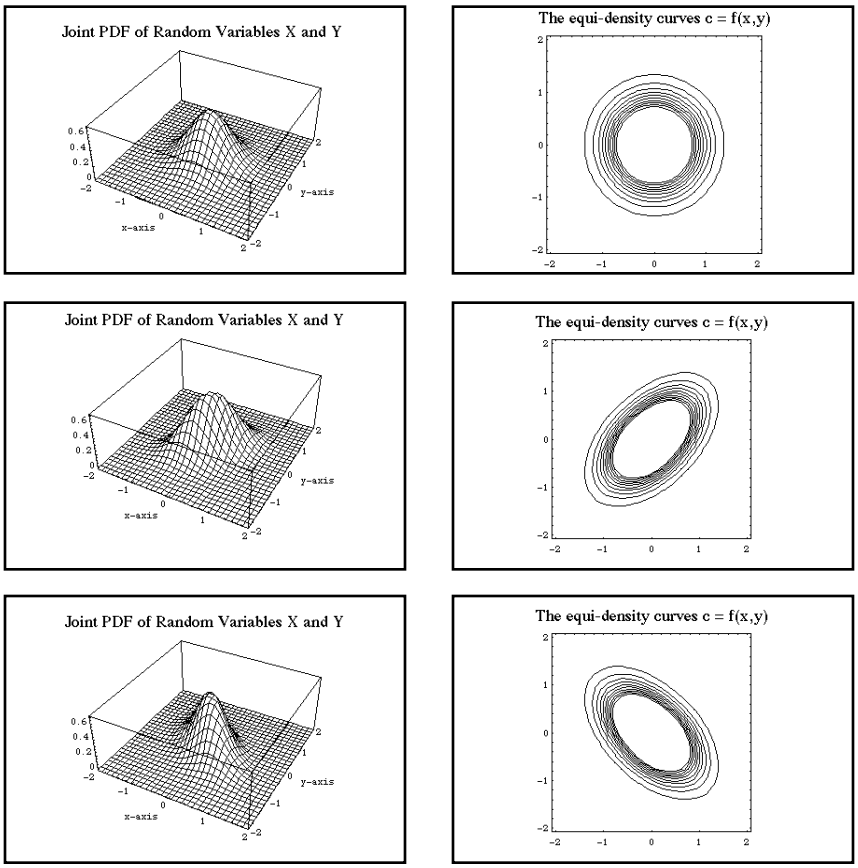
**Definition 12.8.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate normal distribution if its joint probability density function is of the form

$$f(x,y) = \frac{1}{2\,\pi\,\sigma_1\,\sigma_2\,\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}\,Q(x,y)}, \qquad -\infty < x, y < \infty,$$

where $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 \in (0, \infty)$ and $\rho \in (-1, 1)$ are parameters, and

$$Q(x,y) := \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\,\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

As usual, we denote this bivariate normal random variable by writing $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. The graph of $f(x, y)$ has a shape of a "mountain". The pair $(\mu_1, \mu_2)$ tells us where the center of the mountain is located in the $(x, y)$-plane, while $\sigma_1^2$ and $\sigma_2^2$ measure the spread of this mountain in the $x$-direction and $y$-direction, respectively. The parameter $\rho$ determines the shape and orientation on the $(x, y)$-plane of the mountain. The following figures show the graphs of the bivariate normal distributions with different values of correlation coefficient $\rho$. The first two figures illustrate the graph of the bivariate normal distribution with $\rho = 0$, $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 1$ and the equi-density plots. The next two figures illustrate the graph of the bivariate normal distribution with $\rho = 0.5$, $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 0.5$ and the equi-density plots. The last two figures illustrate the graph of the bivariate normal distribution with $\rho = -0.5$, $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 0.5$ and the equi-density plots.




One of the remarkable features of the bivariate normal distribution is that if we vertically slice the graph of $f(x, y)$ along any direction, we obtain a univariate normal distribution. In particular, if we vertically slice the graph of the $f(x, y)$ along the $x$-axis, we obtain a univariate normal distribution. That is the marginal of $f(x, y)$ is again normal. One can show that the marginals of $f(x, y)$ are given by

$$f_1(x) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$$

and

$$f_2(y) = \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}.$$

In view of these, the following theorem is obvious.

**Theorem 12.14.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, then

$$\begin{aligned} E(X) &= \mu_1 \\ E(Y) &= \mu_2 \\ Var(X) &= \sigma_1^2 \\ Var(Y) &= \sigma_2^2 \\ Corr(X, Y) &= \rho \\ M(s, t) &= e^{\mu_1 s + \mu_2 t + \frac{1}{2}(\sigma_1^2 s^2 + 2\rho\sigma_1\sigma_2 st + \sigma_2^2 t^2)}. \end{aligned}$$

**Proof:** It is easy to establish the formulae for $E(X)$, $E(Y)$, $Var(X)$ and $Var(Y)$. Here we only establish the moment generating function. Since $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, we have $X \sim N\left(\mu_1, \sigma_1^2\right)$ and $Y \sim N\left(\mu_2, \sigma_2^2\right)$. Further, for any $s$ and $t$, the random variable $W = sX + tY$ is again normal with

$$\mu_W = s\mu_1 + t\mu_2 \qquad \text{and} \qquad \sigma_W^2 = s^2\sigma_1^2 + 2st\rho\sigma_1\sigma_2 + t^2\sigma_2.$$

Since $W$ is a normal random variable, its moment generating function is given by

$$M(\tau) = e^{\mu_W \tau + \frac{1}{2}\tau^2\sigma_W^2}.$$

The joint moment generating function of $(X, Y)$ is

$$\begin{aligned} M(s, t) &= E\left(e^{sX+tY}\right) \\ &= e^{\mu_W + \frac{1}{2}\sigma_W^2} \\ &= e^{\mu_1 s + \mu_2 t + \frac{1}{2}(\sigma_1^2 s^2 + 2\rho\sigma_1\sigma_2 st + \sigma_2^2 t^2)}. \end{aligned}$$

This completes the proof of the theorem.

It can be shown that the conditional density of $Y$ given $X = x$ is

$$f(y/x) = \frac{1}{\sigma_2\sqrt{2\pi\left(1-\rho^2\right)}}\, e^{-\frac{1}{2}\left(\frac{y-b}{\sigma_2\sqrt{1-\rho^2}}\right)^2}$$

where

$$b = \mu_2 + \rho\,\frac{\sigma_2}{\sigma_1}\,(x - \mu_1).$$

Similarly, the conditional density $f(x/y)$ is

$$f(x/y) = \frac{1}{\sigma_1\sqrt{2\pi\left(1-\rho^2\right)}}\, e^{-\frac{1}{2}\left(\frac{x-c}{\sigma_1\sqrt{1-\rho^2}}\right)^2},$$

where

$$c = \mu_1 + \rho \, \frac{\sigma_1}{\sigma_2} \, (y - \mu_2).$$

In view of the form of $f(y/x)$ and $f(x/y)$, the following theorem is transparent.

**Theorem 12.15.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, then

$$\begin{aligned} E(Y/x) &= \mu_2 + \rho \, \frac{\sigma_2}{\sigma_1} \, (x - \mu_1) \\ E(X/y) &= \mu_1 + \rho \, \frac{\sigma_1}{\sigma_2} \, (y - \mu_2) \\ Var(Y/x) &= \sigma_2^2 \, (1 - \rho^2) \\ Var(X/y) &= \sigma_1^2 \, (1 - \rho^2). \end{aligned}$$

We have seen that if $(X, Y)$ has a bivariate normal distribution, then the distributions of $X$ and $Y$ are also normal. However, the converse of this is not true. That is if $X$ and $Y$ have normal distributions as their marginals, then their joint distribution is not necessarily bivariate normal.

Now we present some characterization theorems concerning the bivariate normal distribution. The first theorem is due to Cramer (1941).

**Theorem 12.16.** The random variables $X$ and $Y$ have a joint bivariate normal distribution if and only if every linear combination of $X$ and $Y$ has a univariate normal distribution.

**Theorem 12.17.** The random variables $X$ and $Y$ with unit variances and correlation coefficient $\rho$ have a joint bivariate normal distribution if and only if

$$\frac{\partial}{\partial \rho} E[g(X, Y)] = E\left[\frac{\partial^2}{\partial X \, \partial Y} g(X, Y)\right]$$

holds for an *arbitrary* function $g(x, y)$ of two variable.

Many interesting characterizations of bivariate normal distribution can be found in the survey paper of Hamedani (1992).

### 12.6. Bivariate Logistic Distributions

In this section, we study two bivariate logistic distributions. A univariate logistic distribution is often considered as an alternative to the univariate normal distribution. The univariate logistic distribution has a shape very close to that of a univariate normal distribution but has heavier tails than

the normal. This distribution is also used as an alternative to the univariate Weibull distribution in life-testing. The univariate logistic distribution has the following probability density function

$$f(x) = \frac{\pi}{\sigma\sqrt{3}} \; \frac{e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}}{\left[1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}\right]^2} \qquad -\infty < x < \infty,$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are parameters. The parameter $\mu$ is the mean and the parameter $\sigma$ is the standard deviation of the distribution. A random variable $X$ with the above logistic distribution will be denoted by $X \sim LOG(\mu, \sigma)$. It is well known that the moment generating function of univariate logistic distribution is given by

$$M(t) = e^{\mu t}\,\Gamma\left(1 + \frac{\sqrt{3}}{\pi}\,\sigma t\right)\,\Gamma\left(1 - \frac{\sqrt{3}}{\pi}\,\sigma t\right)$$

for $|t| < \frac{\pi}{\sigma\sqrt{3}}$. We give brief proof of the above result for $\mu = 0$ and $\sigma = \frac{\pi}{\sqrt{3}}$. Then with these assumptions, the logistic density function reduces to

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

The moment generating function with respect to this density function is

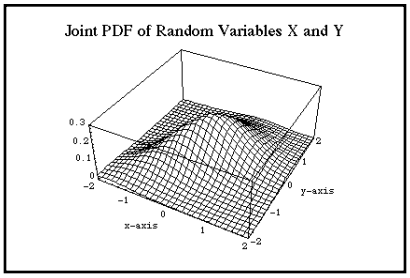
$$\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} e^{tx}\, f(x)\, dx \\
&= \int_{-\infty}^{\infty} e^{tx}\, \frac{e^{-x}}{(1 + e^{-1})^2}\, dx \\
&= \int_{-\infty}^{\infty} \left(e^{-x}\right)^{-t}\, \frac{e^{-x}}{(1 + e^{-1})^2}\, dx \\
&= \int_{0}^{1} \left(z^{-1} - 1\right)^{-t}\, dz \qquad \text{where } z = \frac{1}{1 + e^{-x}} \\
&= \int_{0}^{1} z^{t}\,(1 - z)^{-t}\, dz \\
&= B(1 + t,\, 1 - t) \\
&= \frac{\Gamma(1 + t)\,\Gamma(1 - t)}{\Gamma(1 + t + 1 - t)} \\
&= \frac{\Gamma(1 + t)\,\Gamma(1 - t)}{\Gamma(2)} \\
&= \Gamma(1 + t)\,\Gamma(1 - t) \\
&= t\, cosec(t).
\end{aligned}$$

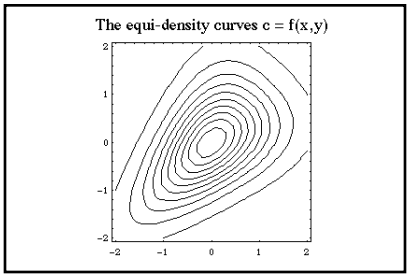
Recall that the marginals and conditionals of the bivariate normal distribution are univariate normal. This beautiful property enjoyed by the bivariate normal distribution are apparently lacking from other bivariate distributions we have discussed so far. If we can not define a bivariate logistic distribution so that the conditionals and marginals are univariate logistic, then we would like to have at least one of the marginal distributions logistic and the conditional distribution of the other variable logistic. The following bivariate logistic distribution is due to Gumble (1961).

**Definition 12.9.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate logistic distribution of first kind if its joint probability density function is of the form

$$f(x,y) = \frac{2\,\pi^2\, e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}+\frac{y-\mu_2}{\sigma_2}\right)}}{3\,\sigma_1\,\sigma_2\left[1+e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}\right)}+e^{-\frac{\pi}{\sqrt{3}}\left(\frac{y-\mu_2}{\sigma_2}\right)}\right]^3} \qquad -\infty < x,\, y < \infty,$$

where $-\infty < \mu_1, \mu_2 < \infty$, and $0 < \sigma_1, \sigma_2 < \infty$ are parameters. If a random variable $(X, Y)$ has a bivariate logistic distribution of first kind, then we express this by writing $(X, Y) \sim LOGF(\mu_1, \mu_2, \sigma_1, \sigma_2)$. The following figures show the graph of $f(x, y)$ with $\mu_1 = 0 = \mu_2$ and $\sigma_1 = 1 = \sigma_2$ and the equi-density plots.




It can be shown that marginally, $X$ is a logistic random variable. That is, $X \sim LOG(\mu_1, \sigma_1)$. Similarly, $Y \sim LOG(\mu_2, \sigma_2)$. These facts lead us to the following theorem.

**Theorem 12.18.** If the random variable $(X, Y) \sim LOGF(\mu_1, \mu_2, \sigma_1, \sigma_2)$,

then

$$\begin{aligned} E(X) &= \mu_1 \\ E(Y) &= \mu_2 \\ Var(X) &= \sigma_1^2 \\ Var(Y) &= \sigma_2^2 \\ E(XY) &= \frac{1}{2}\,\sigma_1\,\sigma_2 + \mu_1\,\mu_2, \end{aligned}$$

and the moment generating function is given by

$$M(s,t) = e^{\mu_1 s + \mu_2 t}\Gamma\left(1 + \frac{(\sigma_1 s + \sigma_2 t)\sqrt{3}}{\pi}\right)\Gamma\left(1 - \frac{\sigma_1 s\sqrt{3}}{\pi}\right)\Gamma\left(1 - \frac{\sigma_2 t\sqrt{3}}{\pi}\right)$$

for $|s| < \frac{\pi}{\sigma_1\sqrt{3}}$ and $|t| < \frac{\pi}{\sigma_2\sqrt{3}}$.

It is an easy exercise to see that if the random variables $X$ and $Y$ have a joint bivariate logistic distribution, then the correlation between $X$ and $Y$ is $\frac{1}{2}$. This can be considered as one of the drawbacks of this distribution in the sense that it limits the dependence between the random variables $X$ and $Y$.

The conditional density of $Y$ given $X = x$ is

$$f(y/x) = \frac{2\,\pi}{\sigma_2\sqrt{3}}\,e^{-\frac{\pi}{\sqrt{3}}\left(\frac{y-\mu_2}{\sigma_2}\right)}\,\frac{\left[1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}\right)}\right]^2}{\left[1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}\right)} + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{y-\mu_2}{\sigma_2}\right)}\right]^3}.$$

Similarly the conditional density of $X$ given $Y = y$ is

$$f(x/y) = \frac{2\,\pi}{\sigma_1\sqrt{3}}\,e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}\right)}\,\frac{\left[1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{y-\mu_2}{\sigma_2}\right)}\right]^2}{\left[1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}\right)} + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{y-\mu_2}{\sigma_2}\right)}\right]^3}.$$

Using these densities, the next theorem offers various conditional properties of the bivariate logistic distribution.

**Theorem 12.19.** If the random variable $(X,Y) \sim LOGF(\mu_1,\mu_2,\sigma_1,\sigma_2)$,

then

$$E(Y/x) = 1 - \ln\left(1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu_1}{\sigma_1}\right)}\right)$$

$$E(X/y) = 1 - \ln\left(1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{y-\mu_2}{\sigma_2}\right)}\right)$$

$$Var(Y/x) = \frac{\pi^3}{3} - 1$$

$$Var(X/y) = \frac{\pi^3}{3} - 1.$$

It was pointed out earlier that one of the drawbacks of this bivariate logistic distribution of first kind is that it limits the dependence of the random variables. The following bivariate logistic distribution was suggested to rectify this drawback.

**Definition 12.10.** A continuous bivariate random variable $(X, Y)$ is said to have the bivariate logistic distribution of second kind if its joint probability density function is of the form

$$f(x,y) = \frac{[\phi_\alpha(x,y)]^{1-2\alpha}}{[1+\phi_\alpha(x,y)]^2}\left(\frac{\phi_\alpha(x,y)-1}{\phi_\alpha(x,y)+1} + \alpha\right) e^{-\alpha(x+y)},\ -\infty < x, y < \infty,$$

where $\alpha > 0$ is a parameter, and $\phi_\alpha(x,y) := (e^{-\alpha x} + e^{-\alpha y})^{\frac{1}{\alpha}}$. As before, we denote a bivariate logistic random variable of second kind $(X, Y)$ by writing $(X, Y) \sim LOGS(\alpha)$.

The marginal densities of $X$ and $Y$ are again logistic and they given by

$$f_1(x) = \frac{e^{-x}}{(1+e^{-x})^2},\ -\infty < x < \infty$$

and

$$f_2(y) = \frac{e^{-y}}{(1+e^{-y})^2},\ -\infty < y < \infty.$$

It was shown by Oliveira (1961) that if $(X, Y) \sim LOGS(\alpha)$, then the correlation between $X$ and $Y$ is

$$\rho(X,Y) = 1 - \frac{1}{2\,\alpha^2}.$$

### 12.7. Review Exercises

**1.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ with $Q(x, y) = x^2 + 2y^2 - 2xy + 2x - 2y + 1$, then what is the value of the conditional variance of $Y$ given the event $X = x$?

**2.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ with

$$Q(x, y) = -\frac{1}{102}\left[(x+3)^2 - 16(x+3)(y-2) + 4(y-2)^2\right],$$

then what is the value of the conditional expectation of $Y$ given $X = x$?

**3.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, then what is the correlation coefficient of the random variables $U$ and $V$, where $U = 2X + 3Y$ and $V = 2X - 3Y$?

**4.** Let the random variables $X$ and $Y$ denote the height and weight of wild turkeys. If the random variables $X$ and $Y$ have a bivariate normal distribution with $\mu_1 = 18$ inches, $\mu_2 = 15$ pounds, $\sigma_1 = 3$ inches, $\sigma_2 = 2$ pounds, and $\rho = 0.75$, then what is the expected weight of one of these wild turkeys that is 17 inches tall?

**5.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, then what is the moment generating function of the random variables $U$ and $V$, where $U = 7X + 3Y$ and $V = 7X - 3Y$?

**6.** Let $(X, Y)$ have a bivariate normal distribution. The mean of $X$ is 10 and the variance of $X$ is 12. The mean of $Y$ is $-5$ and the variance of $Y$ is 5. If the covariance of $X$ and $Y$ is 4, then what is the probability that $X + Y$ is greater than 10?

**7.** Let $X$ and $Y$ have a bivariate normal distribution with means $\mu_X = 5$ and $\mu_Y = 6$, standard deviations $\sigma_X = 3$ and $\sigma_Y = 2$, and covariance $\sigma_{XY} = 2$. Let $\Phi$ denote the cumulative distribution function of a normal random variable with mean 0 and variance 1. What is $P(2 \leq X - Y \leq 5)$ in terms of $\Phi$ ?

**8.** If $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ with $Q(x, y) = -x^2 + xy - 2y^2$, then what is the conditional distributions of $X$ given the event $Y = y$?

**9.** If $(X, Y) \sim GAMK(\alpha, \theta)$, where $0 < \alpha < \infty$ and $0 \leq \theta < 1$ are parameters, then show that the moment generating function is given by

$$M(s, t) = \left(\frac{1}{(1-s)(1-t) - \theta\, s\, t}\right)^{\alpha}.$$

**10.** Let $X$ and $Y$ have a bivariate gamma distribution of Kibble with parameters $\alpha = 1$ and $0 \leq \theta < 0$. What is the probability that the random variable $7X$ is less than $\frac{1}{2}$?

**11.** If $(X, Y) \sim GAMC(\alpha, \beta, \gamma)$, then what are the regression and scedestic curves of $Y$ on $X$?

**12.** The position of a random point $(X, Y)$ is equally probable anywhere on a circle of radius $R$ and whose center is at the origin. What is the probability density function of each of the random variables $X$ and $Y$? Are the random variables $X$ and $Y$ independent?

**13.** If $(X, Y) \sim GAMC(\alpha, \beta, \gamma)$, what is the correlation coefficient of the random variables $X$ and $Y$?

**14.** Let $X$ and $Y$ have a bivariate exponential distribution of Gumble with parameter $\theta > 0$. What is the regression curve of $Y$ on $X$?

**15.** A screen of a navigational radar station represents a circle of radius 12 inches. As a result of noise, a spot may appear with its center at any point of the circle. Find the expected value and variance of the distance between the center of the spot and the center of the circle.

**16.** Let $X$ and $Y$ have a bivariate normal distribution. Which of the following statements must be true?

(I) Any nonzero linear combination of $X$ and $Y$ has a normal distribution.
(II) $E(Y/X = x)$ is a linear function of $x$.
(III) $Var(Y/X = x) \leq Var(Y)$.

**17.** If $(X, Y) \sim LOGS(\alpha)$, then what is the correlation between $X$ and $Y$?

**18.** If $(X, Y) \sim LOGF(\mu_1, \mu_2, \sigma_1, \sigma_2)$, then what is the correlation between the random variables $X$ and $Y$?

**19.** If $(X, Y) \sim LOGF(\mu_1, \mu_2, \sigma_1, \sigma_2)$, then show that marginally $X$ and $Y$ are univariate logistic.

**20.** If $(X, Y) \sim LOGF(\mu_1, \mu_2, \sigma_1, \sigma_2)$, then what is the scedastic curve of the random variable $Y$ and $X$?

# Chapter 13

# SEQUENCES OF RANDOM VARIABLES AND ORDER STASTISTICS

In this chapter, we generalize some of the results we have studied in the previous chapters. We do these generalizations because the generalizations are needed in the subsequent chapters relating mathematical statistics. In this chapter, we also examine the weak law of large numbers, Bernoulli's law of large numbers, the strong law of large numbers, and the central limit theorem. Further, in this chapter, we treat the order statistics and percentiles.

## 13.1. Distribution of sample mean and variance

Consider a random experiment. Let $X$ be the random variable associated with this experiment. Let $f(x)$ be the probability density function of $X$. Let us repeat this experiment $n$ times. Let $X_k$ be the random variable associated with the $k^{\text{th}}$ repetition. Then the collection of the random variables $\{ X_1, X_2, ..., X_n \}$ is a random sample of size $n$. From here after, we simply denote $X_1, X_2, ..., X_n$ as a random sample of size $n$. The random variables $X_1, X_2, ..., X_n$ are independent and identically distributed with the common probability density function $f(x)$.

For a random sample, functions such as the sample mean $\overline{X}$, the sample variance $S^2$ are called *statistics*. In a particular sample, say $x_1, x_2, ..., x_n$, we observed $\overline{x}$ and $s^2$. We may consider

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

as random variables and $\overline{x}$ and $s^2$ are the realizations from a particular sample.

In this section, we are mainly interested in finding the probability distributions of the sample mean $\overline{X}$ and sample variance $S^2$, that is the distribution of the statistics of samples.

**Example 13.1.** Let $X_1$ and $X_2$ be a random sample of size 2 from a distribution with probability density function

$$f(x) = \begin{cases} 6x(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

What are the mean and variance of sample sum $Y = X_1 + X_2$?

**Answer:** The population mean

$$\begin{aligned} \mu_X &= E(X) \\ &= \int_0^1 x\, 6x(1-x)\, dx \\ &= 6 \int_0^1 x^2 (1-x)\, dx \\ &= 6\, B(3,2) \qquad \text{(here } B \text{ denotes the beta function)} \\ &= 6\, \frac{\Gamma(3)\,\Gamma(2)}{\Gamma(5)} \\ &= 6 \left(\frac{1}{12}\right) \\ &= \frac{1}{2}. \end{aligned}$$

Since $X_1$ and $X_2$ have the same distribution, we obtain $\mu_{X_1} = \frac{1}{2} = \mu_{X_2}$. Hence the mean of $Y$ is given by

$$\begin{aligned} E(Y) &= E(X_1 + X_2) \\ &= E(X_1) + E(X_2) \\ &= \frac{1}{2} + \frac{1}{2} \\ &= 1. \end{aligned}$$

Next, we compute the variance of the population $X$. The variance of $X$ is given by

$$\begin{aligned}
Var(X) &= E\left(X^2\right) - E(X)^2 \\
&= \int_0^1 6x^3(1-x)\,dx - \left(\frac{1}{2}\right)^2 \\
&= 6\int_0^1 x^3\,(1-x)\,dx - \left(\frac{1}{4}\right) \\
&= 6\,B(4,2) - \left(\frac{1}{4}\right) \\
&= 6\,\frac{\Gamma(4)\,\Gamma(2)}{\Gamma(6)} - \left(\frac{1}{4}\right) \\
&= 6\,\left(\frac{1}{20}\right) - \left(\frac{1}{4}\right) \\
&= \frac{6}{20} - \frac{5}{20} \\
&= \frac{1}{20}.
\end{aligned}$$

Since $X_1$ and $X_2$ have the same distribution as the population $X$, we get

$$Var(X_1) = \frac{1}{20} = Var(X_2).$$

Hence, the variance of the sample sum $Y$ is given by

$$\begin{aligned}
Var(Y) &= Var\,(X_1 + X_2) \\
&= Var\,(X_1) + Var\,(X_2) + 2\,Cov\,(X_1, X_2) \\
&= Var\,(X_1) + Var\,(X_2) \\
&= \frac{1}{20} + \frac{1}{20} \\
&= \frac{1}{10}.
\end{aligned}$$

**Example 13.2.** Let $X_1$ and $X_2$ be a random sample of size 2 from a distribution with density

$$f(x) = \begin{cases} \frac{1}{4} & \text{for } x = 1, 2, 3, 4 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the distribution of the sample sum $Y = X_1 + X_2$ ?

**Answer:** Since the range space of $X_1$ as well as $X_2$ is $\{1, 2, 3, 4\}$, the range space of $Y = X_1 + X_2$ is

$$R_Y = \{2, 3, 4, 5, 6, 7, 8\}.$$

Let $g(y)$ be the density function of $Y$. We want to find this density function. First, we find $g(2)$, $g(3)$ and so on.

$$\begin{aligned}
g(2) &= P(Y = 2) \\
&= P(X_1 + X_2 = 2) \\
&= P(X_1 = 1 \text{ and } X_2 = 1) \\
&= P(X_1 = 1)\, P(X_2 = 1) \qquad \text{(by independence of } X_1 \text{ and } X_2) \\
&= f(1)\, f(1) \\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}.
\end{aligned}$$

$$\begin{aligned}
g(3) &= P(Y = 3) \\
&= P(X_1 + X_2 = 3) \\
&= P(X_1 = 1 \text{ and } X_2 = 2) + P(X_1 = 2 \text{ and } X_2 = 1) \\
&= P(X_1 = 1)\, P(X_2 = 2) \\
&\quad + P(X_1 = 2)\, P(X_2 = 1) \qquad \text{(by independence of } X_1 \text{ and } X_2) \\
&= f(1)\, f(2) + f(2)\, f(1) \\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{2}{16}.
\end{aligned}$$

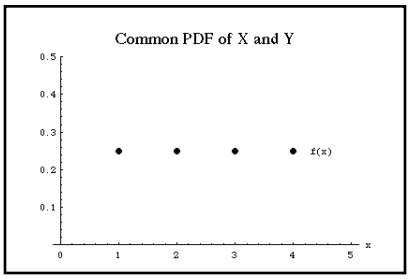
$$\begin{aligned}
g(4) &= P(Y=4)\\
&= P(X_1+X_2=4)\\
&= P(X_1=1 \text{ and } X_2=3)+P(X_1=3 \text{ and } X_2=1)\\
&\quad + P(X_1=2 \text{ and } X_2=2)\\
&= P(X_1=3)\ P(X_2=1)+P(X_1=1)\ P(X_2=3)\\
&\quad + P(X_1=2)\ P(X_2=2) \qquad \text{(by independence of } X_1 \text{ and } X_2)\\
&= f(1)\,f(3)+f(3)\,f(1)+f(2)\,f(2)\\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\\
&= \frac{3}{16}.
\end{aligned}$$

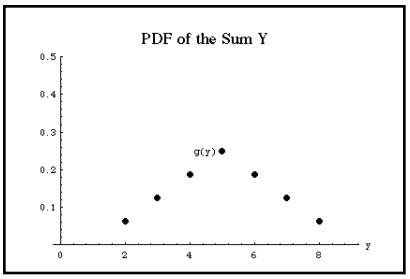
Similarly, we get

$$g(5)=\frac{4}{16}, \qquad g(6)=\frac{3}{16}, \qquad g(7)=\frac{2}{16}, \qquad g(8)=\frac{1}{16}.$$

Thus, putting these into one expression, we get

$$\begin{aligned}
g(y) &= P(Y=y)\\
&= \sum_{k=1}^{y-1} f(k)\,f(y-k)\\
&= \frac{4-|y-5|}{16}, \qquad y=2,3,4,...,8.
\end{aligned}$$




**Remark 13.1.** Note that $g(y)=\sum_{k=1}^{y-1} f(k)\,f(y-k)$ is the discrete convolution of $f$ with itself. The concept of convolution was introduced in chapter 10.

The above example can also be done using the moment generating func-

tion method as follows:

$$\begin{aligned} M_Y(t) &= M_{X_1+X_2}(t) \\ &= M_{X_1}(t)\, M_{X_2}(t) \\ &= \left(\frac{e^t + e^{2t} + e^{3t} + e^{4t}}{4}\right)\left(\frac{e^t + e^{2t} + e^{3t} + e^{4t}}{4}\right) \\ &= \left(\frac{e^t + e^{2t} + e^{3t} + e^{4t}}{4}\right)^2 \\ &= \frac{e^{2t} + 2e^{3t} + 3e^{4t} + 4e^{5t} + 3e^{6t} + 2e^{7t} + e^{8t}}{16}. \end{aligned}$$

Hence, the density of $Y$ is given by

$$g(y) = \frac{4 - |y-5|}{16}, \qquad y = 2, 3, 4, ..., 8.$$

**Theorem 13.1.** If $X_1, X_2, ..., X_n$ are mutually independent random variables with densities $f_1(x_1)$, $f_2(x_2), ..., f_n(x_n)$ and $E[u_i(X_i)]$, $i = 1, 2, ..., n$ exist, then

$$E\left[\prod_{i=1}^{n} u_i(X_i)\right] = \prod_{i=1}^{n} E[u_i(X_i)],$$

where $u_i\,(i = 1, 2, ..., n)$ are arbitrary functions.

**Proof:** We prove the theorem assuming that the random variables $X_1, X_2, ..., X_n$ are continuous. If the random variables are not continuous, then the proof follows exactly in the same manner if one replaces the integrals by summations. Since

$$\begin{aligned} &E\left(\prod_{i=1}^{n} u_i(X_i)\right) \\ &= E(u_1(X_1)\cdots u_n(X_n)) \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} u_1(x_1)\cdots u_n(x_n) f(x_1, ..., x_n)dx_1\cdots dx_n \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} u_1(x_1)\cdots u_n(x_n) f_1(x_1)\cdots f_n(x_n)dx_1\cdots dx_n \\ &= \int_{-\infty}^{\infty} u_1(x_1)f_1(x_1)dx_1\cdots\int_{-\infty}^{\infty} u_n(x_n)f_n(x_n)dx_n \\ &= E\left(u_1(X_1)\right)\cdots E\left(u_n(X_n)\right) \\ &= \prod_{i=1}^{n} E\left(u_i(X_i)\right), \end{aligned}$$

the proof of the theorem is now complete.

**Example 13.3.** Let $X$ and $Y$ be two random variables with the joint density

$$f(x,y) = \begin{cases} e^{-(x+y)} & \text{for } 0 < x,\, y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of the continuous random variable $Z = X^2Y^2 + XY^2 + X^2 + X$?

**Answer:** Since

$$\begin{aligned} f(x,y) &= e^{-(x+y)} \\ &= e^{-x}\, e^{-y} \\ &= f_1(x)\, f_2(y), \end{aligned}$$

the random variables $X$ and $Y$ are mutually independent. Hence, the expected value of $X$ is

$$\begin{aligned} E(X) &= \int_0^\infty x\, f_1(x)\, dx \\ &= \int_0^\infty x e^{-x}\, dx \\ &= \Gamma(2) \\ &= 1. \end{aligned}$$

Similarly, the expected value of $X^2$ is given by

$$\begin{aligned} E\left(X^2\right) &= \int_0^\infty x^2\, f_1(x)\, dx \\ &= \int_0^\infty x^2 e^{-x}\, dx \\ &= \Gamma(3) \\ &= 2. \end{aligned}$$

Since the marginals of $X$ and $Y$ are same, we also get $E(Y) = 1$ and $E(Y^2) = 2$. Further, by Theorem 13.1, we get

$$\begin{aligned} E\left[Z\right] &= E\left[X^2Y^2 + XY^2 + X^2 + X\right] \\ &= E\left[\left(X^2 + X\right)\, \left(Y^2 + 1\right)\right] \\ &= E\left[X^2 + X\right]\, E\left[Y^2 + 1\right] \qquad \text{(by Theorem 13.1)} \\ &= \left(E\left[X^2\right] + E\left[X\right]\right)\, \left(E\left[Y^2\right] + 1\right) \\ &= (2+1)\, (2+1) \\ &= 9. \end{aligned}$$

**Theorem 13.2.** If $X_1, X_2, ..., X_n$ are mutually independent random variables with respective means $\mu_1$, $\mu_2$, ..., $\mu_n$ and variances $\sigma_1^2$, $\sigma_2^2$, ..., $\sigma_n^2$, then the mean and variance of $Y = \sum_{i=1}^{n} a_i X_i$, where $a_1$, $a_2$, ..., $a_n$ are real constants, are given by

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \qquad \text{and} \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

**Proof:** First we show that $\mu_Y = \sum_{i=1}^{n} a_i \mu_i$. Since

$$\begin{aligned} \mu_Y &= E(Y) \\ &= E\left(\sum_{i=1}^{n} a_i X_i\right) \\ &= \sum_{i=1}^{n} a_i E(X_i) \\ &= \sum_{i=1}^{n} a_i \mu_i \end{aligned}$$

we have asserted result. Next we show $\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$. Since $Cov(X_i, X_j) = 0$ for $i \neq j$, we have

$$\begin{aligned} \sigma_Y^2 &= Var(Y) \\ &= Var\left(a_i X_i\right) \\ &= \sum_{i=1}^{n} a_i^2 Var\left(X_i\right) \\ &= \sum_{i=1}^{n} a_i^2 \sigma_i^2. \end{aligned}$$

This completes the proof of the theorem.

**Example 13.4.** Let the independent random variables $X_1$ and $X_2$ have means $\mu_1 = -4$ and $\mu_2 = 3$, respectively and variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$. What are the mean and variance of $Y = 3X_1 - 2X_2$?

**Answer:** The mean of $Y$ is

$$\begin{aligned} \mu_Y &= 3\mu_1 - 2\mu_2 \\ &= 3(-4) - 2(3) \\ &= -18. \end{aligned}$$

Similarly, the variance of $Y$ is

$$\begin{aligned}\sigma_Y^2 &= (3)^2\,\sigma_1^2 + (-2)^2\,\sigma_2^2\\ &= 9\,\sigma_1^2 + 4\,\sigma_2^2\\ &= 9(4) + 4(9)\\ &= 72.\end{aligned}$$

**Example 13.5.** Let $X_1, X_2, ..., X_{50}$ be a random sample of size 50 from a distribution with density

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

What are the mean and variance of the sample mean $\overline{X}$?

**Answer:** Since the distribution of the population $X$ is exponential, the mean and variance of $X$ are given by

$$\mu_X = \theta, \qquad \text{and} \qquad \sigma_X^2 = \theta^2.$$

Thus, the mean of the sample mean is

$$\begin{aligned}E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + \cdots + X_{50}}{50}\right)\\ &= \frac{1}{50}\,\sum_{i=1}^{50} E\left(X_i\right)\\ &= \frac{1}{50}\,\sum_{i=1}^{50} \theta\\ &= \frac{1}{50}\,50\,\theta = \theta.\end{aligned}$$

The variance of the sample mean is given by

$$\begin{aligned}Var\left(\overline{X}\right) &= Var\left(\sum_{i=1}^{50} \frac{1}{50}\, X_i\right)\\ &= \sum_{i=1}^{50} \left(\frac{1}{50}\right)^2\, \sigma_{X_i}^2\\ &= \sum_{i=1}^{50} \left(\frac{1}{50}\right)^2\, \theta^2\\ &= 50\,\left(\frac{1}{50}\right)^2\, \theta^2\\ &= \frac{\theta^2}{50}.\end{aligned}$$

**Theorem 13.3.** If $X_1, X_2, ..., X_n$ are independent random variables with respective moment generating functions $M_{X_i}(t)$, $i = 1, 2, ..., n$, then the moment generating function of $Y = \sum_{i=1}^{n} a_i X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t).$$

**Proof:** Since

$$\begin{aligned} M_Y(t) &= M_{\sum_{i=1}^{n} a_i X_i}(t) \\ &= \prod_{i=1}^{n} M_{a_i X_i}(t) \\ &= \prod_{i=1}^{n} M_{X_i}(a_i t) \end{aligned}$$

we have the asserted result and the proof of the theorem is now complete.

**Example 13.6.** Let $X_1, X_2, ..., X_{10}$ be the observations from a random sample of size 10 from a distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

What is the moment generating function of the sample mean?

**Answer:** The density of the population $X$ is a standard normal. Hence, the moment generating function of each $X_i$ is

$$M_{X_i}(t) = e^{\frac{1}{2}t^2}, \qquad i = 1, 2, ..., 10.$$

The moment generating function of the sample mean is

$$\begin{aligned} M_{\overline{X}}(t) &= M_{\sum_{i=1}^{10} \frac{1}{10} X_i}(t) \\ &= \prod_{i=1}^{10} M_{X_i}\left(\frac{1}{10}\, t\right) \\ &= \prod_{i=1}^{10} e^{\frac{t^2}{200}} \\ &= \left[e^{\frac{t^2}{200}}\right]^{10} = e^{\left(\frac{1}{10}\,\frac{t^2}{2}\right)}. \end{aligned}$$

Hence $\overline{X} \sim N\left(0, \frac{1}{10}\right)$.

The last example tells us that if we take a sample of any size from a standard normal population, then the sample mean also has a normal distribution.

The following theorem says that a linear combination of random variables with normal distributions is again normal.

**Theorem 13.4.** If $X_1, X_2, ..., X_n$ are mutually independent random variables such that

$$X_i \sim N\left(\mu_i,\, \sigma_i^2\right), \qquad i = 1, 2, ..., n.$$

Then the random variable $Y = \sum_{i=1}^{n} a_i X_i$ is a normal random variable with mean

$$\mu_Y = \sum_{i=1}^{n} a_i\, \mu_i \qquad \text{and} \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2\, \sigma_i^2,$$

that is $Y \sim N\left(\sum_{i=1}^{n} a_i\mu_i,\, \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$.

**Proof:** Since each $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, the moment generating function of each $X_i$ is given by

$$M_{X_i}(t) = e^{\mu_i t + \frac{1}{2}\sigma_i^2 t^2}.$$

Hence using Theorem 13.3, we have

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(a_i t) \\ &= \prod_{i=1}^{n} e^{a_i\mu_i t + \frac{1}{2}a_i^2\sigma_i^2 t^2} \\ &= e^{\sum_{i=1}^{n} a_i\mu_i t + \frac{1}{2}\sum_{i=1}^{n} a_i^2\sigma_i^2 t^2}. \end{aligned}$$

Thus the random variable $Y \sim N\left(\sum_{i=1}^{n} a_i\mu_i,\, \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$. The proof of the theorem is now complete.

**Example 13.7.** Let $X_1, X_2, ..., X_n$ be the observations from a random sample of size $n$ from a normal distribution with mean $\mu$ and variance $\sigma^2 > 0$. What are the mean and variance of the sample mean $\overline{X}$?

**Answer:** The expected value (or mean) of the sample mean is given by

$$\begin{aligned} E\left(\overline{X}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned}$$

Similarly, the variance of the sample mean is

$$Var\left(\overline{X}\right) = \sum_{i=1}^{n} Var\left(\frac{X_i}{n}\right) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

This example along with the previous theorem says that if we take a random sample of size $n$ from a normal population with mean $\mu$ and variance $\sigma^2$, then the sample mean is also normal with mean $\mu$ and variance $\frac{\sigma^2}{n}$, that is $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

**Example 13.8.** Let $X_1, X_2, ..., X_{64}$ be a random sample of size 64 from a normal distribution with $\mu = 50$ and $\sigma^2 = 16$. What are $P\left(49 < X_8 < 51\right)$ and $P\left(49 < \overline{X} < 51\right)$?

**Answer:** Since $X_8 \sim N(50, 16)$, we get

$$\begin{aligned} P\left(49 < X_8 < 51\right) &= P\left(49 - 50 < X_8 - 50 < 51 - 50\right) \\ &= P\left(\frac{49-50}{4} < \frac{X_8 - 50}{4} < \frac{51-50}{4}\right) \\ &= P\left(-\frac{1}{4} < \frac{X_8 - 50}{4} < \frac{1}{4}\right) \\ &= P\left(-\frac{1}{4} < Z < \frac{1}{4}\right) \\ &= 2P\left(Z < \frac{1}{4}\right) - 1 \\ &= 0.1974 \qquad \text{(from normal table).} \end{aligned}$$

By the previous theorem, we see that $\overline{X} \sim N\left(50,\, \frac{16}{64}\right)$. Hence

$$\begin{aligned} P\left(49 < \overline{X} < 51\right) &= P\left(49 - 50 < \overline{X} - 50 < 51 - 50\right) \\ &= P\left(\frac{49-50}{\sqrt{\frac{16}{64}}} < \frac{\overline{X}-50}{\sqrt{\frac{16}{64}}} < \frac{51-50}{\sqrt{\frac{16}{64}}}\right) \\ &= P\left(-2 < \frac{\overline{X}-50}{\sqrt{\frac{16}{64}}} < 2\right) \\ &= P\left(-2 < Z < 2\right) \\ &= 2P\left(Z < 2\right) - 1 \\ &= 0.9544 \qquad \text{(from normal table).} \end{aligned}$$

This example tells us that $\overline{X}$ has a greater probability of falling in an interval containing $\mu$, than a single observation, say $X_8$ (or in general any $X_i$).

**Theorem 13.5.** Let the distributions of the random variables $X_1, X_2, ..., X_n$ be $\chi^2(r_1)$, $\chi^2(r_2)$, ..., $\chi^2(r_n)$, respectively. If $X_1, X_2, ..., X_n$ are mutually independent, then $Y = X_1 + X_2 + \cdots + X_n \sim \chi^2\left(\sum_{i=1}^n r_i\right)$.

**Proof:** Since each $X_i \sim \chi^2(r_i)$, the moment generating function of each $X_i$ is given by

$$M_{X_i}(t) = (1-2t)^{-\frac{r_i}{2}}.$$

By Theorem 13.3, we have

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n (1-2t)^{-\frac{r_i}{2}} = (1-2t)^{-\frac{1}{2}\sum_{i=1}^n r_i}.$$

Hence $Y \sim \chi^2\left(\sum_{i=1}^n r_i\right)$ and the proof of the theorem is now complete.

The proof of the following theorem is an easy consequence of Theorem 13.5 and we leave the proof to the reader.

**Theorem 13.6.** If $Z_1, Z_2, ..., Z_n$ are mutually independent and each one is standard normal, then $Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$, that is the sum is chi-square with $n$ degrees of freedom.

The following theorem is very useful in mathematical statistics and its proof is beyond the scope of this introductory book.

**Theorem 13.7.** If $X_1, X_2, ..., X_n$ are observations of a random sample of size $n$ from the normal distribution $N\left(\mu,\, \sigma^2\right)$, then the sample mean $\overline{X} =$

$\frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ have the following properties:

(A) $\overline{X}$ and $S^2$ are independent, and
(B) $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$.

**Remark 13.2.** At first sight the statement (A) might seem odd since the sample mean $\overline{X}$ occurs explicitly in the definition of the sample variance $S^2$. This remarkable independence of $\overline{X}$ and $S^2$ is a unique property that distinguishes normal distribution from all other probability distributions.

**Example 13.9.** Let $X_1, X_2, ..., X_n$ denote a random sample from a normal distribution with variance $\sigma^2 > 0$. If the first percentile of the statistics $W = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$ is 1.24, where $\overline{X}$ denotes the sample mean, what is the sample size $n$?

**Answer:**

$$\begin{aligned}\frac{1}{100} &= P\left(W \leq 1.24\right)\\ &= P\left(\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2} \leq 1.24\right)\\ &= P\left((n-1)\frac{S^2}{\sigma^2} \leq 1.24\right)\\ &= P\left(\chi^2(n-1) \leq 1.24\right).\end{aligned}$$

Thus from $\chi^2$-table, we get

$$n - 1 = 7$$

and hence the sample size $n$ is 8.

**Example 13.10.** Let $X_1, X_2, ..., X_4$ be a random sample from a normal distribution with unknown mean and variance equal to 9. Let $S^2 = \frac{1}{3}\sum_{i=1}^{4}\left(X_i - \overline{X}\right)$. If $P\left(S^2 \leq k\right) = 0.05$, then what is $k$?

**Answer:**

$$\begin{aligned}0.05 &= P\left(S^2 \leq k\right)\\ &= P\left(\frac{3S^2}{9} \leq \frac{3}{9}k\right)\\ &= P\left(\chi^2(3) \leq \frac{3}{9}k\right).\end{aligned}$$

From $\chi^2$-table with 3 degrees of freedom, we get

$$\frac{3}{9}k = 0.35$$

and thus the constant $k$ is given by

$$k = 3(0.35) = 1.05.$$

### 13.2. Laws of Large Numbers

In this section, we mainly examine the weak law of large numbers. The weak law of large numbers states that if $X_1, X_2, ..., X_n$ is a random sample of size $n$ from a population $X$ with mean $\mu$, then the sample mean $\overline{X}$ rarely deviates from the population mean $\mu$ when the sample size $n$ is very large. In other words, the sample mean $\overline{X}$ converges in probability to the population mean $\mu$. We begin this section with a result known as Markov inequality which is needed to establish the weak law of large numbers.

**Theorem 13.8 (Markov Inequality).** Suppose $X$ is a nonnegative random variable with mean $E(X)$. Then

$$P(X \geq t) \leq \frac{E(X)}{t}$$

for all $t > 0$.

**Proof:** We assume the random variable $X$ is continuous. If $X$ is not continuous, then a proof can be obtained for this case by replacing the integrals with summations in the following proof. Since

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_{-\infty}^{t} x f(x) dx + \int_{t}^{\infty} x f(x) dx \\
&\geq \int_{t}^{\infty} x f(x) dx \\
&\geq \int_{t}^{\infty} t f(x) dx \qquad \text{because } x \in [t, \infty) \\
&= t \int_{t}^{\infty} f(x) dx \\
&= t\, P(X \geq t),
\end{aligned}$$

we see that

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

This completes the proof of the theorem.

In Theorem 4.4 of the chapter 4, Chebychev inequality was treated. Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma$. Then Chebychev inequality says that

$$P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

for any nonzero positive constant $k$. This result can be obtained easily using Theorem 13.8 as follows. By Markov inequality, we have

$$P((X-\mu)^2 \geq t^2) \leq \frac{E((X-\mu)^2)}{t^2}$$

for all $t > 0$. Since the events $(X-\mu)^2 \geq t^2$ and $|X-\mu| \geq t$ are same, we get

$$P((X-\mu)^2 \geq t^2) = P(|X-\mu| \geq t) \leq \frac{E((X-\mu)^2)}{t^2}$$

for all $t > 0$. Hence

$$P(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Letting $t = k\sigma$ in the above equality, we see that

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Hence

$$1 - P(|X-\mu| < k\sigma) \leq \frac{1}{k^2}.$$

The last inequality yields the Chebychev inequality

$$P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Now we are ready to treat the weak law of large numbers.

**Theorem 13.9.** Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$ for $i = 1, 2, ..., \infty$. Then

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| \geq \varepsilon) = 0$$

for every $\varepsilon$. Here $\overline{S}_n$ denotes $\frac{X_1+X_2+\cdots+X_n}{n}$.

**Proof:** By Theorem 13.2 (or Example 13.7) we have

$$E(\overline{S}_n) = \mu \qquad \text{and} \qquad Var(\overline{S}_n) = \frac{\sigma^2}{n}.$$

By Chebychev's inequality

$$P(|\overline{S}_n - E(\overline{S}_n)| \geq \varepsilon) \leq \frac{Var(\overline{S}_n)}{\varepsilon^2}$$

for $\varepsilon > 0$. Hence

$$P(|\overline{S}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\,\varepsilon^2}.$$

Taking the limit as $n$ tends to infinity, we get

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\,\varepsilon^2}$$

which yields

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| \geq \varepsilon) = 0$$

and the proof of the theorem is now complete.

It is possible to prove the weak law of large numbers assuming only $E(X)$ to exist and finite but the proof is more involved.

The weak law of large numbers says that the sequence of sample means $\{\overline{S}_n\}_{n=1}^{\infty}$ from a population $X$ stays close to the population mean $E(X)$ most of the time. Let us consider an experiment that consists of tossing a coin infinitely many times. Let $X_i$ be 1 if the $i^{\text{th}}$ toss results in a Head, and 0 otherwise. The weak law of large numbers says that

$$\overline{S}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \frac{1}{2} \qquad \text{as} \quad n \to \infty \tag{13.0}$$

but it is easy to come up with sequences of tosses for which (13.0) is false:

H H H H H H H H H H H H $\cdots$ $\cdots$
H H T H H T H H T H H T $\cdots$ $\cdots$

The strong law of large numbers (Theorem 13.11) states that the set of "bad sequences" like the ones given above has probability zero.

Note that the assertion of Theorem 13.9 for any $\varepsilon > 0$ can also be written as

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| < \varepsilon) = 1.$$

The type of convergence we saw in the weak law of large numbers is not the type of convergence discussed in calculus. This type of convergence is called convergence in probability and defined as follows.

**Definition 13.1.** Suppose $X_1, X_2, ...$ is a sequence of random variables defined on a sample space $S$. The sequence *converges in probability* to the random variable $X$ if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

In view of the above definition, the weak law of large numbers states that the sample mean $\overline{X}$ converges in probability to the population mean $\mu$.

The following theorem is known as the Bernoulli law of large numbers and is a special case of the weak law of large numbers.

**Theorem 13.10.** Let $X_1, X_2, ...$ be a sequence of independent and identically distributed Bernoulli random variables with probability of success $p$. Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{S}_n - p| < \varepsilon) = 1$$

where $\overline{S}_n$ denotes $\frac{X_1+X_2+\cdots+X_n}{n}$.

The fact that the relative frequency of occurrence of an event $E$ is very likely to be close to its probability $P(E)$ for large $n$ can be derived from the weak law of large numbers. Consider a repeatable random experiment repeated large number of time independently. Let $X_i = 1$ if $E$ occurs on the $i^{\text{th}}$ repetition and $X_i = 0$ if $E$ does not occur on $i^{\text{th}}$ repetition. Then

$$\mu = E(X_i) = 1 \cdot P(E) + 0 \cdot P(E) = P(E) \qquad \text{for } i = 1, 2, 3, ...$$

and

$$X_1 + X_2 + \cdots + X_n = N(E)$$

where $N(E)$ denotes the number of times $E$ occurs. Hence by the weak law of large numbers, we have

$$\begin{aligned}\lim_{n\to\infty} P\left(\left|\frac{N(E)}{n} - P(E)\right| \geq \varepsilon\right) &= \lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \\ &= \lim_{n\to\infty} P\left(\left|\overline{S}_n - \mu\right| \geq \varepsilon\right) \\ &= 0.\end{aligned}$$

Hence, for large $n$, the relative frequency of occurrence of the event $E$ is very likely to be close to its probability $P(E)$.

Now we present the strong law of large numbers without a proof.

**Theorem 13.11.** Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$ for $i = 1, 2, ..., \infty$. Then

$$P\left(\lim_{n\to\infty} \overline{S}_n = \mu\right) = 1$$

for every $\varepsilon > 0$. Here $\overline{S}_n$ denotes $\frac{X_1+X_2+\cdots+X_n}{n}$.

The type convergence in Theorem 13.11 is called almost sure convergence. The notion of almost sure convergence is defined as follows.

**Definition 13.2** Suppose the random variable $X$ and the sequence $X_1, X_2, ...,$ of random variables are defined on a sample space $S$. The sequence $X_n(w)$ *converges almost surely* to $X(w)$ if

$$P\left(\left\{w \in S \mid \lim_{n\to\infty} X_n(w) = X(w)\right\}\right) = 1.$$

It can be shown that the convergence in probability implies the almost sure convergence but not the converse.

### 13.3. The Central Limit Theorem

Consider a random sample of measurement $\{X_i\}_{i=1}^n$. The $X_i$'s are identically distributed and their common distribution is the distribution of the population. We have seen that if the population distribution is normal, then the sample mean $\overline{X}$ is also normal. More precisely, if $X_1, X_2, ..., X_n$ is a random sample from a normal distribution with density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

then

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The central limit theorem (also known as Lindeberg-Levy Theorem) states that even though the population distribution may be far from being normal, still for large sample size $n$, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size. Mathematically this can be stated as follows.

**Theorem 13.12 (Central Limit Theorem).** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with mean $\mu$ and variance $\sigma^2 < \infty$, then the limiting distribution of

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is standard normal, that is $Z_n$ converges in distribution to $Z$ where $Z$ denotes a standard normal random variable.

The type of convergence used in the central limit theorem is called the convergence in distribution and is defined as follows.

**Definition 13.3.** Suppose $X$ is a random variable with cumulative density function $F(x)$ and the sequence $X_1, X_2, ...$ of random variables with cumulative density functions $F_1(x), F_2(x), ...$, respectively. The sequence $X_n$ *converges in distribution* to $X$ if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all values $x$ at which $F(x)$ is continuous. The distribution of $X$ is called the *limiting distribution* of $X_n$.

Whenever a sequence of random variables $X_1, X_2, ...$ converges in distribution to the random variable $X$, it will be denoted by $X_n \xrightarrow{d} X$.

**Example 13.11.** Let $Y = X_1 + X_2 + \cdots + X_{15}$ be the sum of a random sample of size 15 from the distribution whose density function is

$$f(x) = \begin{cases} \frac{3}{2}\,x^2 & \text{if } -1 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the approximate value of $P(-0.3 \leq Y \leq 1.5)$ when one uses the central limit theorem?

**Answer:** First, we find the mean $\mu$ and variance $\sigma^2$ for the density function $f(x)$. The mean for this distribution is given by

$$\begin{aligned} \mu &= \int_{-1}^{1} \frac{3}{2} x^3\, dx \\ &= \frac{3}{2}\left[\frac{x^4}{4}\right]_{-1}^{1} \\ &= 0. \end{aligned}$$

Hence the variance of this distribution is given by

$$\begin{aligned} Var(X) &= E(X^2) - [\,E(X)\,]^2 \\ &= \int_{-1}^{1} \frac{3}{2}x^4\,dx \\ &= \frac{3}{2}\left[\frac{x^5}{5}\right]_{-1}^{1} \\ &= \frac{3}{5} \\ &= 0.6. \end{aligned}$$

$$\begin{aligned} P(-0.3 \le Y \le 1.5) &= P(-0.3 - 0 \le Y - 0 \le 1.5 - 0) \\ &= P\left(\frac{-0.3}{\sqrt{15(0.6)}} \le \frac{Y-0}{\sqrt{15(0.6)}} \le \frac{1.5}{\sqrt{15(0.6)}}\right) \\ &= P(-0.10 \le Z \le 0.50) \\ &= P(Z \le 0.50) + P(Z \le 0.10) - 1 \\ &= 0.6915 + 0.5398 - 1 \\ &= 0.2313. \end{aligned}$$

**Example 13.12.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n = 25$ from a population that has a mean $\mu = 71.43$ and variance $\sigma^2 = 56.25$. Let $\overline{X}$ be the sample mean. What is the probability that the sample mean is between 68.91 and 71.97?

**Answer:** The mean of $\overline{X}$ is given by $E\left(\overline{X}\right) = 71.43$. The variance of $\overline{X}$ is given by

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n} = \frac{56.25}{25} = 2.25.$$

In order to find the probability that the sample mean is between 68.91 and 71.97, we need the distribution of the population. However, the population distribution is unknown. Therefore, we use the central limit theorem. The central limit theorem says that $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ as $n$ approaches infinity. Therefore

$$\begin{aligned} &P\left(68.91 \le \overline{X} \le 71.97\right) \\ &= \left(\frac{68.91 - 71.43}{\sqrt{2.25}} \le \frac{\overline{X} - 71.43}{\sqrt{2.25}} \le \frac{71.97 - 71.43}{\sqrt{2.25}}\right) \\ &= P\left(-0.68 \le W \le 0.36\right) \\ &= P\left(W \le 0.36\right) + P\left(W \le 0.68\right) - 1 \\ &= 0.5941. \end{aligned}$$

**Example 13.13.** Light bulbs are installed successively into a socket. If we assume that each light bulb has a mean life of 2 months with a standard deviation of 0.25 months, what is the probability that 40 bulbs last at least 7 years?

**Answer:** Let $X_i$ denote the life time of the $i^{\text{th}}$ bulb installed. The 40 light bulbs last a total time of

$$S_{40} = X_1 + X_2 + \cdots + X_{40}.$$

By the central limit theorem

$$\frac{\sum_{i=1}^{40} X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1) \qquad \text{as} \qquad n \to \infty.$$

Thus

$$\frac{S_{40} - (40)(2)}{\sqrt{(40)(0.25)^2}} \sim N(0,1).$$

That is

$$\frac{S_{40} - 80}{1.581} \sim N(0,1).$$

Therefore

$$\begin{aligned} &P\left(S_{40} \geq 7(12)\right) \\ &\quad = P\left(\frac{S_{40} - 80}{1.581} \geq \frac{84 - 80}{1.581}\right) \\ &\quad = P\left(Z \geq 2.530\right) \\ &\quad = 0.0057. \end{aligned}$$

**Example 13.14.** Light bulbs are installed into a socket. Assume that each has a mean life of 2 months with standard deviation of 0.25 month. How many bulbs $n$ should be bought so that one can be 95% sure that the supply of $n$ bulbs will last 5 years?

**Answer:** Let $X_i$ denote the life time of the $i^{\text{th}}$ bulb installed. The $n$ light bulbs last a total time of

$$S_n = X_1 + X_2 + \cdots + X_n.$$

The total average life span $S_n$ has

$$E\left(S_n\right) = 2n \qquad \text{and} \qquad Var(S_n) = \frac{n}{16}.$$

By the central limit theorem, we get

$$\frac{S_n - E(S_n)}{\frac{\sqrt{n}}{4}} \sim N(0,1).$$

Thus, we seek $n$ such that

$$\begin{aligned} 0.95 &= P(S_n \geq 60) \\ &= P\left(\frac{S_n - 2n}{\frac{\sqrt{n}}{4}} \geq \frac{60 - 2n}{\frac{\sqrt{n}}{4}}\right) \\ &= P\left(Z \geq \frac{240 - 8n}{\sqrt{n}}\right) \\ &= 1 - P\left(Z \leq \frac{240 - 8n}{\sqrt{n}}\right). \end{aligned}$$

From the standard normal table, we get

$$\frac{240 - 8n}{\sqrt{n}} = -1.645$$

which implies

$$1.645\sqrt{n} + 8n - 240 = 0.$$

Solving this quadratic equation for $\sqrt{n}$, we get

$$\sqrt{n} = -5.375 \quad \text{or} \quad 5.581.$$

Thus $n = 31.15$. So we should buy 32 bulbs.

**Example 13.15.** American Airlines claims that the average number of people who pay for in-flight movies, when the plane is fully loaded, is 42 with a standard deviation of 8. A sample of 36 fully loaded planes is taken. What is the probability that fewer than 38 people paid for the in-flight movies?

**Answer:** Here, we like to find $P(\overline{X} < 38)$. Since, we do not know the distribution of $\overline{X}$, we will use the central limit theorem. We are given that the population mean is $\mu = 42$ and population standard deviation is $\sigma = 8$. Moreover, we are dealing with sample of size $n = 36$. Thus

$$\begin{aligned} P(\overline{X} < 38) &= P\left(\frac{\overline{X} - 42}{\frac{8}{6}} < \frac{38 - 42}{\frac{8}{6}}\right) \\ &= P(Z < -3) \\ &= 1 - P(Z < 3) \\ &= 1 - 0.9987 \\ &= 0.0013. \end{aligned}$$

Since we have not yet seen the proof of the central limit theorem, first let us go through some examples to see the main idea behind the proof of the central limit theorem. Later, at the end of this section a proof of the central limit theorem will be given. We know from the central limit theorem that if $X_1, X_2, ..., X_n$ is a random sample of size $n$ from a distribution with mean $\mu$ and variance $\sigma^2$, then

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} Z \sim N(0, 1) \qquad \text{as} \qquad n \to \infty.$$

However, the above expression is *not equivalent* to

$$\overline{X} \xrightarrow{d} Z \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{as} \qquad n \to \infty$$

as the following example shows.

**Example 13.16.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a gamma distribution with parameters $\theta = 1$ and $\alpha = 1$. What is the distribution of the sample mean $\overline{X}$? Also, what is the limiting distribution of $\overline{X}$ as $n \to \infty$?

**Answer:** Since, each $X_i \sim GAM(1, 1)$, the probability density function of each $X_i$ is given by

$$f(x) = \begin{cases} e^{-x} & \text{if } \ x \geq 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

and hence the moment generating function of each $X_i$ is

$$M_{X_i}(t) = \frac{1}{1-t}.$$

First we determine the moment generating function of the sample mean $\overline{X}$, and then examine this moment generating function to find the probability distribution of $\overline{X}$. Since

$$\begin{aligned} M_{\overline{X}}(t) &= M_{\frac{1}{n}\sum_{i=1}^{n} X_i}(t) \\ &= \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) \\ &= \prod_{i=1}^{n} \frac{1}{\left(1 - \frac{t}{n}\right)} \\ &= \frac{1}{\left(1 - \frac{t}{n}\right)^n}, \end{aligned}$$

therefore $\overline{X} \sim GAM\left(\frac{1}{n},\, n\right)$.

Next, we find the limiting distribution of $\overline{X}$ as $n \to \infty$. This can be done again by finding the limiting moment generating function of $\overline{X}$ and identifying the distribution of $\overline{X}$. Consider

$$\begin{aligned}\lim_{n\to\infty} M_{\overline{X}}(t) &= \lim_{n\to\infty} \frac{1}{\left(1-\frac{t}{n}\right)^n} \\ &= \frac{1}{\lim_{n\to\infty}\left(1-\frac{t}{n}\right)^n} \\ &= \frac{1}{e^{-t}} \\ &= e^t.\end{aligned}$$

Thus, the sample mean $\overline{X}$ has a degenerate distribution, that is all the probability mass is concentrated at one point of the space of $\overline{X}$.

**Example 13.17.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a gamma distribution with parameters $\theta = 1$ and $\alpha = 1$. What is the distribution of

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{as} \qquad n \to \infty$$

where $\mu$ and $\sigma$ are the population mean and variance, respectively?

**Answer:** From Example 13.7, we know that

$$M_{\overline{X}}(t) = \frac{1}{\left(1-\frac{t}{n}\right)^n}.$$

Since the population distribution is gamma with $\theta = 1$ and $\alpha = 1$, the population mean $\mu$ is 1 and population variance $\sigma^2$ is also 1. Therefore

$$\begin{aligned}M_{\frac{\overline{X}-1}{\frac{1}{\sqrt{n}}}}(t) &= M_{\sqrt{n}\overline{X}-\sqrt{n}}(t) \\ &= e^{-\sqrt{n}t}\, M_{\overline{X}}\left(\sqrt{n}\,t\right) \\ &= e^{-\sqrt{n}t}\, \frac{1}{\left(1-\frac{\sqrt{n}t}{n}\right)^n} \\ &= \frac{1}{e^{\sqrt{n}t}\,\left(1-\frac{t}{\sqrt{n}}\right)^n}.\end{aligned}$$

The limiting moment generating function can be obtained by taking the limit of the above expression as $n$ tends to infinity. That is,

$$\begin{aligned}\lim_{n\to\infty} M_{\frac{\overline{X}-1}{\frac{1}{\sqrt{n}}}}(t) &= \lim_{n\to\infty} \frac{1}{e^{\sqrt{n}t}\left(1-\frac{t}{\sqrt{n}}\right)^n} \\ &= e^{\frac{1}{2}t^2} \qquad \text{(using MAPLE)} \\ &= \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).\end{aligned}$$

The following theorem is used to prove the central limit theorem.

**Theorem 13.13 (Lévy Continuity Theorem).** Let $X_1, X_2, ...$ be a sequence of random variables with distribution functions $F_1(x), F_2(x), ...$ and moment generating functions $M_{X_1}(t), M_{X_2}(t), ...$, respectively. Let $X$ be a random variable with distribution function $F(x)$ and moment generating function $M_X(t)$. If for all $t$ in the open interval $(-h, h)$ for some $h > 0$

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t),$$

then at the points of continuity of $F(x)$

$$\lim_{n\to\infty} F_n(x) = F(x).$$

The proof of this theorem is beyond the scope of this book.

The following limit

$$\lim_{n\to\infty}\left[1+\frac{t}{n}+\frac{d(n)}{n}\right]^n = e^t, \qquad \text{if } \lim_{n\to\infty} d(n) = 0, \tag{13.1}$$

whose proof we leave it to the reader, can be established using advanced calculus. Here $t$ is independent of $n$.

Now we proceed to prove the central limit theorem assuming that the moment generating function of the population $X$ exists. Let $M_{X-\mu}(t)$ be the moment generating function of the random variable $X - \mu$. We denote $M_{X-\mu}(t)$ as $M(t)$ when there is no danger of confusion. Then

$$\left.\begin{aligned} M(0) &= 1, \\ M'(0) &= E(X-\mu) = E(X) - \mu = \mu - \mu = 0, \\ M''(0) &= E\left((X-\mu)^2\right) = \sigma^2. \end{aligned}\right\} \tag{13.2}$$

By Taylor series expansion of $M(t)$ about 0, we get

$$M(t) = M(0) + M'(0)\,t + \frac{1}{2}\,M''(\eta)\,t^2$$

where $\eta \in (0, t)$. Hence using (13.2), we have

$$\begin{aligned} M(t) &= 1 + \frac{1}{2}\,M''(\eta)\,t^2 \\ &= 1 + \frac{1}{2}\,\sigma^2\,t^2 + \frac{1}{2}\,M''(\eta)\,t^2 - \frac{1}{2}\,\sigma^2\,t^2 \\ &= 1 + \frac{1}{2}\,\sigma^2\,t^2 + \frac{1}{2}\,\left[M''(\eta) - \sigma^2\right]\,t^2. \end{aligned}$$

Now using $M(t)$ we compute the moment generating function of $Z_n$. Note that

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Hence

$$\begin{aligned} M_{Z_n}(t) &= \prod_{i=1}^{n} M_{X_i-\mu}\left(\frac{t}{\sigma\,\sqrt{n}}\right) \\ &= \prod_{i=1}^{n} M_{X-\mu}\left(\frac{t}{\sigma\,\sqrt{n}}\right) \\ &= \left[M\left(\frac{t}{\sigma\,\sqrt{n}}\right)\right]^n \\ &= \left[1 + \frac{t^2}{2\,n} + \frac{(M''(\eta) - \sigma^2)\,t^2}{2\,n\,\sigma^2}\right]^n \end{aligned}$$

for $0 < |\eta| < \frac{1}{\sigma\,\sqrt{n}}\,|t|$. Note that since $0 < |\eta| < \frac{1}{\sigma\,\sqrt{n}}\,|t|$, we have

$$\lim_{n\to\infty} \frac{t}{\sigma\,\sqrt{n}} = 0, \quad \lim_{n\to\infty} \eta = 0, \quad \text{and} \quad \lim_{n\to\infty} M''(\eta) - \sigma^2 = 0. \tag{13.3}$$

Letting

$$d(n) = \frac{(M''(\eta) - \sigma^2)\,t^2}{2\,\sigma^2}$$

and using (13.3), we see that $\lim\limits_{n\to\infty} d(n) = 0$, and

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2\,n} + \frac{d(n)}{n}\right]^n. \tag{13.4}$$

Using (13.1) we have

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \left[1 + \frac{t^2}{2\,n} + \frac{d(n)}{n}\right]^n = e^{\frac{1}{2}\,t^2}.$$

Hence by the Lévy continuity theorem, we obtain

$$\lim_{n\to\infty} F_n(x) = \Phi(x)$$

where $\Phi(x)$ is the cumulative density function of the standard normal distribution. Thus $Z_n \xrightarrow{d} Z$ and the proof of the theorem is now complete.

Now we give another proof of the central limit theorem using L'Hospital rule. This proof is essentially due to Tardiff (1981).

As before, let $Z_n = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$. Then $M_{Z_n}(t) = \left[ M\left(\frac{t}{\sigma\sqrt{n}}\right) \right]^n$ where $M(t)$ is the moment generating function of the random variable $X - \mu$. Hence from (13.2), we have $M(0) = 1,\ M'(0) = 0$, and $M''(0) = \sigma^2$. Now applying the L'Hospital rule twice we compute

$$\begin{aligned}
&\lim_{n\to\infty} M_{Z_n}(t) \\
&= \lim_{n\to\infty} \left[ M\left(\frac{t}{\sigma\sqrt{n}}\right) \right]^n \\
&= \lim_{n\to\infty} exp\left( n \ln\left( M\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\right) \\
&= \lim_{n\to\infty} exp\left( \frac{\ln\left( M\left(\frac{t}{\sigma\sqrt{n}}\right)\right)}{\frac{1}{n}} \right) \qquad \left(\frac{0}{0} \text{ form since M(0) = 1}\right) \\
&= \lim_{n\to\infty} exp\left( \frac{t}{2\,\sigma} \frac{M\left(\frac{t}{\sigma\sqrt{n}}\right)^{-1} M'\left(\frac{t}{\sigma\sqrt{n}}\right)\left(-\frac{1}{n\sqrt{n}}\right)}{-\frac{1}{n^2}} \right) \text{(L'Hospital rule)} \\
&= \lim_{n\to\infty} exp\left( \frac{t}{2\,\sigma} \frac{M\left(\frac{t}{\sigma\sqrt{n}}\right)^{-1} M'\left(\frac{t}{\sigma\sqrt{n}}\right)}{\frac{1}{\sqrt{n}}} \right) \left(\frac{0}{0} \text{ form since M'(0) = 0}\right) \\
&= \lim_{n\to\infty} exp\left( \frac{t^2}{2\,\sigma^2} \frac{M\left(\frac{t}{\sigma\sqrt{n}}\right) M''\left(\frac{t}{\sigma\sqrt{n}}\right) - \left\{M'\left(\frac{t}{\sigma\sqrt{n}}\right)\right\}^2}{M\left(\frac{t}{\sigma\sqrt{n}}\right)^2} \right) \\
&= exp\left( \frac{t^2}{2\,\sigma^2} \frac{M(0)\ M''(0) - \{M'(0)\}^2}{M(0)^2} \right) \\
&= exp\left( \frac{t^2}{2\,\sigma^2} \left[1 \cdot \sigma^2 - 0^2\right] \right) \\
&= exp\left( \frac{1}{2}\, t^2 \right).
\end{aligned}$$

Hence by the Lévy continuity theorem, we obtain

$$\lim_{n\to\infty} F_n(x) = \Phi(x)$$

where $\Phi(x)$ is the cumulative density function of the standard normal distribution. Thus as $n \to \infty$, the random variable $Z_n \xrightarrow{d} Z$, where $Z \sim N(0,1)$.

**Remark 13.3.** In contrast to the moment generating function, since the characteristic function of a random variable always exists, the original proof of the central limit theorem involved the characteristic function (see for example *An Introduction to Probability Theory and Its Applications, Volume II* by Feller). In 1988, Brown gave an elementary proof using very clever Taylor series expansions, where the use of the characteristic function has been avoided.

### 13.4. Order Statistics

Often, sample values such as the smallest, largest, or middle observation from a random sample provide important information. For example, the highest flood water or lowest winter temperature recorded during the last 50 years might be useful when planning for future emergencies. The median price of houses sold during the previous month might be useful for estimating the cost of living. The statistics highest, lowest or median are examples of order statistics.

**Definition 13.4.** Let $X_1, X_2, ..., X_n$ be observations from a random sample of size $n$ from a distribution $f(x)$. Let $X_{(1)}$ denote the smallest of $\{X_1, X_2, ..., X_n\}$, $X_{(2)}$ denote the second smallest of $\{X_1, X_2, ..., X_n\}$, and similarly $X_{(r)}$ denote the $r^{\text{th}}$ smallest of $\{X_1, X_2, ..., X_n\}$. Then the random variables $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are called the order statistics of the sample $X_1, X_2, ..., X_n$. In particular, $X_{(r)}$ is called the $r^{\text{th}}$-order statistic of $X_1, X_2, ..., X_n$.

The sample range, $R$, is the distance between the smallest and the largest observation. That is,

$$R = X_{(n)} - X_{(1)}.$$

This is an important statistic which is defined using order statistics.

The distribution of the order statistics are very important when one uses these in any statistical investigation. The next theorem gives the distribution of an order statistic.

**Theorem 13.14.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with density function $f(x)$. Then the probability density function of the $r^{\text{th}}$ order statistic, $X_{(r)}$, is

$$g(x) = \frac{n!}{(r-1)!\,(n-r)!}\,[F(x)]^{r-1}\,f(x)\,[1-F(x)]^{n-r},$$

where $F(x)$ denotes the cdf of $f(x)$.

**Proof:** We prove the theorem assuming $f(x)$ continuous. In the case $f(x)$ is discrete the proof has to be modified appropriately. Let $h$ be a positive real number and $x$ be an arbitrary point in the domain of $f$. Let us divide the real line into three segments, namely

$$\mathbb{R} = (-\infty,\, x) \bigcup [x,\, x+h) \bigcup [x+h,\, \infty).$$

The probability, say $p_1$, of a sample value falls into the first interval $(-\infty,\, x]$ and is given by

$$p_1 = \int_{-\infty}^{x} f(t)\,dt = F(x).$$

Similarly, the probability $p_2$ of a sample value falls into the second interval $[x,\, x+h)$ is

$$p_2 = \int_{x}^{x+h} f(t)\,dt = F(x+h) - F(x).$$

In the same token, we can compute the probability $p_3$ of a sample value which falls into the third interval

$$p_3 = \int_{x+h}^{\infty} f(t)\,dt = 1 - F(x+h).$$

Then the probability, $P_h(x)$, that $(r-1)$ sample values fall in the first interval, one falls in the second interval, and $(n-r)$ fall in the third interval is

$$P_h(x) = \binom{n}{r-1,\,1,\,n-r} p_1^{r-1}\,p_2^1\,p_3^{n-r} = \frac{n!}{(r-1)!\,(n-r)!}\,p_1^{r-1}\,p_2\,p_3^{n-r}.$$

Hence the probability density function $g(x)$ of the $r^{\text{th}}$ statistics is given by

$$\begin{aligned}
&g(x)\\
&= \lim_{h\to 0} \frac{P_h(x)}{h}\\
&= \lim_{h\to 0} \left[\frac{n!}{(r-1)!\,(n-r)!}\, p_1^{r-1}\ \frac{p_2}{h}\ p_3^{n-r}\right]\\
&= \frac{n!}{(r-1)!\,(n-r)!}\,[F(x)]^{r-1} \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} \lim_{h\to 0}[1-F(x+h)]^{n-r}\\
&= \frac{n!}{(r-1)!\,(n-r)!}\ [F(x)]^{r-1}\ F'(x)\ [1-F(x)]^{n-r}\\
&= \frac{n!}{(r-1)!\,(n-r)!}\ [F(x)]^{r-1}\ f(x)\ [1-F(x)]^{n-r}\,.
\end{aligned}$$

**Example 13.18.** Let $X_1, X_2$ be a random sample from a distribution with density function

$$f(x) = \begin{cases} e^{-x} & \text{for } 0 \le x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the density function of $Y = \min\{X_1, X_2\}$ where nonzero?

**Answer:** The cumulative distribution function of $f(x)$ is

$$\begin{aligned}
F(x) &= \int_0^x e^{-t}\,dt\\
&= 1 - e^{-x}
\end{aligned}$$

In this example, $n = 2$ and $r = 1$. Hence, the density of $Y$ is

$$\begin{aligned}
g(y) &= \frac{2!}{0!\,1!}\ [F(y)]^0\ f(y)\ [1-F(y)]\\
&= 2f(y)\ [1-F(y)]\\
&= 2\,e^{-y}\ \left(1-1+e^{-y}\right)\\
&= 2\,e^{-2y}.
\end{aligned}$$

**Example 13.19.** Let $Y_1 < Y_2 < \cdots < Y_6$ be the order statistics from a random sample of size 6 from a distribution with density function

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of $Y_6$?

**Answer:**

$$\begin{aligned} f(x) &= 2x \\ F(x) &= \int_0^x 2t\,dt \\ &= x^2. \end{aligned}$$

The density function of $Y_6$ is given by

$$\begin{aligned} g(y) &= \frac{6!}{5!\,0!}\ [F(y)]^5\ f(y) \\ &= 6\left(y^2\right)^5\ 2y \\ &= 12y^{11}. \end{aligned}$$

Hence, the expected value of $Y_6$ is

$$\begin{aligned} E\left(Y_6\right) &= \int_0^1 y\,g(y)\,dy \\ &= \int_0^1 y\,12y^{11}\,dy \\ &= \frac{12}{13}\ \left[y^{13}\right]_0^1 \\ &= \frac{12}{13}. \end{aligned}$$

**Example 13.20.** Let $X, Y$ and $Z$ be independent uniform random variables on the interval $(0, a)$. Let $W = \min\{X, Y, Z\}$. What is the expected value of $\left(1 - \frac{W}{a}\right)^2$?

**Answer:** The probability distribution of $X$ (or $Y$ or $Z$) is

$$f(x) = \begin{cases} \frac{1}{a} & \text{if } 0 < x < a \\ 0 & \text{otherwise.} \end{cases}$$

Thus the cumulative distribution of function of $f(x)$ is given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{x}{a} & \text{if } 0 < x < a \\ 1 & \text{if } x \geq a. \end{cases}$$

Since $W = \min\{X, Y, Z\}$, $W$ is the first order statistic of the random sample $X, Y, Z$. Thus, the density function of $W$ is given by

$$\begin{aligned} g(w) &= \frac{3!}{0!\,1!\,2!}\ [F(w)]^0\ f(w)\ [1-F(w)]^2 \\ &= 3f(w)\ [1-F(w)]^2 \\ &= 3\left(1-\frac{w}{a}\right)^2\ \left(\frac{1}{a}\right) \\ &= \frac{3}{a}\ \left(1-\frac{w}{a}\right)^2. \end{aligned}$$

Thus, the pdf of $W$ is given by

$$g(w) = \begin{cases} \frac{3}{a}\ \left(1-\frac{w}{a}\right)^2 & \text{if } 0 < w < a \\ \\ 0 & \text{otherwise.} \end{cases}$$

The expected value of $W$ is

$$\begin{aligned} & E\left[\left(1-\frac{W}{a}\right)^2\right] \\ &= \int_0^a \left(1-\frac{w}{a}\right)^2\ g(w)\,dw \\ &= \int_0^a \left(1-\frac{w}{a}\right)^2\ \frac{3}{a}\ \left(1-\frac{w}{a}\right)^2\ dw \\ &= \int_0^a \frac{3}{a}\ \left(1-\frac{w}{a}\right)^4\ dw \\ &= -\frac{3}{5}\ \left[\left(1-\frac{w}{a}\right)^5\right]_0^a \\ &= \frac{3}{5}. \end{aligned}$$

**Example 13.21.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with uniform distribution on the interval $[0, 1]$. What is the probability distribution of the sample range $W := X_{(n)} - X_{(1)}$?

**Answer:** To find the distribution of $W$, we need the joint distribution of the random variable $\left(X_{(n)}, X_{(1)}\right)$. The joint distribution of $\left(X_{(n)}, X_{(1)}\right)$ is given by

$$h(x_1, x_n) = n(n-1)f(x_1)f(x_n)\,[F(x_n) - F(x_1)]^{n-2},$$

where $x_n \geq x_1$ and $f(x)$ is the probability density function of $X$. To determine the probability distribution of the sample range $W$, we consider the transformation

$$\left.\begin{aligned} U &= X_{(1)} \\ W &= X_{(n)} - X_{(1)} \end{aligned}\right\}$$

which has an inverse

$$\left.\begin{aligned} X_{(1)} &= U \\ X_{(n)} &= U + W. \end{aligned}\right\}$$

The Jacobian of this transformation is

$$J = det \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = 1.$$

Hence the joint density of $(U, W)$ is given by

$$\begin{aligned} g(u,w) &= |J|\, h(x_1, x_n) \\ &= n(n-1) f(u) f(u+w) [F(u+w) - F(u)]^{n-2} \end{aligned}$$

where $w \geq 0$. Since $f(u)$ and $f(u+w)$ are simultaneously nonzero if $0 \leq u \leq 1$ and $0 \leq u + w \leq 1$. Hence $f(u)$ and $f(u+w)$ are simultaneously nonzero if $0 \leq u \leq 1 - w$. Thus, the probability of $W$ is given by

$$\begin{aligned} j(w) &= \int_{-\infty}^{\infty} g(u,w)\, du \\ &= \int_{-\infty}^{\infty} n(n-1) f(u) f(u+w) [F(u+w) - F(u)]^{n-2}\, du \\ &= n(n-1)\, w^{n-2} \int_0^{1-w} du \\ &= n(n-1)\,(1-w)\, w^{n-2} \end{aligned}$$

where $0 \leq w \leq 1$.

### 13.5. Sample Percentiles

The sample median, $M$, is a number such that approximately one-half of the observations are less than $M$ and one-half are greater than $M$.

**Definition 13.5.** Let $X_1, X_2, ..., X_n$ be a random sample. The sample median $M$ is defined as

$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}\right] & \text{if } n \text{ is even.} \end{cases}$$

The median is a measure of location like sample mean.

Recall that for continuous distribution, $100p^{\text{th}}$ percentile, $\pi_p$, is a number such that

$$p = \int_{-\infty}^{\pi_p} f(x)\,dx.$$

**Definition 13.6.** The $100p^{\text{th}}$ sample percentile is defined as

$$\pi_p = \begin{cases} X_{([np])} & \text{if } p < 0.5 \\ M & \text{if } p = 0.5 \\ X_{(n+1-[n(1-p)])} & \text{if } p > 0.5. \end{cases}$$

where $[b]$ denote the number $b$ rounded to the nearest integer.

**Example 13.22.** Let $X_1, X_2, ..., X_{12}$ be a random sample of size 12. What is the $65^{\text{th}}$ percentile of this sample?

**Answer:**

$$\begin{aligned} 100p &= 65 \\ p &= 0.65 \\ n(1-p) &= (12)(1-0.65) = 4.2 \\ [n(1-p)] &= [4.2] = 4 \end{aligned}$$

Hence by definition of $65^{\text{th}}$ percentile is

$$\begin{aligned} \pi_{0.65} &= X_{(n+1-[n(1-p)])} \\ &= X_{(13-4)} \\ &= X_{(9)}. \end{aligned}$$

Thus, the $65^{\text{th}}$ percentile of the random sample $X_1, X_2, ..., X_{12}$ is the $9^{\text{th}}$-order statistic.

For any number $p$ between 0 and 1, the $100p^{\text{th}}$ sample percentile is an observation such that approximately $np$ observations are less than this observation and $n(1-p)$ observations are greater than this.

**Definition 13.7.** The $25^{\text{th}}$ percentile is called the lower quartile while the $75^{\text{th}}$ percentile is called the upper quartile. The distance between these two quartiles is called the interquartile range.

**Example 13.23.** If a sample of size 3 from a uniform distribution over $[0, 1]$ is observed, what is the probability that the sample median is between $\frac{1}{4}$ and $\frac{3}{4}$?

**Answer:** When a sample of $(2n + 1)$ random variables are observed, the $(n + 1)^{\text{th}}$ smallest random variable is called the sample median. For our problem, the sample median is given by

$$X_{(2)} = 2^{\text{nd}} \text{ smallest } \{X_1, X_2, X_3\}.$$

Let $Y = X_{(2)}$. The density function of each $X_i$ is given by

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the cumulative density function of $f(x)$ is

$$F(x) = x.$$

Thus the density function of $Y$ is given by

$$\begin{aligned} g(y) &= \frac{3!}{1!\,1!}\ [F(y)]^{2-1}\ f(y)\ [1 - F(y)]^{3-2} \\ &= 6\,F(y)\,f(y)\ [1 - F(y)] \\ &= 6y\,(1 - y). \end{aligned}$$

Therefore

$$\begin{aligned} P\left(\frac{1}{4} < Y < \frac{3}{4}\right) &= \int_{\frac{1}{4}}^{\frac{3}{4}} g(y)\,dy \\ &= \int_{\frac{1}{4}}^{\frac{3}{4}} 6\,y\,(1 - y)\,dy \\ &= 6\left[\frac{y^2}{2} - \frac{y^3}{3}\right]_{\frac{1}{4}}^{\frac{3}{4}} \\ &= \frac{11}{16}. \end{aligned}$$

### 13.6. Review Exercises

**1.** Suppose we roll a die 1000 times. What is the probability that the sum of the numbers obtained lies between 3000 and 4000?

**2.** Suppose Kathy flip a coin 1000 times. What is the probability she will get at least 600 heads?

**3.** At a certain large university the weight of the male students and female students are approximately normally distributed with means and standard deviations of 180, and 20, and 130 and 15, respectively. If a male and female are selected at random, what is the probability that the sum of their weights is less than 280?

**4.** Seven observations are drawn from a population with an unknown continuous distribution. What is the probability that the least and the greatest observations bracket the median?

**5.** If the random variable $X$ has the density function

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

what is the probability that the larger of 2 independent observations of $X$ will exceed $\frac{1}{2}$?

**6.** Let $X_1, X_2, X_3$ be a random sample from the uniform distribution on the interval (0, 1). What is the probability that the sample median is less than 0.4?

**7.** Let $X_1, X_2, X_3, X_4, X_5$ be a random sample from the uniform distribution on the interval $(0, \theta)$, where $\theta$ is unknown, and let $X_{max}$ denote the largest observation. For what value of the constant $k$, the expected value of the random variable $kX_{max}$ is equal to $\theta$?

**8.** A random sample of size 16 is to be taken from a normal population having mean 100 and variance 4. What is the $90^{\text{th}}$ percentile of the distribution of the sample mean?

**9.** If the density function of a random variable $X$ is given by

$$f(x) = \begin{cases} \frac{1}{2x} & \text{for } \frac{1}{e} < x < e \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the probability that one of the two independent observations of $X$ is less than 2 and the other is greater than 1?

**10.** Five observations have been drawn independently and at random from a continuous distribution. What is the probability that the next observation will be less than all of the first 5?

**11.** Let the random variable $X$ denote the length of time it takes to complete a mathematics assignment. Suppose the density function of $X$ is given by

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is a positive constant that represents the minimum time to complete a mathematics assignment. If $X_1, X_2, ..., X_5$ is a random sample from this distribution. What is the expected value of $X_{(1)}$?

**12.** Let $X$ and $Y$ be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the probability density function of $W = \max\{X, Y\}$ ?

**13.** Let $X$ and $Y$ be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} \frac{3\,x^2}{\theta^3} & \text{for } 0 \leq x \leq \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

for some $\theta > 0$. What is the probability density function of $W = \min\{X, Y\}$?

**14.** Let $X_1, X_2, ..., X_n$ be a random sample from a uniform distribution on the interval from 0 to 5. What is the limiting moment generating function of $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ as $n \to \infty$?

**15.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a normal distribution with mean $\mu$ and variance 1. If the $75^{\text{th}}$ percentile of the statistic $W = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ is 28.24, what is the sample size $n$ ?

**16.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a Bernoulli distribution with probability of success $p = \frac{1}{2}$. What is the limiting distribution the sample mean $\overline{X}$ ?

**17.** Let $X_1, X_2, ..., X_{1995}$ be a random sample of size 1995 from a distribution with probability density function

$$f(x) = \frac{e^{-\lambda}\,\lambda^x}{x!} \qquad x = 0, 1, 2, 3, ..., \infty.$$

What is the distribution of $1995\overline{X}$ ?

**18.** Suppose $X_1, X_2, ..., X_n$ is a random sample from the uniform distribution on $(0, 1)$ and $Z$ be the sample range. What is the probability that $Z$ is less than or equal to 0.5?

**19.** Let $X_1, X_2, ..., X_9$ be a random sample from a uniform distribution on the interval $[1, 12]$. Find the probability that the next to smallest is greater than or equal to 4?

**20.** A machine needs 4 out of its 6 independent components to operate. Let $X_1, X_2, ..., X_6$ be the lifetime of the respective components. Suppose each is exponentially distributed with parameter $\theta$. What is the probability density function of the machine lifetime?

**21.** Suppose $X_1, X_2, ..., X_{2n+1}$ is a random sample from the uniform distribution on $(0, 1)$. What is the probability density function of the sample median $X_{(n+1)}$?

**22.** Let $X$ and $Y$ be two random variables with joint density

$$f(x,\, y) = \begin{cases} 12x & \text{if } 0 < y < 2x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of the random variable $Z = X^2Y^3 + X^2 - X - Y^3$?

**23.** Let $X_1, X_2, ..., X_{50}$ be a random sample of size 50 from a distribution with density

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\,\theta^\alpha}\, x^{\alpha-1} e^{-\frac{x}{\theta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What are the mean and variance of the sample mean $\overline{X}$?

**24.** Let $X_1, X_2, ..., X_{100}$ be a random sample of size 100 from a distribution with density

$$f(x) = \begin{cases} \frac{e^{-\lambda}\,\lambda^x}{x!} & \text{for } x = 0, 1, 2, ..., \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that $\overline{X}$ greater than or equal to 1?

# Chapter 14

# SAMPLING DISTRIBUTIONS ASSOCIATED WITH THE NORMAL POPULATIONS

Given a random sample $X_1, X_2, ..., X_n$ from a population $X$ with probability distribution $f(x;\theta)$, where $\theta$ is a parameter, a *statistic* is a function $T$ of $X_1, X_2, ..., X_n$, that is

$$T = T(X_1, X_2, ..., X_n)$$

which is free of the parameter $\theta$. If the distribution of the population is known, then sometimes it is possible to find the probability distribution of the statistic $T$. The probability distribution of the statistic $T$ is called the sampling distribution of $T$. The joint distribution of the random variables $X_1, X_2, ..., X_n$ is called the distribution of the sample. The distribution of the sample is the joint density

$$f(x_1, x_2, ..., x_n;\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

since the random variables $X_1, X_2, ..., X_n$ are independent and identically distributed.

Since the normal population is very important in statistics, the sampling distributions associated with the normal population are very important. The most important sampling distributions which are associated with the normal

population are the followings: the chi-square distribution, the student's t-distribution, the F-distribution, and the beta distribution. In this chapter, we only consider the first three distributions, since the last distribution was considered earlier.

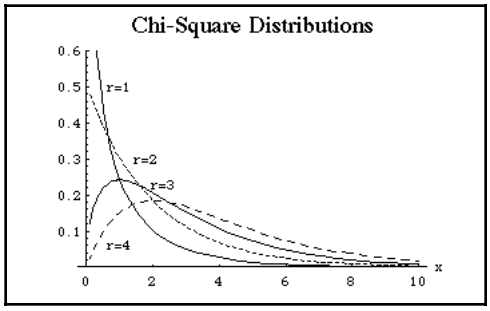
### 14.1. Chi-square distribution

In this section, we treat the Chi-square distribution, which is one of the very useful sampling distributions.

**Definition 14.1.** A continuous random variable $X$ is said to have a chi-square distribution with $r$ degrees of freedom if its probability density function is of the form

$$f(x;\, r) = \begin{cases} \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}}\, x^{\frac{r}{2}-1}\, e^{-\frac{x}{2}} & \text{if } 0 \le x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $r > 0$. If $X$ has chi-square distribution, then we denote it by writing $X \sim \chi^2(r)$. Recall that a gamma distribution reduces to chi-square distribution if $\alpha = \frac{r}{2}$ and $\theta = 2$. The mean and variance of $X$ are $r$ and $2r$, respectively.




Thus, chi-square distribution is also a special case of gamma distribution. Further, if $r \to \infty$, then chi-square distribution tends to normal distribution.

**Example 14.1.** If $X \sim GAM(1, 1)$, then what is the probability density function of the random variable $2X$?

**Answer:** We will use the moment generating method to find the distribution of $2X$. The moment generating function of a gamma random variable is given by

$$M(t) = (1 - \theta\, t)^{-\alpha}, \qquad \text{if} \qquad t < \frac{1}{\theta}.$$

Since $X \sim GAM(1,1)$, the moment generating function of $X$ is given by

$$M_X(t) = \frac{1}{1-t}, \qquad t < 1.$$

Hence, the moment generating function of $2X$ is

$$\begin{aligned} M_{2X}(t) &= M_X(2t) \\ &= \frac{1}{1-2t} \\ &= \frac{1}{(1-2t)^{\frac{2}{2}}} \\ &= \text{MGF of } \chi^2(2). \end{aligned}$$

Hence, if $X$ is $GAM(1,1)$ or is an exponential with parameter 1, then $2X$ is chi-square with 2 degrees of freedom.

**Example 14.2.** If $X \sim \chi^2(5)$, then what is the probability that $X$ is between 1.145 and 12.83?

**Answer:** The probability of $X$ between 1.145 and 12.83 can be calculated from the following:

$$\begin{aligned} &P(1.145 \leq X \leq 12.83) \\ &\quad = P(X \leq 12.83) - P(X \leq 1.145) \\ &\quad = \int_0^{12.83} f(x)\,dx - \int_0^{1.145} f(x)\,dx \\ &\quad = \int_0^{12.83} \frac{1}{\Gamma\left(\frac{5}{2}\right) 2^{\frac{5}{2}}}\, x^{\frac{5}{2}-1}\, e^{-\frac{x}{2}}\,dx - \int_0^{1.145} \frac{1}{\Gamma\left(\frac{5}{2}\right) 2^{\frac{5}{2}}}\, x^{\frac{5}{2}-1}\, e^{-\frac{x}{2}}\,dx \\ &\quad = 0.975 - 0.050 \qquad \text{(from } \chi^2 \text{ table)} \\ &\quad = 0.925. \end{aligned}$$

The above integrals are hard to evaluate and thus their values are taken from the chi-square table.

**Example 14.3.** If $X \sim \chi^2(7)$, then what are values of the constants $a$ and $b$ such that $P(a < X < b) = 0.95$?

**Answer:** Since

$$0.95 = P(a < X < b) = P(X < b) - P(X < a),$$

we get

$$P(X < b) = 0.95 + P(X < a).$$

We choose $a = 1.690$, so that

$$P(X < 1.690) = 0.025.$$

From this, we get

$$P(X < b) = 0.95 + 0.025 = 0.975$$

Thus, from chi-square table, we get $b = 16.01$.

The following theorems were studied earlier in Chapters 6 and 13 and they are very useful in finding the sampling distributions of many statistics. We state these theorems here for the convenience of the reader.

**Theorem 14.1.** If $X \sim N(\mu, \sigma^2)$, then $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$.

**Theorem 14.2.** If $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, ..., X_n$ is a random sample from the population $X$, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).$$

**Theorem 14.3.** If $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, ..., X_n$ is a random sample from the population $X$, then

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2(n-1).$$

**Theorem 14.4.** If $X \sim GAM(\theta, \alpha)$, then

$$\frac{2}{\theta} X \sim \chi^2(2\alpha).$$

**Example 14.4.** A new component is placed in service and $n$ spares are available. If the times to failure in days are independent exponential variables, that is $X_i \sim EXP(100)$, how many spares would be needed to be 95% sure of successful operation for at least two years ?

**Answer:** Since $X_i \sim EXP(100)$,

$$\sum_{i=1}^{n} X_i \sim GAM(100, n).$$

Hence, by Theorem 14.4, the random variable

$$Y = \frac{2}{100} \sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

We have to find the number of spares $n$ such that

$$\begin{aligned}
0.95 &= P\left(\sum_{i=1}^{n} X_i \geq 2 \text{ years}\right) \\
&= P\left(\sum_{i=1}^{n} X_i \geq 730 \text{ days}\right) \\
&= P\left(\frac{2}{100} \sum_{i=1}^{n} X_i \geq \frac{2}{100} 730 \text{ days}\right) \\
&= P\left(\frac{2}{100} \sum_{i=1}^{n} X_i \geq \frac{730}{50}\right) \\
&= P\left(\chi^2(2n) \geq 14.6\right). \\
2n &= 25 \qquad\qquad (\text{from } \chi^2 \text{ table})
\end{aligned}$$

Hence $n = 13$ (after rounding up to the next integer). Thus, 13 spares are needed to be 95% sure of successful operation for at least two years.

**Example 14.5.** If $X \sim N(10, 25)$ and $X_1, X_2, ..., X_{501}$ is a random sample of size 501 from the population $X$, then what is the expected value of the sample variance $S^2$ ?

**Answer:** We will use the Theorem 14.3, to do this problem. By Theorem 14.3, we see that

$$\frac{(501-1)\,S^2}{\sigma^2} \sim \chi^2(500).$$

Hence, the expected value of $S^2$ is given by

$$\begin{aligned}
E\left[S^2\right] &= E\left[\left(\frac{25}{500}\right)\left(\frac{500}{25}\right) S^2\right] \\
&= \left(\frac{25}{500}\right) E\left[\left(\frac{500}{25}\right) S^2\right] \\
&= \left(\frac{1}{20}\right) E\left[\chi^2(500)\right] \\
&= \left(\frac{1}{20}\right) 500 \\
&= 25.
\end{aligned}$$

### 14.2. Student's $t$-distribution

Here we treat the Student's $t$-distribution, which is also one of the very useful sampling distributions.

**Definition 14.2.** A continuous random variable $X$ is said to have a $t$-distribution with $\nu$ degrees of freedom if its probability density function is of the form

$$f(x;\,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\,\nu}\;\Gamma\left(\frac{\nu}{2}\right)\;\left(1+\frac{x^2}{\nu}\right)^{\left(\frac{\nu+1}{2}\right)}}, \qquad -\infty < x < \infty$$

where $\nu > 0$. If $X$ has a $t$-distribution with $\nu$ degrees of freedom, then we denote it by writing $X \sim t(\nu)$.

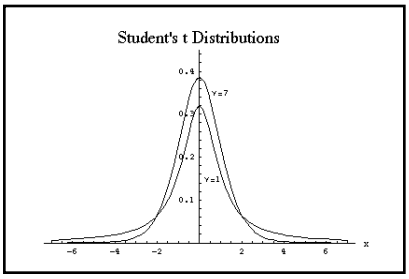
The $t$-distribution was discovered by W.S. Gosset (1876-1936) of England who published his work under the pseudonym of student. Therefore, this distribution is known as Student's $t$-distribution. This distribution is a generalization of the Cauchy distribution and the normal distribution. That is, if $\nu = 1$, then the probability density function of $X$ becomes

$$f(x;\,1) = \frac{1}{\pi\,(1+x^2)} \qquad -\infty < x < \infty,$$

which is the Cauchy distribution. Further, if $\nu \to \infty$, then

$$\lim_{\nu\to\infty} f(x;\,\nu) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}x^2} \qquad -\infty < x < \infty,$$

which is the probability density function of the standard normal distribution. The following figure shows the graph of $t$-distributions with various degrees of freedom.




**Example 14.6.** If $T \sim t(10)$, then what is the probability that $T$ is at least 2.228 ?

**Answer:** The probability that $T$ is at least 2.228 is given by

$$\begin{aligned} P(T \geq 2.228) &= 1 - P(T < 2.228) \\ &= 1 - 0.975 \qquad (\text{from t} - \text{table}) \\ &= 0.025. \end{aligned}$$

**Example 14.7.** If $T \sim t(19)$, then what is the value of the constant $c$ such that $P(|T| \leq c) = 0.95$ ?

**Answer:**

$$\begin{aligned} 0.95 &= P(|T| \leq c) \\ &= P(-c \leq T \leq c) \\ &= P(T \leq c) - 1 + P(T \leq c) \\ &= 2\,P(T \leq c) - 1. \end{aligned}$$

Hence

$$P(T \leq c) = 0.975.$$

Thus, using the t-table, we get for 19 degrees of freedom

$$c = 2.093.$$

**Theorem 14.5.** If the random variable $X$ has a $t$-distribution with $\nu$ degrees of freedom, then

$$E[X] = \begin{cases} 0 & \text{if} \quad \nu \geq 2 \\ \\ DNE & \text{if} \quad \nu = 1 \end{cases}$$

and

$$Var[X] = \begin{cases} \frac{\nu}{\nu-2} & \text{if} \quad \nu \geq 3 \\ \\ DNE & \text{if} \quad \nu = 1, 2 \end{cases}$$

where DNE means does not exist.

**Theorem 14.6.** If $Z \sim N(0,1)$ and $U \sim \chi^2(\nu)$ and in addition, $Z$ and $U$ are independent, then the random variable $W$ defined by

$$W = \frac{Z}{\sqrt{\frac{U}{\nu}}}$$

has a $t$-distribution with $\nu$ degrees of freedom.

**Theorem 14.7.** If $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, ..., X_n$ be a random sample from the population $X$, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

**Proof:** Since each $X_i \sim N(\mu, \sigma^2)$,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Thus,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Further, from Theorem 14.3 we know that

$$(n-1)\,\frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

Hence

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)\,S^2}{(n-1)\,\sigma^2}}} \sim t(n-1) \qquad \text{(by Theorem 14.6).}$$

This completes the proof of the theorem.

**Example 14.8.** Let $X_1, X_2, X_3, X_4$ be a random sample of size 4 from a standard normal distribution. If the statistic $W$ is given by

$$W = \frac{X_1 - X_2 + X_3}{\sqrt{X_1^2 + X_2^2 + X_3^2 + X_4^2}},$$

then what is the expected value of $W$ ?

**Answer:** Since $X_i \sim N(0,1)$, we get

$$X_1 - X_2 + X_3 \sim N(0,3)$$

and

$$\frac{X_1 - X_2 + X_3}{\sqrt{3}} \sim N(0,1).$$

Further, since $X_i \sim N(0,1)$, we have

$$X_i^2 \sim \chi^2(1)$$

and hence

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 \sim \chi^2(4)$$

Thus,

$$\frac{\frac{X_1 - X_2 + X_3}{\sqrt{3}}}{\sqrt{\frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{4}}} = \left(\frac{2}{\sqrt{3}}\right) W \sim t(4).$$

Now using the distribution of $W$, we find the expected value of $W$.

$$\begin{aligned} E[W] &= \left(\frac{\sqrt{3}}{2}\right) E\left[\frac{2}{\sqrt{3}} W\right] \\ &= \left(\frac{\sqrt{3}}{2}\right) E[t(4)] \\ &= \left(\frac{\sqrt{3}}{2}\right) 0 \\ &= 0. \end{aligned}$$

**Example 14.9.** If $X \sim N(0,1)$ and $X_1, X_2$ is random sample of size two from the population $X$, then what is the $75^{\text{th}}$ percentile of the statistic $W = \frac{X_1}{\sqrt{X_2^2}}$?

**Answer:** Since each $X_i \sim N(0,1)$, we have

$$\begin{aligned} X_1 &\sim N(0,1) \\ X_2^2 &\sim \chi^2(1). \end{aligned}$$

Hence

$$W = \frac{X_1}{\sqrt{X_2^2}} \sim t(1).$$

The $75^{\text{th}}$ percentile $a$ of $W$ is then given by

$$0.75 = P(W \leq a)$$

Hence, from the $t$-table, we get

$$a = 1.0$$

Hence the $75^{\text{th}}$ percentile of $W$ is 1.0.

**Example 14.10.** Suppose $X_1, X_2, ...., X_n$ is a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$. If $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $V^2 =$

$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, and $X_{n+1}$ is an additional observation, what is the value of $m$ so that the statistics $\frac{m(\overline{X}-X_{n+1})}{V}$ has a $t$-distribution.

**Answer:** Since

$$\begin{aligned}
&X_i \sim N(\mu,\sigma^2)\\
\Rightarrow\ &\overline{X} \sim N\left(\mu,\frac{\sigma^2}{n}\right)\\
\Rightarrow\ &\overline{X}-X_{n+1} \sim N\left(\mu-\mu,\ \frac{\sigma^2}{n}+\sigma^2\right)\\
\Rightarrow\ &\overline{X}-X_{n+1} \sim N\left(0,\ \left(\frac{n+1}{n}\right)\sigma^2\right)\\
\Rightarrow\ &\frac{\overline{X}-X_{n+1}}{\sigma\sqrt{\frac{n+1}{n}}} \sim N(0,1)
\end{aligned}$$

Now, we establish a relationship between $V^2$ and $S^2$. We know that

$$\begin{aligned}
(n-1)\,S^2 &= (n-1)\,\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\overline{X})^2\\
&= \sum_{i=1}^{n}(X_i-\overline{X})^2\\
&= n\ \left(\frac{1}{n}\ \sum_{i=1}^{n}(X_i-\overline{X})^2\right)\\
&= n\,V^2.
\end{aligned}$$

Hence, by Theorem 14.3

$$\frac{n\,V^2}{\sigma^2}=\frac{(n-1)\,S^2}{\sigma^2}\sim\chi^2(n-1).$$

Thus

$$\left(\sqrt{\frac{n-1}{n+1}}\right)\ \frac{\overline{X}-X_{n+1}}{V}=\frac{\frac{\overline{X}-X_{n+1}}{\sigma\sqrt{\frac{n+1}{n}}}}{\sqrt{\frac{\frac{n\,V^2}{\sigma^2}}{(n-1)}}}\sim t(n-1).$$

Thus by comparison, we get

$$m=\sqrt{\frac{n-1}{n+1}}.$$

### 14.3. Snedecor's $F$-distribution

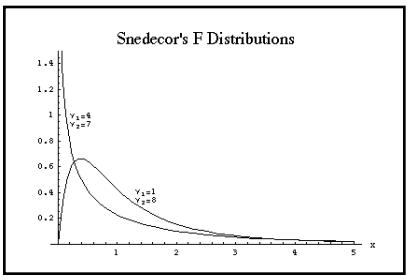
The next sampling distribution to be discussed in this chapter is Snedecor's $F$-distribution. This distribution has many applications in mathematical statistics. In the analysis of variance, this distribution is used to develop the technique for testing the equalities of sample means.

**Definition 14.3.** A continuous random variable $X$ is said to have a $F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom if its probability density function is of the form

$$f(x;\,\nu_1,\nu_2) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}\, x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}\,x\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}} & \text{if } 0 \le x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\nu_1,\ \nu_2 > 0$. If $X$ has a $F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom, then we denote it by writing $X \sim F\,(\nu_1,\nu_2)$.

The $F$-distribution was named in honor of Sir Ronald Fisher by George Snedecor. $F$-distribution arises as the distribution of a ratio of variances. Like, the other two distributions this distribution also tends to normal distribution as $\nu_1$ and $\nu_2$ become very large. The following figure illustrates the shape of the graph of this distribution for various degrees of freedom.




The following theorem gives us the mean and variance of Snedecor's $F$-distribution.

**Theorem 14.8.** If the random variable $X \sim F\,(\nu_1,\nu_2)$, then

$$E[X] = \begin{cases} \frac{\nu_2}{\nu_2-2} & \text{if} \quad \nu_2 \ge 3 \\ DNE & \text{if} \quad \nu_2 = 1,2 \end{cases}$$

and

$$Var[X] = \begin{cases} \frac{2\,\nu_2^2\,(\nu_1+\nu_2-2)}{\nu_1\,(\nu_2-2)^2\,(\nu_2-4)} & \text{if} \quad \nu_2 \ge 5 \\ DNE & \text{if} \quad \nu_2 = 1,2,3,4. \end{cases}$$

Here DNE means does not exist.

**Example 14.11.** If $X \sim F(9, 10)$, what $P(X \geq 3.02)$ ? Also, find the mean and variance of $X$.

**Answer:**

$$P(X \geq 3.02) = 1 - P(X \leq 3.02)$$

$$= 1 - P(F(9, 10) \leq 3.02)$$

$$= 1 - 0.95 \qquad (\text{from F} - \text{table})$$

$$= 0.05.$$

Next, we determine the mean and variance of $X$ using the Theorem 14.8. Hence,

$$E(X) = \frac{\nu_2}{\nu_2 - 2} = \frac{10}{10 - 2} = \frac{10}{8} = 1.25$$

and

$$Var(X) = \frac{2\,\nu_2^2\,(\nu_1 + \nu_2 - 2)}{\nu_1\,(\nu_2 - 2)^2\,(\nu_2 - 4)}$$

$$= \frac{2\,(10)^2\,(19 - 2)}{9\,(8)^2\,(6)}$$

$$= \frac{(25)\,(17)}{(27)\,(16)}$$

$$= \frac{425}{432} = 0.9838.$$

**Theorem 14.9.** If $X \sim F(\nu_1, \nu_2)$, then the random variable $\frac{1}{X} \sim F(\nu_2, \nu_1)$.

This theorem is very useful for computing probabilities like $P(X \leq 0.2439)$. If you look at a $F$-table, you will notice that the table start with values bigger than 1. Our next example illustrates how to find such probabilities using Theorem 14.9.

**Example 14.12.** If $X \sim F(6, 9)$, what is the probability that $X$ is less than or equal to 0.2439 ?

**Answer:** We use the above theorem to compute

$$\begin{aligned} P\left(X \leq 0.2439\right) &= P\left(\frac{1}{X} \geq \frac{1}{0.2439}\right) \\ &= P\left(F(9,6) \geq \frac{1}{0.2439}\right) \qquad \text{(by Theorem 14.9)} \\ &= 1 - P\left(F(9,6) \leq \frac{1}{0.2439}\right) \\ &= 1 - P\left(F(9,6) \leq 4.10\right) \\ &= 1 - 0.95 \\ &= 0.05. \end{aligned}$$

The following theorem says that $F$-distribution arises as the distribution of a random variable which is the quotient of two independently distributed chi-square random variables, each of which is divided by its degrees of freedom.

**Theorem 14.10.** If $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$, and the random variables $U$ and $V$ are independent, then the random variable

$$\frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F\left(\nu_1, \nu_2\right).$$

**Example 14.13.** Let $X_1, X_2, ..., X_4$ and $Y_1, Y_2, ..., Y_5$ be two random samples of size 4 and 5 respectively, from a standard normal population. What is the variance of the statistic $T = \left(\frac{5}{4}\right) \frac{X_1^2+X_2^2+X_3^2+X_4^2}{Y_1^2+Y_2^2+Y_3^2+Y_4^2+Y_5^2}$ ?

**Answer:** Since the population is standard normal, we get

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 \sim \chi^2(4).$$

Similarly,

$$Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2 \sim \chi^2(5).$$

Thus

$$\begin{aligned} T &= \left(\frac{5}{4}\right) \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2} \\ &= \frac{\frac{X_1^2+X_2^2+X_3^2+X_4^2}{4}}{\frac{Y_1^2+Y_2^2+Y_3^2+Y_4^2+Y_5^2}{5}} \\ &= T \sim F(4,5). \end{aligned}$$

Therefore

$$\begin{aligned} Var(T) &= Var[\,F(4,5)\,] \\ &= \frac{2\,(5)^2\,(7)}{4\,(3)^2\,(1)} \\ &= \frac{350}{36} \\ &= 9.72. \end{aligned}$$

**Theorem 14.11.** Let $X \sim N(\mu_1,\, \sigma_1^2)$ and $X_1, X_2, ..., X_n$ be a random sample of size $n$ from the population $X$. Let $Y \sim N(\mu_2,\, \sigma_2^2)$ and $Y_1, Y_2, ..., Y_m$ be a random sample of size $m$ from the population $Y$. Then the statistic

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n-1,\, m-1),$$

where $S_1^2$ and $S_2^2$ denote the sample variances of the first and the second sample, respectively.

**Proof:** Since,

$$X_i \sim N(\mu_1, \sigma_1^2)$$

we have by Theorem 14.3, we get

$$(n-1)\,\frac{S_1^2}{\sigma_1^2} \sim \chi^2(n-1).$$

Similarly, since

$$Y_i \sim N(\mu_2, \sigma_2^2)$$

we have by Theorem 14.3, we get

$$(m-1)\,\frac{S_2^2}{\sigma_2^2} \sim \chi^2(m-1).$$

Therefore

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\frac{(n-1)\,S_1^2}{(n-1)\,\sigma_1^2}}{\frac{(m-1)\,S_2^2}{(m-1)\,\sigma_2^2}} \sim F(n-1,\, m-1).$$

This completes the proof of the theorem.

Because of this theorem, the $F$-distribution is also known as the variance-ratio distribution.

### 14.4. Review Exercises

**1.** Let $X_1, X_2, ..., X_5$ be a random sample of size 5 from a normal distribution with mean zero and standard deviation 2. Find the sampling distribution of the statistic $X_1 + 2X_2 - X_3 + X_4 + X_5$.

**2.** Let $X_1, X_2, X_3$ be a random sample of size 3 from a standard normal distribution. Find the distribution of $X_1^2 + X_2^2 + X_3^2$. If possible, find the sampling distribution of $X_1^2 - X_2^2$. If not, justify why you can not determine it's distribution.

**3.** Let $X_1, X_2, X_3$ be a random sample of size 3 from a standard normal distribution. Find the sampling distribution of the statistics $\frac{X_1+X_2+X_3}{\sqrt{X_1^2+X_2^2+X_3^2}}$ and $\frac{X_1-X_2-X_3}{\sqrt{X_1^2+X_2^2+X_3^2}}$.

**4.** Let $X_1, X_2, X_3$ be a random sample of size 3 from an exponential distribution with a parameter $\theta > 0$. Find the distribution of the sample (that is the joint distribution of the random variables $X_1, X_2, X_3$).

**5.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a normal population with mean $\mu$ and variance $\sigma^2 > 0$. What is the expected value of the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$?

**6.** Let $X_1, X_2, X_3, X_4$ be a random sample of size 4 from a standard normal population. Find the distribution of the statistic $\frac{X_1+X_4}{\sqrt{X_2^2+X_3^2}}$.

**7.** Let $X_1, X_2, X_3, X_4$ be a random sample of size 4 from a standard normal population. Find the sampling distribution (if possible) and moment generating function of the statistic $2X_1^2+3X_2^2+X_3^2+4X_4^2$. What is the probability distribution of the sample?

**8.** Let $X$ equal the maximal oxygen intake of a human on a treadmill, where the measurement are in milliliters of oxygen per minute per kilogram of weight. Assume that for a particular population the mean of $X$ is $\mu = 54.03$ and the standard deviation is $\sigma = 5.8$. Let $\bar{X}$ be the sample mean of a random sample $X_1, X_2, ..., X_{47}$ of size 47 drawn from $X$. Find the probability that the sample mean is between 52.761 and 54.453.

**9.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$. What is the variance of $V^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ ?

**10.** If $X$ is a random variable with mean $\mu$ and variance $\sigma^2$, then $\mu - 2\sigma$ is called the lower $2\sigma$ point of $X$. Suppose a random sample $X_1, X_2, X_3, X_4$ is

drawn from a chi-square distribution with two degrees of freedom. What is the lower $2\sigma$ point of $X_1 + X_2 + X_3 + X_4$ ?

**11.** Let $X$ and $Y$ be independent normal random variables such that the mean and variance of $X$ are 2 and 4, respectively, while the mean and variance of $Y$ are 6 and $k$, respectively. A sample of size 4 is taken from the $X$-distribution and a sample of size 9 is taken from the $Y$-distribution. If $P\left(\overline{Y} - \overline{X} > 8\right) = 0.0228$, then what is the value of the constant $k$ ?

**12.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with density function

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the distribution of the statistic $Y = 2\lambda \sum_{i=1}^{n} X_i$ ?

**13.** Suppose $X$ has a normal distribution with mean 0 and variance 1, $Y$ has a chi-square distribution with $n$ degrees of freedom, $W$ has a chi-square distribution with $p$ degrees of freedom, and $W, X$, and $Y$ are independent. What is the sampling distribution of the statistic $V = \frac{X}{\sqrt{\frac{W+Y}{p+n}}}$?

**14.** A random sample $X_1, X_2, ..., X_n$ of size $n$ is selected from a normal population with mean $\mu$ and standard deviation 1. Later an additional independent observation $X_{n+1}$ is obtained from the same population. What is the distribution of the statistic $(X_{n+1} - \mu)^2 + \sum_{i=1}^{n}(X_i - \overline{X})^2$, where $\overline{X}$ denote the sample mean?

**15.** Let $T = \frac{k(X+Y)}{\sqrt{Z^2+W^2}}$, where $X$, $Y$, $Z$, and $W$ are independent normal random variables with mean 0 and variance $\sigma^2 > 0$. For exactly one value of $k$, $T$ has a t-distribution. If $r$ denotes the degrees of freedom of that distribution, then what is the value of the pair $(k, r)$?

**16.** Let $X$ and $Y$ be joint normal random variables with common mean 0, common variance 1, and covariance $\frac{1}{2}$. What is the probability of the event $\left(X + Y \leq \sqrt{3}\right)$, that is $P\left(X + Y \leq \sqrt{3}\right)$?

**17.** Suppose $X_j = Z_j - Z_{j-1}$, where $j = 1, 2, ..., n$ and $Z_0, Z_1, ..., Z_n$ are independent and identically distributed with common variance $\sigma^2$. What is the variance of the random variable $\frac{1}{n}\sum_{j=1}^{n} X_j$ ?

**18.** A random sample of size 5 is taken from a normal distribution with mean 0 and standard deviation 2. Find the constant $k$ such that 0.05 is equal to the

probability that the sum of the squares of the sample observations exceeds the constant $k$.

**19.** Let $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ be two random sample from the independent normal distributions with $Var[X_i] = \sigma^2$ and $Var[Y_i] = 2\sigma^2$, for $i = 1, 2, ..., n$ and $\sigma^2 > 0$. If $U = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ and $V = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$, then what is the sampling distribution of the statistic $\frac{2U+V}{2\sigma^2}$?

**20.** Suppose $X_1, X_2, ..., X_6$ and $Y_1, Y_2, ..., Y_9$ are independent, identically distributed normal random variables, each with mean zero and variance $\sigma^2 > 0$. What is the $95^{\text{th}}$ percentile of the statistics $W = \left[\sum_{i=1}^{6} X_i^2\right] / \left[\sum_{j=1}^{9} Y_j^2\right]$?

**21.** Let $X_1, X_2, ..., X_6$ and $Y_1, Y_2, ..., Y_8$ be independent random samples from a normal distribution with mean 0 and variance 1, and $Z = \left[4\sum_{i=1}^{6} X_i^2\right] / \left[3\sum_{j=1}^{8} Y_j^2\right]$?

# Chapter 15
# SOME TECHNIQUES FOR FINDING POINT ESTIMATORS OF PARAMETERS

A statistical population consists of all the measurements of interest in a statistical investigation. Usually a population is described by a random variable $X$. If we can gain some knowledge about the probability density function $f(x;\theta)$ of $X$, then we also gain some knowledge about the population under investigation.

A sample is a portion of the population usually chosen by method of random sampling and as such it is a set of random variables $X_1, X_2, ..., X_n$ with the same probability density function $f(x;\theta)$ as the population. Once the sampling is done, we get

$$X_1 = x_1,\ X_2 = x_2,\ \cdots,\ X_n = x_n$$

where $x_1, x_2, ..., x_n$ are the *sample data*.

Every statistical method employs a random sample to gain information about the population. Since the population is characterized by the probability density function $f(x;\theta)$, in statistics one makes statistical inferences about the population distribution $f(x;\theta)$ based on sample information. A statistical inference is a statement based on sample information about the population. There are three types of statistical inferences (1) estimation (2)

hypothesis testing and (3) prediction. The goal of this chapter is to examine some well known point estimation methods.

In point estimation, we try to find the parameter $\theta$ of the population distribution $f(x;\theta)$ from the sample information. Thus, in the parametric point estimation one assumes the functional form of the pdf $f(x;\theta)$ to be known and only estimate the unknown parameter $\theta$ of the population using information available from the sample.

**Definition 15.1.** Let $X$ be a population with the density function $f(x;\theta)$, where $\theta$ is an unknown parameter. The set of all admissible values of $\theta$ is called a parameter space and it is denoted by $\Omega$, that is

$$\Omega = \{\theta \in \mathbb{R}^n \,|\, f(x;\theta) \text{ is a pdf } \}$$

for some natural number $m$.

**Example 15.1.** If $X \sim EXP(\theta)$, then what is the parameter space of $\theta$ ?

**Answer:** Since $X \sim EXP(\theta)$, the density function of $X$ is given by

$$f(x;\theta) = \frac{1}{\theta}\, e^{-\frac{x}{\theta}}.$$

If $\theta$ is zero or negative then $f(x;\theta)$ is not a density function. Thus, the admissible values of $\theta$ are all the positive real numbers. Hence

$$\begin{aligned}\Omega &= \{\theta \in \mathbb{R} \mid 0 < \theta < \infty\} \\ &= \mathbb{R}_+.\end{aligned}$$

**Example 15.2.** If $X \sim N\left(\mu, \sigma^2\right)$, what is the parameter space?

**Answer:** The parameter space $\Omega$ is given by

$$\begin{aligned}\Omega &= \left\{\theta \in \mathbb{R}^2 \mid f(x;\theta) \sim N\left(\mu, \sigma^2\right)\right\} \\ &= \left\{(\mu,\sigma) \in \mathbb{R}^2 \mid -\infty < \mu < \infty,\ 0 < \sigma < \infty\right\} \\ &= \mathbb{R} \times \mathbb{R}_+ \\ &= \text{upper half plane.}\end{aligned}$$

In general, a parameter space is a subset of $\mathbb{R}^m$. Statistics concerns with the estimation of the unknown parameter $\theta$ from a random sample $X_1, X_2, ..., X_n$. Recall that a statistic is a function of $X_1, X_2, ..., X_n$ and free of the population parameter $\theta$.

**Definition 15.2.** Let $X \sim f(x;\theta)$ and $X_1, X_2, ..., X_n$ be a random sample from the population $X$. Any statistic that can be used to guess the parameter $\theta$ is called an estimator of $\theta$. The numerical value of this statistic is called an estimate of $\theta$. The estimator of the parameter $\theta$ is denoted by $\widehat{\theta}$.

One of the basic problems is how to find an estimator of population parameter $\theta$. There are several methods for finding an estimator of $\theta$. Some of these methods are:

(1) Moment Method
(2) Maximum Likelihood Method
(3) Bayes Method
(4) Least Squares Method
(5) Minimum Chi-Squares Method
(6) Minimum Distance Method

In this chapter, we only discuss the first three methods of estimating a population parameter.

### 15.1. Moment Method

Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with probability density function $f(x;\theta_1, \theta_2, ..., \theta_m)$, where $\theta_1, \theta_2, ..., \theta_m$ are $m$ unknown parameters. Let

$$E\left(X^k\right) = \int_{-\infty}^{\infty} x^k \, f(x;\theta_1, \theta_2, ..., \theta_m)\, dx$$

be the $k^{\text{th}}$ population moment about 0. Further, let

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

be the $k^{\text{th}}$ sample moment about 0.

In moment method, we find the estimator for the parameters $\theta_1, \theta_2, ..., \theta_m$ by equating the first $m$ population moments (if they exist) to the first $m$ sample moments, that is

$$\begin{aligned} E\left(X\right) &= M_1 \\ E\left(X^2\right) &= M_2 \\ E\left(X^3\right) &= M_3 \\ &\vdots \\ E\left(X^m\right) &= M_m \end{aligned}$$

The moment method is one of the classical methods for estimating parameters and motivation comes from the fact that the sample moments are in some sense estimates for the population moments. The moment method was first discovered by British statistician Karl Pearson in 1902. Now we provide some examples to illustrate this method.

**Example 15.3.** Let $X \sim N\left(\mu, \sigma^2\right)$ and $X_1, X_2, ..., X_n$ be a random sample of size $n$ from the population $X$. What are the estimators of the population parameters $\mu$ and $\sigma^2$ if we use the moment method?

**Answer:** Since the population is normal, that is

$$X \sim N\left(\mu, \sigma^2\right)$$

we know that

$$\begin{aligned} E\left(X\right) &= \mu \\ E\left(X^2\right) &= \sigma^2 + \mu^2. \end{aligned}$$

Hence

$$\begin{aligned} \mu &= E\left(X\right) \\ &= M_1 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \overline{X}. \end{aligned}$$

Therefore, the estimator of the parameter $\mu$ is $\overline{X}$, that is

$$\widehat{\mu} = \overline{X}.$$

Next, we find the estimator of $\sigma^2$ equating $E(X^2)$ to $M_2$. Note that

$$\begin{aligned} \sigma^2 &= \sigma^2 + \mu^2 - \mu^2 \\ &= E\left(X^2\right) - \mu^2 \\ &= M_2 - \mu^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2. \end{aligned}$$

The last line follows from the fact that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2-2\,X_i\,\overline{X}+\overline{X}^2\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\sum_{i=1}^{n}2\,X_i\,\overline{X}+\frac{1}{n}\sum_{i=1}^{n}\overline{X}^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-2\,\overline{X}\,\frac{1}{n}\sum_{i=1}^{n}X_i+\overline{X}^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-2\,\overline{X}\,\overline{X}+\overline{X}^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2.\end{aligned}$$

Thus, the estimator of $\sigma^2$ is $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, that is

$$\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2.$$

**Example 15.4.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population $X$ with probability density function

$$f(x;\theta)=\begin{cases}\theta\,x^{\theta-1} & \text{if } 0<x<1\\ \\ 0 & \text{otherwise,}\end{cases}$$

where $0<\theta<\infty$ is an unknown parameter. Using the method of moment find an estimator of $\theta$ ? If $x_1=0.2, x_2=0.6, x_3=0.5, x_4=0.3$ is a random sample of size 4, then what is the estimate of $\theta$ ?

**Answer:** To find an estimator, we shall equate the population moment to the sample moment. The population moment $E(X)$ is given by

$$\begin{aligned}E(X) &= \int_0^1 x\,f(x;\theta)\,dx\\ &= \int_0^1 x\,\theta\,x^{\theta-1}\,dx\\ &= \theta\int_0^1 x^{\theta}\,dx\\ &= \frac{\theta}{\theta+1}\left[\,x^{\theta+1}\,\right]_0^1\\ &= \frac{\theta}{\theta+1}.\end{aligned}$$

We know that $M_1 = \overline{X}$. Now setting $M_1$ equal to $E(X)$ and solving for $\theta$, we get

$$\overline{X} = \frac{\theta}{\theta + 1}$$

that is

$$\theta = \frac{\overline{X}}{1 - \overline{X}},$$

where $\overline{X}$ is the sample mean. Thus, the statistic $\frac{\overline{X}}{1-\overline{X}}$ is an estimator of the parameter $\theta$. Hence

$$\widehat{\theta} = \frac{\overline{X}}{1 - \overline{X}}.$$

Since $x_1 = 0.2, x_2 = 0.6, x_3 = 0.5, x_4 = 0.3$, we have $\overline{X} = 0.4$ and

$$\widehat{\theta} = \frac{0.4}{1 - 0.4} = \frac{2}{3}$$

is an estimate of the $\theta$.

**Example 15.5.** What is the basic principle of the moment method?

**Answer:** To choose a value for the unknown population parameter for which the observed data have the same moments as the population.

**Example 15.6.** Suppose $X_1, X_2, ..., X_7$ is a random sample from a population $X$ with density function

$$f(x;\beta) = \begin{cases} \frac{x^6 e^{-\frac{x}{\beta}}}{\Gamma(7)\,\beta^7} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Find an estimator of $\beta$ by the moment method.

**Answer:** Since, we have only one parameter, we need to compute only the first population moment $E(X)$ about 0. Thus,

$$\begin{aligned} E(X) &= \int_0^\infty x\, f(x;\beta)\, dx \\ &= \int_0^\infty x\, \frac{x^6\, e^{-\frac{x}{\beta}}}{\Gamma(7)\,\beta^7}\, dx \\ &= \frac{1}{\Gamma(7)} \int_0^\infty \left(\frac{x}{\beta}\right)^7 e^{-\frac{x}{\beta}}\, dx \\ &= \beta\, \frac{1}{\Gamma(7)} \int_0^\infty y^7\, e^{-y}\, dy \\ &= \beta\, \frac{1}{\Gamma(7)}\, \Gamma(8) \\ &= 7\,\beta. \end{aligned}$$

Since $M_1 = \overline{X}$, equating $E(X)$ to $M_1$, we get

$$7\,\beta = \overline{X}$$

that is

$$\beta = \frac{1}{7}\,\overline{X}.$$

Therefore, the estimator of $\beta$ by the moment method is given by

$$\widehat{\beta} = \frac{1}{7}\,\overline{X}.$$

**Example 15.7.** Suppose $X_1, X_2, ..., X_n$ is a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find an estimator of $\theta$ by the moment method.

**Answer:** Examining the density function of the population $X$, we see that $X \sim UNIF(0, \theta)$. Therefore

$$E(X) = \frac{\theta}{2}.$$

Now, equating this population moment to the sample moment, we obtain

$$\frac{\theta}{2} = E(X) = M_1 = \overline{X}.$$

Therefore, the estimator of $\theta$ is

$$\widehat{\theta} = 2\,\overline{X}.$$

**Example 15.8.** Suppose $X_1, X_2, ..., X_n$ is a random sample from a population $X$ with density function

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the estimators of $\alpha$ and $\beta$ by the moment method.

**Answer:** Examining the density function of the population $X$, we see that $X \sim UNIF(\alpha, \beta)$. Since, the distribution has two unknown parameters, we need the first two population moments. Therefore

$$E(X) = \frac{\alpha + \beta}{2} \qquad \text{and} \qquad E(X^2) = \frac{(\beta - \alpha)^2}{12} + E(X)^2.$$

Equating these moments to the corresponding sample moments, we obtain

$$\frac{\alpha + \beta}{2} = E(X) = M_1 = \overline{X}$$

that is

$$\alpha + \beta = 2\overline{X} \tag{1}$$

and

$$\frac{(\beta - \alpha)^2}{12} + E(X)^2 = E(X^2) = M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

which is

$$\begin{aligned}(\beta - \alpha)^2 &= 12\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - E(X)^2\right] \\ &= 12\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right] \\ &= 12\left[\frac{1}{n}\sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2\right].\end{aligned}$$

Hence, we get

$$\beta - \alpha = \sqrt{\frac{12}{n}\sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2}. \tag{2}$$

Adding equation (1) to equation (2), we obtain

$$2\beta = 2\overline{X} \pm 2\sqrt{\frac{3}{n}\sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2}$$

that is

$$\beta = \overline{X} \pm \sqrt{\frac{3}{n}\sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2}.$$

Similarly, subtracting (2) from (1), we get

$$\alpha = \overline{X} \mp \sqrt{\frac{3}{n}\sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2}.$$

Since, $\alpha < \beta$, we see that the estimators of $\alpha$ and $\beta$ are

$$\widehat{\alpha} = \overline{X} - \sqrt{\frac{3}{n}\sum_{i=1}^{n}\left(X_i^2 - \overline{X}\right)^2} \qquad \text{and} \qquad \widehat{\beta} = \overline{X} + \sqrt{\frac{3}{n}\sum_{i=1}^{n}\left(X_i^2 - \overline{X}\right)^2}.$$

## 15.2. Maximum Likelihood Method

The maximum likelihood method was first used by Sir Ronald Fisher in 1912 for finding estimator of a unknown parameter. However, the method originated in the works of Gauss and Bernoulli. Next, we describe the method in detail.

**Definition 15.3.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with probability density function $f(x;\theta)$, where $\theta$ is an unknown parameter. The likelihood function, $L(\theta)$, is the distribution of the sample. That is

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

This definition says that the likelihood function of a random sample $X_1, X_2, ..., X_n$ is the joint density of the random variables $X_1, X_2, ..., X_n$.

The $\theta$ that maximizes the likelihood function $L(\theta)$ is called the maximum likelihood estimator of $\theta$, and it is denoted by $\widehat{\theta}$. Hence

$$\widehat{\theta} = Arg \sup_{\theta\in\Omega} L(\theta),$$

where $\Omega$ is the parameter space of $\theta$ so that $L(\theta)$ is the joint density of the sample.

The method of maximum likelihood in a sense picks out of all the possible values of $\theta$ the one most likely to have produced the given observations $x_1, x_2, ..., x_n$. The method is summarized below: (1) Obtain a random sample $x_1, x_2, ..., x_n$ from the distribution of a population $X$ with probability density function $f(x;\theta)$; (2) define the likelihood function for the sample $x_1, x_2, ..., x_n$ by $L(\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta)$; (3) find the expression for $\theta$ that maximizes $L(\theta)$. This can be done directly or by maximizing $\ln L(\theta)$; (4) replace $\theta$ by $\widehat{\theta}$ to obtain an expression for the maximum likelihood estimator for $\theta$; (5) find the observed value of this estimator for a given sample.

**Example 15.9.** If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} (1-\theta)\, x^{-\theta} & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

what is the maximum likelihood estimator of $\theta$ ?

**Answer:** The likelihood function of the sample is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Therefore

$$\begin{aligned} \ln L(\theta) &= \ln\left(\prod_{i=1}^{n} f(x_i;\theta)\right) \\ &= \sum_{i=1}^{n} \ln f(x_i;\theta) \\ &= \sum_{i=1}^{n} \ln\left[(1-\theta)\, {x_i}^{-\theta}\right] \\ &= n\, \ln(1-\theta) - \theta \sum_{i=1}^{n} \ln x_i. \end{aligned}$$

Now we maximize $\ln L(\theta)$ with respect to $\theta$.

$$\begin{aligned} \frac{d \ln L(\theta)}{d\theta} &= \frac{d}{d\theta}\left(n\, \ln(1-\theta) - \theta \sum_{i=1}^{n} \ln x_i\right) \\ &= -\frac{n}{1-\theta} - \sum_{i=1}^{n} \ln x_i. \end{aligned}$$

Setting this derivative $\frac{d \ln L(\theta)}{d\theta}$ to 0, we get

$$\frac{d \ln L(\theta)}{d\theta} = -\frac{n}{1-\theta} - \sum_{i=1}^{n} \ln x_i = 0$$

that is

$$\frac{1}{1-\theta} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

or

$$\frac{1}{1-\theta} = -\frac{1}{n}\sum_{i=1}^{n}\ln x_i = -\overline{\ln x}.$$

or

$$\theta = 1 + \frac{1}{\overline{\ln x}}.$$

This $\theta$ can be shown to be maximum by the second derivative test and we leave this verification to the reader. Therefore, the estimator of $\theta$ is

$$\widehat{\theta} = 1 + \frac{1}{\overline{\ln X}}.$$

**Example 15.10.** If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\beta) = \begin{cases} \frac{x^6\, e^{-\frac{x}{\beta}}}{\Gamma(7)\,\beta^7} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

then what is the maximum likelihood estimator of $\beta$ ?

**Answer:** The likelihood function of the sample is given by

$$L(\beta) = \prod_{i=1}^{n} f(x_i;\beta).$$

Thus,

$$\begin{aligned} \ln L(\beta) &= \sum_{i=1}^{n} \ln f(x_i,\beta) \\ &= 6\sum_{i=1}^{n}\ln x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i - n\ln(6!) - 7n\ln(\beta). \end{aligned}$$

Therefore

$$\frac{d}{d\beta}\ln L(\beta) = \frac{1}{\beta^2}\sum_{i=1}^{n} x_i - \frac{7n}{\beta}.$$

Setting this derivative $\frac{d}{d\beta}\ln L(\beta)$ to zero, we get

$$\frac{1}{\beta^2}\sum_{i=1}^{n} x_i - \frac{7n}{\beta} = 0$$

which yields

$$\beta = \frac{1}{7n}\sum_{i=1}^{n} x_i.$$

This $\beta$ can be shown to be maximum by the second derivative test and again we leave this verification to the reader. Hence, the estimator of $\beta$ is given by

$$\widehat{\beta} = \frac{1}{7}\overline{X}.$$

**Remark 15.1.** Note that this maximum likelihood estimator of $\beta$ is same as the one found for $\beta$ using the moment method in Example 15.6. However, in general the estimators by different methods are different as the following example illustrates.

**Example 15.11.** If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the maximum likelihood estimator of $\theta$ ?

**Answer:** The likelihood function of the sample is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \left(\frac{1}{\theta}\right) \qquad \theta > x_i \quad (i = 1, 2, 3, ..., n) \\ &= \left(\frac{1}{\theta}\right)^n \qquad \theta > \max\{x_1, x_2, ..., x_n\}. \end{aligned}$$

Hence the parameter space of $\theta$ with respect to $L(\theta)$ is given by

$$\Omega = \{\theta \in \mathbb{R} \mid x_{\max} < \theta < \infty\} = (x_{\max}, \infty).$$

Now we maximize $L(\theta)$ on $\Omega$. First, we compute $\ln L(\theta)$ and then differentiate it to get

$$\ln L(\theta) = -n \ln(\theta)$$

and

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{n}{\theta} < 0.$$

Therefore $\ln L(\theta)$ is a decreasing function of $\theta$ and as such the maximum of $\ln L(\theta)$ occurs at the left end point of the interval $(x_{\max}, \infty)$. Therefore, at

$\theta = x_{\max}$ the likelihood function achieve maximum. Hence the likelihood estimator of $\theta$ is given by

$$\widehat{\theta} = X_{(n)}$$

where $X_{(n)}$ denotes the $n^{\text{th}}$ order statistic of the given sample.

Thus, Example 15.7 and Example 15.11 say that the if we estimate the parameter $\theta$ of a distribution with uniform density on the interval $(0, \theta)$, then the maximum likelihood estimator is given by

$$\widehat{\theta} = X_{(n)}$$

where as

$$\widehat{\theta} = 2\,\overline{X}$$

is the estimator obtained by the method of moment. Hence, in general these two methods do not provide the same estimator of an unknown parameter.

**Example 15.12.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \sqrt{\frac{2}{\pi}}\; e^{-\frac{1}{2}(x-\theta)^2} & \text{if } x \geq \theta \\ 0 & \text{elsewhere.} \end{cases}$$

What is the maximum likelihood estimator of $\theta$ ?

**Answer:** The likelihood function $L(\theta)$ is given by

$$L(\theta) = \left(\sqrt{\frac{2}{\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{1}{2}(x_i-\theta)^2} \qquad x_i \geq \theta \ (i = 1, 2, 3, ..., n).$$

Hence the parameter space of $\theta$ is given by

$$\Omega = \{\theta \in \mathbb{R} \mid 0 \leq \theta \leq x_{\min}\} = [0, x_{\min}],,$$

where $x_{\min} = \min\{x_1, x_2, ..., x_n\}$. Now we evaluate the logarithm of the likelihood function.

$$\ln L(\theta) = \frac{n}{2}\, \ln\left(\frac{2}{\pi}\right) - \frac{1}{2}\, \sum_{i=1}^{n}(x_i - \theta)^2,$$

where $\theta$ is on the interval $[0,\, x_{\min}\,]$. Now we maximize $\ln L(\theta)$ subject to the condition $0 \leq \theta \leq x_{\min}$. Taking the derivative, we get

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{1}{2} \sum_{i=1}^{n}(x_i - \theta)\, 2(-1) = \sum_{i=1}^{n}(x_i - \theta).$$

In this example, if we equate the derivative to zero, then we get $\theta = \overline{x}$. But this value of $\theta$ is not on the parameter space $\Omega$. Thus, $\theta = \overline{x}$ is not the solution. Hence to find the solution of this optimization process, we examine the behavior of the $\ln L(\theta)$ on the interval $[0,\, x_{\min}\,]$. Note that

$$\frac{d}{d\theta}\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)\,2(-1) = \sum_{i=1}^{n}(x_i-\theta) > 0$$

since each $x_i$ is bigger than $\theta$. Therefore, the function $\ln L(\theta)$ is an increasing function on the interval $[0,\, x_{\min}\,]$ and as such it will achieve maximum at the right end point of the interval $[0,\, x_{\min}\,]$. Therefore, the maximum likelihood estimator of $\theta$ is given by

$$\widehat{X} = X_{(1)}$$

where $X_{(1)}$ denotes the smallest observation in the random sample $X_1, X_2, ..., X_n$.

**Example 15.13.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2$. What are the maximum likelihood estimators of $\mu$ and $\sigma^2$ ?

**Answer:** Since $X \sim N(\mu, \sigma^2)$, the probability density function of $X$ is given by

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The likelihood function of the sample is given by

$$L(\mu,\sigma) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

Hence, the logarithm of this likelihood function is given by

$$\ln L(\mu,\sigma) = -\frac{n}{2}\,\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\,\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the partial derivatives of $\ln L(\mu,\sigma)$ with respect to $\mu$ and $\sigma$, we get

$$\frac{\partial}{\partial\mu}\ln L(\mu,\sigma) = -\frac{1}{2\sigma^2}\,\sum_{i=1}^{n}(x_i-\mu)\,(-2) = \frac{1}{\sigma^2}\,\sum_{i=1}^{n}(x_i-\mu).$$

and

$$\frac{\partial}{\partial\sigma}\ln L(\mu,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\,\sum_{i=1}^{n}(x_i-\mu)^2.$$

Setting $\frac{\partial}{\partial \mu} \ln L(\mu, \sigma) = 0$ and $\frac{\partial}{\partial \sigma} \ln L(\mu, \sigma) = 0$, and solving for the unknown $\mu$ and $\sigma$, we get

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

Thus the maximum likelihood estimator of $\mu$ is

$$\widehat{\mu} = \overline{X}.$$

Similarly, we get

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

implies

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Again $\mu$ and $\sigma^2$ found by the first derivative test can be shown to be maximum using the second derivative test for the functions of two variables. Hence, using the estimator of $\mu$ in the above expression, we get the estimator of $\sigma^2$ to be

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

**Example 15.14.** Suppose $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise.} \end{cases}$$

Find the estimators of $\alpha$ and $\beta$ by the method of maximum likelihood.

**Answer:** The likelihood function of the sample is given by

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\beta - \alpha} = \left( \frac{1}{\beta - \alpha} \right)^n$$

for all $\alpha \leq x_i$ for $(i = 1, 2, ..., n)$ and for all $\beta \geq x_i$ for $(i = 1, 2, ..., n)$. Hence, the domain of the likelihood function is

$$\Omega = \{(\alpha, \beta) \mid 0 < \alpha \leq x_{(1)} \quad \text{and} \quad x_{(n)} \leq \beta < \infty\}.$$

It is easy to see that $L(\alpha, \beta)$ is maximum if $\alpha = x_{(1)}$ and $\beta = x_{(n)}$. Therefore, the maximum likelihood estimator of $\alpha$ and $\beta$ are

$$\widehat{\alpha} = X_{(1)} \qquad \text{and} \qquad \widehat{\beta} = X_{(n)}.$$

The maximum likelihood estimator $\widehat{\theta}$ of a parameter $\theta$ has a remarkable property known as the invariance property. This invariance property says that if $\widehat{\theta}$ is a maximum likelihood estimator of $\theta$, then $g(\widehat{\theta})$ is the maximum likelihood estimator of $g(\theta)$, where $g$ is a function from $\mathbb{R}^k$ to a subset of $\mathbb{R}^m$. This result was proved by Zehna in 1966. We state this result as a theorem without a proof.

**Theorem 15.1.** Let $\widehat{\theta}$ be a maximum likelihood estimator of a parameter $\theta$ and let $g(\theta)$ be a function of $\theta$. Then the maximum likelihood estimator of $g(\theta)$ is given by $g\left(\widehat{\theta}\right)$.

Now we give two examples to illustrate the importance of this theorem.

**Example 15.15.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2$. What are the maximum likelihood estimators of $\sigma$ and $\mu - \sigma$?

**Answer:** From Example 15.13, we have the maximum likelihood estimator of $\mu$ and $\sigma^2$ to be

$$\widehat{\mu} = \overline{X}$$

and

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 =: \Sigma^2 \text{ (say)}.$$

Now using the invariance property of the maximum likelihood estimator we have

$$\widehat{\sigma} = \Sigma$$

and

$$\widehat{\mu - \sigma} = \overline{X} - \Sigma.$$

**Example 15.16.** Suppose $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise.} \end{cases}$$

Find the estimator of $\sqrt{\alpha^2 + \beta^2}$ by the method of maximum likelihood.

**Answer:** From Example 15.14, we have the maximum likelihood estimator of $\alpha$ and $\beta$ to be

$$\widehat{\alpha} = X_{(1)} \qquad \text{and} \qquad \widehat{\beta} = X_{(n)},$$

respectively. Now using the invariance property of the maximum likelihood estimator we see that the maximum likelihood estimator of $\sqrt{\alpha^2 + \beta^2}$ is $\sqrt{X_{(1)}^2 + X_{(n)}^2}$.

The concept of information in statistics was introduced by Sir Ronald Fisher, and it is known as Fisher information.

**Definition 15.4.** Let $X$ be an observation from a population with probability density function $f(x;\theta)$. Suppose $f(x;\theta)$ is continuous, twice differentiable and it's support does not depend on $\theta$. Then the Fisher information, $I(\theta)$, in a single observation $X$ about $\theta$ is given by

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{d \ln f(x;\theta)}{d\theta}\right]^2 f(x;\theta)\, dx.$$

Thus $I(\theta)$ is the expected value of the square of the random variable $\frac{d \ln f(X;\theta)}{d\theta}$. That is,

$$I(\theta) = E\left(\left[\frac{d \ln f(X;\theta)}{d\theta}\right]^2\right).$$

In the following lemma, we give an alternative formula for the Fisher information.

**Lemma 15.1.** The Fisher information contained in a single observation about the unknown parameter $\theta$ can be given alternatively as

$$I(\theta) = -\int_{-\infty}^{\infty} \left[\frac{d^2 \ln f(x;\theta)}{d\theta^2}\right] f(x;\theta)\, dx.$$

**Proof:** Since $f(x;\theta)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x;\theta)\, dx = 1. \tag{3}$$

Differentiating (3) with respect to $\theta$, we get

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x;\theta)\, dx = 0.$$

Rewriting the last equality, we obtain

$$\int_{-\infty}^{\infty} \frac{df(x;\theta)}{d\theta} \frac{1}{f(x;\theta)} f(x;\theta)\, dx = 0$$

which is

$$\int_{-\infty}^{\infty} \frac{d \ln f(x;\theta)}{d\theta} f(x;\theta)\, dx = 0. \tag{4}$$

Now differentiating (4) with respect to $\theta$, we see that

$$\int_{-\infty}^{\infty} \left[ \frac{d^2 \ln f(x;\theta)}{d\theta^2} f(x;\theta) + \frac{d \ln f(x;\theta)}{d\theta} \frac{df(x;\theta)}{d\theta} \right] dx = 0.$$

Rewriting the last equality, we have

$$\int_{-\infty}^{\infty} \left[ \frac{d^2 \ln f(x;\theta)}{d\theta^2} f(x;\theta) + \frac{d \ln f(x;\theta)}{d\theta} \frac{df(x;\theta)}{d\theta} \frac{1}{f(x;\theta)} f(x;\theta) \right] dx = 0$$

which is

$$\int_{-\infty}^{\infty} \left( \frac{d^2 \ln f(x;\theta)}{d\theta^2} + \left[ \frac{d \ln f(x;\theta)}{d\theta} \right]^2 \right) f(x;\theta)\, dx = 0.$$

The last equality implies that

$$\int_{-\infty}^{\infty} \left[ \frac{d \ln f(x;\theta)}{d\theta} \right]^2 f(x;\theta)\, dx = -\int_{-\infty}^{\infty} \left[ \frac{d^2 \ln f(x;\theta)}{d\theta^2} \right] f(x;\theta)\, dx.$$

Hence using the definition of Fisher information, we have

$$I(\theta) = -\int_{-\infty}^{\infty} \left[ \frac{d^2 \ln f(x;\theta)}{d\theta^2} \right] f(x;\theta)\, dx$$

and the proof of the lemma is now complete.

The following two examples illustrate how one can determine Fisher information.

**Example 15.17.** Let $X$ be a single observation taken from a normal population with unknown mean $\mu$ and known variance $\sigma^2$. Find the Fisher information in a single observation $X$ about $\mu$.

**Answer:** Since $X \sim N(\mu, \sigma^2)$, the probability density of $X$ is given by

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Hence

$$\ln f(x;\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}.$$

Therefore

$$\frac{d\ln f(x;\mu)}{d\mu} = \frac{x-\mu}{\sigma^2}$$

and

$$\frac{d^2\ln f(x;\mu)}{d\mu^2} = -\frac{1}{\sigma^2}.$$

Hence

$$I(\mu) = -\int_{-\infty}^{\infty}\left(-\frac{1}{\sigma^2}\right) f(x;\mu)\,dx = \frac{1}{\sigma^2}.$$

**Example 15.18.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with unknown mean $\mu$ and known variance $\sigma^2$. Find the Fisher information in this sample of size $n$ about $\mu$.

**Answer:** Let $I_n(\mu)$ be the required Fisher information. Then from the definition, we have

$$\begin{aligned}
I_n(\mu) &= -E\left(\frac{d^2\ln f(X_1, X_2, ..., X_n;\mu}{d\mu^2}\right)\\
&= -E\left(\frac{d^2}{d\mu^2}\{\ln f(X_1;\mu) + \cdots + \ln f(X_n;\mu)\}\right)\\
&= -E\left(\frac{d^2\ln f(X_1;\mu)}{d\mu^2}\right) - \cdots - E\left(\frac{d^2\ln f(X_n;\mu)}{d\mu^2}\right)\\
&= I(\mu) + \cdots + I(\mu)\\
&= n\,I(\mu)\\
&= n\,\frac{1}{\sigma^2} \qquad \text{(using Example 15.17).}
\end{aligned}$$

This example shows that if $X_1, X_2, ..., X_n$ is a random sample from a population $X \sim f(x;\theta)$, then the Fisher information, $I_n(\theta)$, in a sample of size $n$ about the parameter $\theta$ is equal to $n$ times the Fisher information in $X$ about $\theta$. Thus

$$I_n(\theta) = n\,I(\theta).$$

If $X$ is a random variable with probability density function $f(x;\theta)$, where $\theta = (\theta_1, ..., \theta_n)$ is an unknown parameter vector then the Fisher information,

$I(\theta)$, is a $n \times n$ matrix given by

$$\begin{aligned} I(\theta) &= (I_{ij}(\theta)) \\ &= \left( -E\left( \frac{\partial^2 \ln f(X;\theta)}{\partial\theta_i \, \partial\theta_j} \right) \right). \end{aligned}$$

**Example 15.19.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2$. What is the Fisher information matrix, $I_n(\mu, \sigma^2)$, of the sample of size $n$ about the parameters $\mu$ and $\sigma^2$?

**Answer:** Let us write $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. The Fisher information, $I_n(\theta)$, in a sample of size $n$ about the parameter $(\theta_1, \theta_2)$ is equal to $n$ times the Fisher information in the population about $(\theta_1, \theta_2)$, that is

$$I_n(\theta_1, \theta_2) = n\, I(\theta_1, \theta_2). \tag{5}$$

Since there are two parameters $\theta_1$ and $\theta_2$, the Fisher information matrix $I(\theta_1, \theta_2)$ is a $2 \times 2$ matrix given by

$$I(\theta_1, \theta_2) = \begin{pmatrix} I_{11}(\theta_1, \theta_2) & I_{12}(\theta_1, \theta_2) \\ \\ I_{21}(\theta_1, \theta_2) & I_{22}(\theta_1, \theta_2) \end{pmatrix} \tag{6}$$

where

$$I_{ij}(\theta_1, \theta_2) = -E\left( \frac{\partial^2 \ln f(X;\theta_1, \theta_2)}{\partial\theta_i \, \partial\theta_j} \right)$$

for $i = 1, 2$ and $j = 1, 2$. Now we proceed to compute $I_{ij}$. Since

$$f(x;\theta_1, \theta_2) = \frac{1}{\sqrt{2\,\pi\,\theta_2}}\, e^{-\frac{(x-\theta_1)^2}{2\,\theta_2}}$$

we have

$$\ln f(x;\theta_1, \theta_2) = -\frac{1}{2}\, \ln(2\,\pi\,\theta_2) - \frac{(x-\theta_1)^2}{2\,\theta_2}.$$

Taking partials of $\ln f(x;\theta_1, \theta_2)$, we have

$$\begin{aligned} \frac{\partial \ln f(x;\theta_1, \theta_2)}{\partial\theta_1} &= \frac{x-\theta_1}{\theta_2}, \\ \frac{\partial \ln f(x;\theta_1, \theta_2)}{\partial\theta_2} &= -\frac{1}{2\,\theta_2} + \frac{(x-\theta_1)^2}{2\,\theta_2^2}, \\ \frac{\partial^2 \ln f(x;\theta_1, \theta_2)}{\partial\theta_1^2} &= -\frac{1}{\theta_2}, \\ \frac{\partial^2 \ln f(x;\theta_1, \theta_2)}{\partial\theta_2^2} &= \frac{1}{2\,\theta_2^2} - \frac{(x-\theta_1)^2}{\theta_2^3}, \\ \frac{\partial^2 \ln f(x;\theta_1, \theta_2)}{\partial\theta_1\, \partial\theta_2} &= -\frac{x-\theta_1}{\theta_2^2}. \end{aligned}$$

Hence

$$I_{11}(\theta_1, \theta_2) = -E\left(-\frac{1}{\theta_2}\right) = \frac{1}{\theta_2} = \frac{1}{\sigma^2}.$$

Similarly,

$$I_{21}(\theta_1, \theta_2) = I_{12}(\theta_1, \theta_2) = -E\left(-\frac{X-\theta_1}{\theta_2^2}\right) = \frac{E(X)}{\theta_2^2} - \frac{\theta_1}{\theta_2^2} = \frac{\theta_1}{\theta_2^2} - \frac{\theta_1}{\theta_2^2} = 0$$

and

$$\begin{aligned} I_{22}(\theta_1, \theta_2) &= -E\left(-\frac{(X-\theta_1)^2}{\theta_2^3} + \frac{1}{2\theta_2^2}\right) \\ &= \frac{E\left((X-\theta_1)^2\right)}{\theta_2^3} - \frac{1}{2\theta_2^2} = \frac{\theta_2}{\theta_2^3} - \frac{1}{2\theta_2^2} = \frac{1}{2\theta_2^2} = \frac{1}{2\sigma^4}. \end{aligned}$$

Thus from (5), (6) and the above calculations, the Fisher information matrix is given by

$$I_n(\theta_1, \theta_2) = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Now we present an important theorem about the maximum likelihood estimator without a proof.

**Theorem 15.2.** Under certain regularity conditions on the $f(x;\theta)$ the maximum likelihood estimator $\widehat{\theta}$ of $\theta$ based on a random sample of size $n$ from a population $X$ with probability density $f(x;\theta)$ is asymptotically normally distributed with mean $\theta$ and variance $\frac{1}{n\,I(\theta)}$. That is

$$\widehat{\theta}_{ML} \sim N\left(\theta, \frac{1}{n\,I(\theta)}\right) \qquad \text{as } n \to \infty.$$

The following example shows that the maximum likelihood estimator of a parameter is not necessarily unique.

**Example 15.20.** If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{2} & \text{if } \theta - 1 \le x \le \theta + 1 \\ 0 & \text{otherwise,} \end{cases}$$

then what is the maximum likelihood estimator of $\theta$?

**Answer:** The likelihood function of this sample is given by

$$L(\theta) = \begin{cases} \left(\frac{1}{2}\right)^n & \text{if } \max\{x_1, ..., x_n\} - 1 \leq \theta \leq \min\{x_1, ..., x_n\} + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Since the likelihood function is a constant, any value in the interval $[\max\{x_1, ..., x_n\} - 1,\ \min\{x_1, ..., x_n\} + 1]$ is a maximum likelihood estimate of $\theta$.

**Example 15.21.** What is the basic principle of maximum likelihood estimation?

**Answer:** To choose a value of the parameter for which the observed data have as high a probability or density as possible. In other words a maximum likelihood estimate is a parameter value under which the sample data have the highest probability.

### 15.3. Bayesian Method

In the classical approach, the parameter $\theta$ is assumed to be an unknown, but fixed quantity. A random sample $X_1, X_2, ..., X_n$ is drawn from a population with probability density function $f(x; \theta)$ and based on the observed values in the sample, knowledge about the value of $\theta$ is obtained.

In Bayesian approach $\theta$ is considered to be a quantity whose variation can be described by a probability distribution (known as the prior distribution). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (and hence the name prior distribution). A sample is then taken from a population where $\theta$ is a parameter and the prior distribution is updated with this sample information. This updated prior is called the posterior distribution. The updating is done with the help of Bayes' theorem and hence the name Bayesian method.

In this section, we shall denote the population density $f(x; \theta)$ as $f(x/\theta)$, that is the density of the population $X$ given the parameter $\theta$.

**Definition 15.5.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where $\theta$ is the unknown parameter to be estimated. The probability density function of the random variable $\theta$ is called the prior distribution of $\theta$ and usually denoted by $h(\theta)$.

**Definition 15.6.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where $\theta$ is the unknown parameter to be estimated. The

conditional density, $k(\theta/x_1, x_2, ..., x_n)$, of $\theta$ given the sample $x_1, x_2, ..., x_n$ is called the posterior distribution of $\theta$.

**Example 15.22.** Let $X_1 = 1$, $X_2 = 2$ be a random sample of size 2 from a distribution with probability density function

$$f(x/\theta) = \binom{3}{x}\theta^x(1-\theta)^{3-x}, \qquad x = 0, 1, 2, 3.$$

If the prior density of $\theta$ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{otherwise,} \end{cases}$$

what is the posterior distribution of $\theta$ ?

**Answer:** Since $h(\theta)$ is the probability density of $\theta$, we should get

$$\int_{\frac{1}{2}}^{1} h(\theta)\, d\theta = 1$$

which implies

$$\int_{\frac{1}{2}}^{1} k\, d\theta = 1.$$

Therefore $k = 2$. The joint density of the sample and the parameter is given by

$$\begin{aligned} u(x_1, x_2, \theta) &= f(x_1/\theta)f(x_2/\theta)h(\theta) \\ &= \binom{3}{x_1}\theta^{x_1}(1-\theta)^{3-x_1}\binom{3}{x_2}\theta^{x_2}(1-\theta)^{3-x_2}\, 2 \\ &= 2\binom{3}{x_1}\binom{3}{x_2}\theta^{x_1+x_2}(1-\theta)^{6-x_1-x_2}. \end{aligned}$$

Hence,

$$\begin{aligned} u(1, 2, \theta) &= 2\binom{3}{1}\binom{3}{2}\theta^3(1-\theta)^3 \\ &= 18\,\theta^3(1-\theta)^3. \end{aligned}$$

The marginal distribution of the sample

$$\begin{aligned} g(1,2) &= \int_{\frac{1}{2}}^{1} u(1,2,\theta)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} 18\,\theta^3(1-\theta)^3\, d\theta \\ &= 18\int_{\frac{1}{2}}^{1} \theta^3\left(1+3\theta^2-3\theta-\theta^3\right)\, d\theta \\ &= 18\int_{\frac{1}{2}}^{1} \left(\theta^3+3\theta^5-3\theta^4-\theta^6\right)\, d\theta \\ &= \frac{9}{140}. \end{aligned}$$

The conditional distribution of the parameter $\theta$ given the sample $X_1 = 1$ and $X_2 = 2$ is given by

$$\begin{aligned} k(\theta/x_1=1,x_2=2) &= \frac{u(1,2,\theta)}{g(1,2)} \\ &= \frac{18\,\theta^3\,(1-\theta)^3}{\frac{9}{140}} \\ &= 280\,\theta^3\,(1-\theta)^3. \end{aligned}$$

Therefore, the posterior distribution of $\theta$ is

$$k(\theta/x_1=1,x_2=2) = \begin{cases} 280\,\theta^3\,(1-\theta)^3 & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

**Remark 15.2.** If $X_1, X_2, ..., X_n$ is a random sample from a population with density $f(x/\theta)$, then the joint density of the sample and the parameter is given by

$$u(x_1, x_2, ..., x_n, \theta) = h(\theta)\prod_{i=1}^{n} f(x_i/\theta).$$

Given this joint density, the marginal density of the sample can be computed using the formula

$$g(x_1, x_2, ..., x_n) = \int_{-\infty}^{\infty} h(\theta)\prod_{i=1}^{n} f(x_i/\theta)\, d\theta.$$

Now using the Bayes rule, the posterior distribution of $\theta$ can be computed as follows:

$$k(\theta/x_1, x_2, ..., x_n) = \frac{h(\theta)\prod_{i=1}^{n} f(x_i/\theta)}{\int_{-\infty}^{\infty} h(\theta)\prod_{i=1}^{n} f(x_i/\theta)\, d\theta}.$$

In Bayesian method, we use two types of loss functions.

**Definition 15.7.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where $\theta$ is the unknown parameter to be estimated. Let $\widehat{\theta}$ be an estimator of $\theta$. The function

$$\mathcal{L}_2\left(\widehat{\theta}, \theta\right) = \left(\widehat{\theta} - \theta\right)^2$$

is called the squared error loss. The function

$$\mathcal{L}_1\left(\widehat{\theta}, \theta\right) = \left|\widehat{\theta} - \theta\right|$$

is called the absolute error loss.

The loss function $\mathcal{L}$ represents the 'loss' incurred when $\widehat{\theta}$ is used in place of the parameter $\theta$.

**Definition 15.8.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where $\theta$ is the unknown parameter to be estimated. Let $\widehat{\theta}$ be an estimator of $\theta$ and let $\mathcal{L}\left(\widehat{\theta}, \theta\right)$ be a given loss function. The expected value of this loss function with respect to the population distribution $f(x/\theta)$, that is

$$R_{\mathcal{L}}(\theta) = \int \mathcal{L}\left(\widehat{\theta}, \theta\right)\, f(x/\theta)\, dx$$

is called the risk.

The posterior density of the parameter $\theta$ given the sample $x_1, x_2, ..., x_n$, that is

$$k(\theta/x_1, x_2, ..., x_n)$$

contains all information about $\theta$. In Bayesian estimation of parameter one chooses an estimate $\widehat{\theta}$ for $\theta$ such that

$$k(\widehat{\theta}/x_1, x_2, ..., x_n)$$

is maximum subject to a loss function. Mathematically, this is equivalent to minimizing the integral

$$\int_{\Omega} \mathcal{L}\left(\widehat{\theta}, \theta\right)\, k(\theta/x_1, x_2, ..., x_n)\, d\theta$$

with respect to $\widehat{\theta}$, where $\Omega$ denotes the support of the prior density $h(\theta)$ of the parameter $\theta$.

**Example 15.23.** Suppose one observation was taken of a random variable $X$ which yielded the value 2. The density function for $X$ is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

and prior distribution for parameter $\theta$ is

$$h(\theta) = \begin{cases} \frac{3}{\theta^4} & \text{if } 1 < \theta < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

If the loss function is $\mathcal{L}(z, \theta) = (z - \theta)^2$, then what is the Bayes' estimate for $\theta$ ?

**Answer:** The prior density of the random variable $\theta$ is

$$h(\theta) = \begin{cases} \frac{3}{\theta^4} & \text{if } 1 < \theta < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

The probability density function of the population is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the joint probability density function of the sample and the parameter is given by

$$\begin{aligned} u(x, \theta) &= h(\theta)\, f(x/\theta) \\ &= \frac{3}{\theta^4}\,\frac{1}{\theta} \\ &= \begin{cases} 3\,\theta^{-5} & \text{if } 0 < x < \theta \quad \text{and} \quad 1 < \theta < \infty \\ \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The marginal density of the sample is given by

$$\begin{aligned} g(x) &= \int_x^\infty u(x, \theta)\, d\theta \\ &= \int_x^\infty 3\,\theta^{-5}\, d\theta \\ &= \frac{3}{4}\, x^{-4} \\ &= \frac{3}{4\, x^4}. \end{aligned}$$

Thus, if $x = 2$, then $g(2) = \frac{3}{64}$. The posterior density of $\theta$ when $x = 2$ is given by

$$\begin{aligned} k(\theta/x = 2) &= \frac{u(2,\theta)}{g(2)} \\ &= \frac{64}{3}\, 3\, \theta^{-5} \\ &= \begin{cases} 64\, \theta^{-5} & \text{if } 2 < \theta < \infty \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

Now, we find the Bayes estimator by minimizing the expression $E\left[\mathcal{L}(\theta, z)/x = 2\right]$. That is

$$\widehat{\theta} = Arg \max_{z \in \Omega} \int_{\Omega} \mathcal{L}(\theta, z)\, k(\theta/x = 2)\, d\theta.$$

Let us call this integral $\psi(z)$. Then

$$\begin{aligned} \psi(z) &= \int_{\Omega} \mathcal{L}(\theta, z)\, k(\theta/x = 2)\, d\theta \\ &= \int_{2}^{\infty} (z - \theta)^2\, k(\theta/x = 2)\, d\theta \\ &= \int_{2}^{\infty} (z - \theta)^2\, 64\theta^{-5}\, d\theta. \end{aligned}$$

We want to find the value of $z$ which yields a minimum of $\psi(z)$. This can be done by taking the derivative of $\psi(z)$ and evaluating where the derivative is zero.

$$\begin{aligned} \frac{d}{dz}\psi(z) &= \frac{d}{dz} \int_{2}^{\infty} (z - \theta)^2\, 64\theta^{-5}\, d\theta \\ &= 2 \int_{2}^{\infty} (z - \theta)\, 64\theta^{-5}\, d\theta \\ &= 2 \int_{2}^{\infty} z\, 64\theta^{-5}\, d\theta - 2 \int_{2}^{\infty} \theta\, 64\theta^{-5}\, d\theta \\ &= 2\, z - \frac{16}{3}. \end{aligned}$$

Setting this derivative of $\psi(z)$ to zero and solving for $z$, we get

$$\begin{aligned} 2z - \frac{16}{3} &= 0 \\ \Rightarrow z &= \frac{8}{3}. \end{aligned}$$

Since $\frac{d^2\psi(z)}{dz^2} = 2$, the function $\psi(z)$ has a minimum at $z = \frac{8}{3}$. Hence, the Bayes' estimate of $\theta$ is $\frac{8}{3}$.

In Example 15.23, we have found the Bayes' estimate of $\theta$ by directly minimizing the $\int_\Omega \mathcal{L}\left(\widehat{\theta},\theta\right) k(\theta/x_1,x_2,...,x_n)\,d\theta$ with respect to $\widehat{\theta}$. The next result is very useful while finding the Bayes' estimate using a quadratic loss function. Notice that if $\mathcal{L}(\widehat{\theta},\theta) = (\theta-\widehat{\theta})^2$, then $\int_\Omega \mathcal{L}\left(\widehat{\theta},\theta\right) k(\theta/x_1,x_2,...,x_n)\,d\theta$ is $E\left((\theta-\widehat{\theta})^2/x_1,x_2,...,x_n\right)$. The following theorem is based on the fact that the function $\phi$ defined by $\phi(c) = E\left[(X-c)^2\right]$ attains minimum if $c = E[X]$.

**Theorem 15.3.** Let $X_1,X_2,...,X_n$ be a random sample from a distribution with density $f(x/\theta)$, where $\theta$ is the unknown parameter to be estimated. If the loss function is squared error, then the Bayes' estimator $\widehat{\theta}$ of parameter $\theta$ is given by

$$\widehat{\theta} = E(\theta/x_1,x_2,...,x_n),$$

where the expectation is taken with respect to density $k(\theta/x_1,x_2,...,x_n)$.

Now we give several examples to illustrate the use of this theorem.

**Example 15.24.** Suppose the prior distribution of $\theta$ is uniform over the interval $(0,1)$. Given $\theta$, the population $X$ is uniform over the interval $(0,\theta)$. If the squared error loss function is used, find the Bayes' estimator of $\theta$ based on a sample of size one.

**Answer:** The prior density of $\theta$ is given by

$$h(\theta) = \begin{cases} 1 & \text{if } 0<\theta<1 \\ \\ 0 & \text{otherwise .} \end{cases}$$

The density of population is given by

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0<x<\theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

The joint density of the sample and the parameter is given by

$$\begin{aligned} u(x,\theta) &= h(\theta)\,f(x/\theta) \\ &= 1\ \left(\frac{1}{\theta}\right) \\ &= \begin{cases} \frac{1}{\theta} & \text{if } 0<x<\theta<1 \\ \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

The marginal density of the sample is

$$\begin{aligned} g(x) &= \int_x^1 u(x,\theta)\, d\theta \\ &= \int_x^1 \frac{1}{\theta}\, d\theta \\ &= \begin{cases} -\ln x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The conditional density of $\theta$ given the sample is

$$k(\theta/x) = \frac{u(x,\theta)}{g(x)} = \begin{cases} -\frac{1}{\theta \ln x} & \text{if } 0 < x < \theta < 1 \\ 0 & \text{elsewhere .} \end{cases}$$

Since the loss function is quadratic error, therefore the Bayes' estimator of $\theta$ is

$$\begin{aligned} \widehat{\theta} &= E[\theta/x] \\ &= \int_x^1 \theta\, k(\theta/x)\, d\theta \\ &= \int_x^1 \theta\, \frac{-1}{\theta \ln x}\, d\theta \\ &= -\frac{1}{\ln x} \int_x^1 d\theta \\ &= \frac{x-1}{\ln x}. \end{aligned}$$

Thus, the Bayes' estimator of $\theta$ based on one observation $X$ is

$$\widehat{\theta} = \frac{X-1}{\ln X}.$$

**Example 15.25.** Given $\theta$, the random variable $X$ has a binomial distribution with $n = 2$ and probability of success $\theta$. If the prior density of $\theta$ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of $\theta$ for a squared error loss if $X = 1$ ?

**Answer:** Note that $\theta$ is uniform on the interval $\left(\frac{1}{2}, 1\right)$, hence $k = 2$. Therefore, the prior density of $\theta$ is

$$h(\theta) = \begin{cases} 2 & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The population density is given by

$$f(x/\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{2}{x} \theta^x (1-\theta)^{2-x}, \qquad x = 0, 1, 2.$$

The joint density of the sample and the parameter $\theta$ is

$$\begin{aligned} u(x, \theta) &= h(\theta)\, f(x/\theta) \\ &= 2 \binom{2}{x} \theta^x (1-\theta)^{2-x} \end{aligned}$$

where $\frac{1}{2} < \theta < 1$ and $x = 0, 1, 2$. The marginal density of the sample is given by

$$g(x) = \int_{\frac{1}{2}}^{1} u(x, \theta)\, d\theta.$$

This integral is easy to evaluate if we substitute $X = 1$ now. Hence

$$\begin{aligned} g(1) &= \int_{\frac{1}{2}}^{1} 2\binom{2}{1} \theta\,(1-\theta)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} \left(4\theta - 4\theta^2\right)\, d\theta \\ &= 4 \left[\frac{\theta^2}{2} - \frac{\theta^3}{3}\right]_{\frac{1}{2}}^{1} \\ &= \frac{2}{3} \left[3\theta^2 - 2\theta^3\right]_{\frac{1}{2}}^{1} \\ &= \frac{2}{3} \left[(3-2) - \left(\frac{3}{4} - \frac{2}{8}\right)\right] \\ &= \frac{1}{3}. \end{aligned}$$

Therefore, the posterior density of $\theta$ given $x = 1$, is

$$k(\theta/x = 1) = \frac{u(1, \theta)}{g(1)} = 12\,(\theta - \theta^2),$$

where $\frac{1}{2} < \theta < 1$. Since the loss function is quadratic error, therefore the

Bayes' estimate of $\theta$ is

$$\begin{aligned}\widehat{\theta} &= E[\theta/x=1] \\ &= \int_{\frac{1}{2}}^{1} \theta\, k(\theta/x=1)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} 12\,\theta\,(\theta-\theta^2)\, d\theta \\ &= \left[4\,\theta^3 - 3\,\theta^4\right]_{\frac{1}{2}}^{1} \\ &= 1 - \frac{5}{16} \\ &= \frac{11}{16}.\end{aligned}$$

Hence, based on the sample of size one with $X=1$, the Bayes' estimate of $\theta$ is $\frac{11}{16}$, that is

$$\widehat{\theta} = \frac{11}{16}.$$

The following theorem help us to evaluate the Bayes estimate of a sample if the loss function is absolute error loss. This theorem is based the fact that a function $\phi(c) = E\,[\,|X-c|\,]$ is minimum if $c$ is the median of $X$.

**Theorem 15.4.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where $\theta$ is the unknown parameter to be estimated. If the loss function is absolute error, then the Bayes estimator $\widehat{\theta}$ of the parameter $\theta$ is given by

$$\widehat{\theta} = \text{median of } k(\theta/x_1, x_2, ..., x_n)$$

where $k(\theta/x_1, x_2, ..., x_n)$ is the posterior distribution of $\theta$.

The followings are some examples to illustrate the above theorem.

**Example 15.26.** Given $\theta$, the random variable $X$ has a binomial distribution with $n=3$ and probability of success $\theta$. If the prior density of $\theta$ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of $\theta$ for an *absolute difference error loss* if the sample consists of one observation $x=3$?

**Answer:** Since, the prior density of $\theta$ is

$$h(\theta) = \begin{cases} 2 & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{otherwise ,} \end{cases}$$

and the population density is

$$f(x/\theta) = \binom{3}{x} \theta^x (1-\theta)^{3-x},$$

the joint density of the sample and the parameter is given by

$$u(3, \theta) = h(\theta)\, f(3/\theta) = 2\,\theta^3,$$

where $\frac{1}{2} < \theta < 1$. The marginal density of the sample (at $x = 3$) is given by

$$\begin{aligned} g(3) &= \int_{\frac{1}{2}}^{1} u(3, \theta)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} 2\,\theta^3\, d\theta \\ &= \left[\frac{\theta^4}{2}\right]_{\frac{1}{2}}^{1} \\ &= \frac{15}{32}. \end{aligned}$$

Therefore, the conditional density of $\theta$ given $X = 3$ is

$$k(\theta/x = 3) = \frac{u(3, \theta)}{g(3)} = \begin{cases} \frac{64}{15}\,\theta^3 & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Since, the loss function is absolute error, the Bayes' estimator is the median of the probability density function $k(\theta/x = 3)$. That is

$$\begin{aligned} \frac{1}{2} &= \int_{\frac{1}{2}}^{\widehat{\theta}} \frac{64}{15}\,\theta^3\, d\theta \\ &= \frac{64}{60}\, \left[\theta^4\right]_{\frac{1}{2}}^{\widehat{\theta}} \\ &= \frac{64}{60}\, \left[\left(\widehat{\theta}\right)^4 - \frac{1}{16}\right]. \end{aligned}$$

Solving the above equation for $\widehat{\theta}$, we get

$$\widehat{\theta} = \sqrt[4]{\frac{17}{32}} = 0.8537.$$

**Example 15.27.** Suppose the prior distribution of $\theta$ is uniform over the interval $(2, 5)$. Given $\theta$, $X$ is uniform over the interval $(0, \theta)$. What is the Bayes' estimator of $\theta$ for *absolute error loss* if $X = 1$ ?

**Answer:** Since, the prior density of $\theta$ is

$$h(\theta) = \begin{cases} \frac{1}{3} & \text{if } 2 < \theta < 5 \\ 0 & \text{otherwise ,} \end{cases}$$

and the population density is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{elsewhere,} \end{cases}$$

the joint density of the sample and the parameter is given by

$$u(x, \theta) = h(\theta)\, f(x/\theta) = \frac{1}{3\theta},$$

where $2 < \theta < 5$ and $0 < x < \theta$. The marginal density of the sample (at $x = 1$) is given by

$$\begin{aligned} g(1) &= \int_1^5 u(1, \theta)\, d\theta \\ &= \int_1^2 u(1, \theta)\, d\theta + \int_2^5 u(1, \theta)\, d\theta \\ &= \int_2^5 \frac{1}{3\theta}\, d\theta \\ &= \frac{1}{3} \ln\left(\frac{5}{2}\right). \end{aligned}$$

Therefore, the conditional density of $\theta$ given the sample $x = 1$, is

$$\begin{aligned} k(\theta/x = 1) &= \frac{u(1, \theta)}{g(1)} \\ &= \frac{1}{\theta \ln\left(\frac{5}{2}\right)}. \end{aligned}$$

Since, the loss function is absolute error, the Bayes estimate of $\theta$ is the median of $k(\theta/x=1)$. Hence

$$\frac{1}{2} = \int_2^{\widehat{\theta}} \frac{1}{\theta \ln\left(\frac{5}{2}\right)}\, d\theta$$

$$= \frac{1}{\ln\left(\frac{5}{2}\right)} \ln\left(\frac{\widehat{\theta}}{2}\right).$$

Solving for $\widehat{\theta}$, we get

$$\widehat{\theta} = \sqrt{10} = 3.16.$$

**Example 15.28.** What is the basic principle of Bayesian estimation?

**Answer:** The basic principle behind the Bayesian estimation method consists of choosing a value of the parameter $\theta$ for which the observed data have as high a posterior probability $k(\theta/x_1, x_2, ..., x_n)$ of $\theta$ as possible subject to a loss function.

### 15.4. Review Exercises

**1.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{2\theta} & \text{if} \quad -\theta < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the moment method find an estimator for the parameter $\theta$.

**2.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} (\theta+1)\, x^{-\theta-2} & \text{if} \quad 1 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the moment method find an estimator for the parameter $\theta$.

**3.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta^2\, x\, e^{-\theta\, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the moment method find an estimator for the parameter $\theta$.

**4.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if} \quad 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter $\theta$.

**5.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} (\theta+1)\, x^{-\theta-2} & \text{if} \quad 1 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter $\theta$.

**6.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta^2\, x\, e^{-\theta\, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter $\theta$.

**7.** Let $X_1, X_2, X_3, X_4$ be a random sample from a distribution with density function

$$f(x;\beta) = \begin{cases} \frac{1}{\beta}\, e^{\frac{-(x-4)}{\beta}} & \text{for} \quad x > 4 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta > 0$. If the data from this random sample are 8.2, 9.1, 10.6 and 4.9, respectively, what is the maximum likelihood estimate of $\beta$?

**8.** Given $\theta$, the random variable $X$ has a binomial distribution with $n = 2$ and probability of success $\theta$. If the prior density of $\theta$ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of $\theta$ for a squared error loss if the sample consists of $x_1 = 1$ and $x_2 = 2$.

**9.** Suppose two observations were taken of a random variable $X$ which yielded the values 2 and 3. The density function for $X$ is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

and prior distribution for the parameter $\theta$ is

$$h(\theta) = \begin{cases} 3\,\theta^{-4} & \text{if } \theta > 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

If the loss function is quadratic, then what is the Bayes' estimate for $\theta$?

**10.** The Pareto distribution is often used in study of incomes and has the *cumulative density function*

$$F(x;\alpha,\theta) = \begin{cases} 1 - \left(\frac{\alpha}{x}\right)^{\theta} & \text{if } \alpha \leq x \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \alpha < \infty$ and $1 < \theta < \infty$ are parameters. Find the maximum likelihood estimates of $\alpha$ and $\theta$ based on a sample of size 5 for value 3, 5, 2, 7, 8.

**11.** The Pareto distribution is often used in study of incomes and has the *cumulative density function*

$$F(x;\alpha,\theta) = \begin{cases} 1 - \left(\frac{\alpha}{x}\right)^{\theta} & \text{if } \alpha \leq x \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \alpha < \infty$ and $1 < \theta < \infty$ are parameters. Using moment methods find estimates of $\alpha$ and $\theta$ based on a sample of size 5 for value 3, 5, 2, 7, 8.

**12.** Suppose one observation was taken of a random variable $X$ which yielded the value 2. The density function for $X$ is

$$f(x/\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu)^2} \qquad -\infty < x < \infty,$$

and prior distribution of $\mu$ is

$$h(\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\mu^2} \qquad -\infty < \mu < \infty.$$

If the loss function is quadratic, then what is the Bayes' estimate for $\mu$?

**13.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with probability density

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{if } 2\theta \leq x \leq 3\theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. What is the maximum likelihood estimator of $\theta$?

**14.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with probability density

$$f(x) = \begin{cases} 1 - \theta^2 & \text{if } 0 \leq x \leq \frac{1}{1-\theta^2} \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. What is the maximum likelihood estimator of $\theta$?

**15.** Given $\theta$, the random variable $X$ has a binomial distribution with $n = 3$ and probability of success $\theta$. If the prior density of $\theta$ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of $\theta$ for a *absolute difference error loss* if the sample consists of one observation $x = 1$?

**16.** Suppose the random variable $X$ has the cumulative density function $F(x)$. Show that the expected value of the random variable $(X - c)^2$ is minimum if $c$ equals the expected value of $X$.

**17.** Suppose the continuous random variable $X$ has the cumulative density function $F(x)$. Show that the expected value of the random variable $|X - c|$ is minimum if $c$ equals the median of $X$ (that is, $F(c) = 0.5$).

**18.** Eight independent trials are conducted of a given system with the following results: $S$, $F$, $S$, $F$, $S$, $S$, $S$, $S$ where $S$ denotes the success and $F$ denotes the failure. What is the maximum likelihood estimate of the probability of successful operation $p$ ?

**19.** What is the maximum likelihood estimate of $\beta$ if the 5 values $\frac{4}{5}$, $\frac{2}{3}$, 1, $\frac{3}{2}$, $\frac{5}{4}$ were drawn from the population for which $f(x;\beta) = \frac{1}{2}(1+\beta)^5 \left(\frac{x}{2}\right)^{\beta}$ ?

**20.** If a sample of five values of $X$ is taken from the population for which $f(x;t) = 2(t-1)t^x$, what is the maximum likelihood estimator of $t$ ?

**21.** A sample of size $n$ is drawn from a gamma distribution

$$f(x;\beta) = \begin{cases} \frac{x^3\, e^{-\frac{x}{\beta}}}{6\beta^4} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is the maximum likelihood estimator of $\beta$ ?

**22.** The probability density function of the random variable $X$ is defined by

$$f(x;\lambda) = \begin{cases} 1 - \frac{2}{3}\lambda + \lambda\sqrt{x} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the maximum likelihood estimate of the parameter $\lambda$ based on two independent observations $x_1 = \frac{1}{4}$ and $x_2 = \frac{9}{16}$ ?

**23.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function $f(x;\sigma) = \frac{\sigma}{2}\, e^{-\sigma|x-\mu|}$. What is the maximum likelihood estimator of $\sigma$ ?

**24.** Suppose $X_1, X_2, ...$ are independent random variables, each with probability of success $p$ and probability of failure $1-p$, where $0 \le p \le 1$. Let $N$ be the number of observation needed to obtain the first success. What is the maximum likelihood estimator of $p$ in term of $N$ ?

**25.** Let $X_1, X_2, X_3$ and $X_4$ be a random sample from the discrete distribution $X$ such that

$$P(X = x) = \begin{cases} \frac{\theta^{2x}\, e^{-\theta^2}}{x!} & \text{for } x = 0, 1, 2, ..., \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. If the data are 17, 10, 32, 5, what is the maximum likelihood estimate of $\theta$ ?

**26.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population with a probability density function

$$f(x;\alpha,\lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\lambda x} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha$ and $\lambda$ are parameters. Using the moment method find the estimators for the parameters $\alpha$ and $\lambda$.

**27.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population distribution with the probability density function

$$f(x;p) = \binom{10}{x} p^x \, (1-p)^{10-x}$$

for $x = 0, 1, ..., 10$, where $p$ is a parameter. Find the Fisher information in the sample about the parameter $p$.

**28.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population distribution with the probability density function

$$f(x;\theta) = \begin{cases} \theta^2 \, x \, e^{-\theta \, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Find the Fisher information in the sample about the parameter $\theta$.

**29.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population distribution with the probability density function

$$f(x;\mu,\sigma^2) = \begin{cases} \frac{1}{x \, \sigma \, \sqrt{2 \, \pi}} \, e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}, & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ are unknown parameters. Find the Fisher information matrix in the sample about the parameters $\mu$ and $\sigma^2$.

**30.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population distribution with the probability density function

$$f(x;\mu,\lambda) = \begin{cases} \sqrt{\frac{\lambda}{2\pi}} \, x^{-\frac{3}{2}} \; e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}, & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $0 < \mu < \infty$ and $0 < \lambda < \infty$ are unknown parameters. Find the Fisher information matrix in the sample about the parameters $\mu$ and $\lambda$.

**31.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha) \, \theta^\alpha} \, x^{\alpha-1} \, e^{-\frac{x}{\theta}} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $\theta > 0$ are parameters. Using the moment method find estimators for parameters $\alpha$ and $\beta$.

**32.** Let $X_1, X_2, ..., X_n$ be a random sample of size$n$ from a distribution with a probability density function

$$f(x;\theta) = \frac{1}{\pi\ [1 + (x-\theta)^2]}, \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter $\theta$.

**33.** Let $X_1, X_2, ..., X_n$ be a random sample of size$n$ from a distribution with a probability density function

$$f(x;\theta) = \frac{1}{2}\, e^{-|x-\theta|}, \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter $\theta$.

**34.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population distribution with the probability density function

$$f(x;\lambda) = \begin{cases} \frac{\lambda^x\, e^{-\lambda}}{x!} & \text{if } x = 0, 1, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda > 0$ is an unknown parameter. Find the Fisher information matrix in the sample about the parameter $\lambda$.

**35.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population distribution with the probability density function

$$f(x;p) = \begin{cases} (1-p)^{x-1}p & \text{if } x = 1, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p < 1$ is an unknown parameter. Find the Fisher information matrix in the sample about the parameter $p$.

# Chapter 16

# CRITERIA FOR EVALUATING THE GOODNESS OF ESTIMATORS

We have seen in Chapter 15 that, in general, different parameter estimation methods yield different estimators. For example, if $X \sim UNIF(0, \theta)$ and $X_1, X_2, ..., X_n$ is a random sample from the population $X$, then the estimator of $\theta$ obtained by moment method is

$$\widehat{\theta}_{MM} = 2\overline{X}$$

where as the estimator obtained by the maximum likelihood method is

$$\widehat{\theta}_{ML} = X_{(n)}$$

where $\overline{X}$ and $X_{(n)}$ are the sample average and the $n^{\text{th}}$ order statistic, respectively. Now the question arises: which of the two estimators is better? Thus, we need some criteria to evaluate the goodness of an estimator. Some well known criteria for evaluating the goodness of an estimator are: (1) Unbiasedness, (2) Efficiency and Relative Efficiency, (3) Uniform Minimum Variance Unbiasedness, (4) Sufficiency, and (5) Consistency.

In this chapter, we shall examine only the first four criteria in details. The concepts of unbiasedness, efficiency and sufficiency were introduced by Sir Ronald Fisher.

### 16.1. The Unbiased Estimator

Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population with probability density function $f(x;\theta)$. An estimator $\widehat{\theta}$ of $\theta$ is a function of the random variables $X_1, X_2, ..., X_n$ which is free of the parameter $\theta$. An estimate is a realized value of an estimator that is obtained when a sample is actually taken.

**Definition 16.1.** An estimator $\widehat{\theta}$ of $\theta$ is said to be an unbiased estimator of $\theta$ if and only if

$$E\left(\widehat{\theta}\right) = \theta.$$

If $\widehat{\theta}$ is not unbiased, then it is called a biased estimator of $\theta$.

An estimator of a parameter may not equal to the actual value of the parameter for every realization of the sample $X_1, X_2, ..., X_n$, but if it is unbiased then on an average it will equal to the parameter.

**Example 16.1.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2 > 0$. Is the sample mean $\overline{X}$ an unbiased estimator of the parameter $\mu$ ?

**Answer:** Since, each $X_i \sim N(\mu, \sigma^2)$, we have

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

That is, the sample mean is normal with mean $\mu$ and variance $\frac{\sigma^2}{n}$. Thus

$$E\left(\overline{X}\right) = \mu.$$

Therefore, the sample mean $\overline{X}$ is an unbiased estimator of $\mu$.

**Example 16.2.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2 > 0$. What is the maximum likelihood estimator of $\sigma^2$ ? Is this maximum likelihood estimator an unbiased estimator of the parameter $\sigma^2$ ?

**Answer:** In Example 15.13, we have shown that the maximum likelihood estimator of $\sigma^2$ is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Now, we examine the unbiasedness of this estimator

$$\begin{aligned}
E\left[\widehat{\sigma^2}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= E\left[\frac{n-1}{n}\,\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= \frac{n-1}{n}\,E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= \frac{n-1}{n}\,E\left[S^2\right] \\
&= \frac{\sigma^2}{n}\,E\left[\frac{n-1}{\sigma^2}\,S^2\right] \qquad \left(\text{since } \frac{n-1}{\sigma^2}\,S^2 \sim \chi^2(n-1)\right) \\
&= \frac{\sigma^2}{n}E\left[\chi^2(n-1)\right] \\
&= \frac{\sigma^2}{n}\,(n-1) \\
&= \frac{n-1}{n}\,\sigma^2 \\
&\neq \sigma^2.
\end{aligned}$$

Therefore, the maximum likelihood estimator of $\sigma^2$ is a biased estimator.

Next, in the following example, we show that the sample variance $S^2$ given by the expression

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

is an unbiased estimator of the population variance $\sigma^2$ irrespective of the population distribution.

**Example 16.3.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with mean $\mu$ and variance $\sigma^2 > 0$. Is the sample variance $S^2$ an unbiased estimator of the population variance $\sigma^2$ ?

**Answer:** Note that the distribution of the population is not given. However, we are given $E(X_i) = \mu$ and $E[(X_i - \mu)^2] = \sigma^2$. In order to find $E\left(S^2\right)$, we need $E\left(\overline{X}\right)$ and $E\left(\overline{X}^2\right)$. Thus we proceed to find these two expected

values. Consider

$$E\left(\overline{X}\right) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Similarly,

$$Var\left(\overline{X}\right) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Therefore

$$E\left(\overline{X}^2\right) = Var\left(\overline{X}\right) + E\left(\overline{X}\right)^2 = \frac{\sigma^2}{n} + \mu^2.$$

Consider

$$E\left(S^2\right) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]$$
$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i^2 - 2\overline{X}X_i + \overline{X}^2\right)\right]$$
$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2\right]$$
$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left[X_i^2\right] - E\left[n\,\overline{X}^2\right]\right\}$$
$$= \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right]$$
$$= \frac{1}{n-1}\left[(n-1)\,\sigma^2\right]$$
$$= \sigma^2.$$

Therefore, the sample variance $S^2$ is an unbiased estimator of the population variance $\sigma^2$.

**Example 16.4.** Let $X$ be a random variable with mean 2. Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be unbiased estimators of the second and third moments, respectively, of $X$ about the origin. Find an unbiased estimator of the third moment of $X$ about its mean in terms of $\widehat{\theta_1}$ and $\widehat{\theta_2}$ .

**Answer:** Since, $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are the unbiased estimators of the second and third moments of $X$ about origin, we get

$$E\left(\widehat{\theta_1}\right) = E(X^2) \qquad \text{and} \qquad E\left(\widehat{\theta_2}\right) = E\left(X^3\right).$$

The unbiased estimator of the third moment of $X$ about its mean is

$$\begin{aligned} E\left[(X-2)^3\right] &= E\left[X^3 - 6X^2 + 12X - 8\right] \\ &= E\left[X^3\right] - 6E\left[X^2\right] + 12E\left[X\right] - 8 \\ &= \widehat{\theta_2} - 6\widehat{\theta_1} + 24 - 8 \\ &= \widehat{\theta_2} - 6\widehat{\theta_1} + 16. \end{aligned}$$

Thus, the unbiased estimator of the third moment of $X$ about its mean is $\widehat{\theta_2} - 6\widehat{\theta_1} + 16$.

**Example 16.5.** Let $X_1, X_2, ..., X_5$ be a sample of size 5 from the uniform distribution on the interval $(0, \theta)$, where $\theta$ is unknown. Let the estimator of $\theta$ be $k\, X_{\max}$, where $k$ is some constant and $X_{\max}$ is the largest observation. In order $k\, X_{\max}$ to be an unbiased estimator, what should be the value of the constant $k$ ?

**Answer:** The probability density function of $X_{\max}$ is given by

$$\begin{aligned} g(x) &= \frac{5!}{4!\,0!}\,\left[F(x)\right]^4\, f(x) \\ &= 5\,\left(\frac{x}{\theta}\right)^4\,\frac{1}{\theta} \\ &= \frac{5}{\theta^5}\,x^4. \end{aligned}$$

If $k\, X_{\max}$ is an unbiased estimator of $\theta$, then

$$\begin{aligned} \theta &= E\left(k\, X_{\max}\right) \\ &= k\, E\left(X_{\max}\right) \\ &= k\,\int_0^\theta x\, g(x)\, dx \\ &= k\,\int_0^\theta \frac{5}{\theta^5}\, x^5\, dx \\ &= \frac{5}{6}\, k\, \theta. \end{aligned}$$

Hence,

$$k = \frac{6}{5}.$$

**Example 16.6.** Let $X_1, X_2, ..., X_n$ be a sample of size $n$ from a distribution with unknown mean $-\infty < \mu < \infty$, and unknown variance $\sigma^2 > 0$. Show that the statistic $\overline{X}$ and $Y = \frac{X_1+2X_2+\cdots+nX_n}{\frac{n(n+1)}{2}}$ are both unbiased estimators of $\mu$. Further, show that $Var\left(\overline{X}\right) < Var(Y)$.

**Answer:** First, we show that $\overline{X}$ is an unbiased estimator of $\mu$

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu. \end{aligned}$$

Hence, the sample mean $\overline{X}$ is an unbiased estimator of the population mean irrespective of the distribution of $X$. Next, we show that $Y$ is also an unbiased estimator of $\mu$.

$$\begin{aligned} E\left(Y\right) &= E\left(\frac{X_1 + 2X_2 + \cdots + nX_n}{\frac{n(n+1)}{2}}\right) \\ &= \frac{2}{n\left(n+1\right)}\sum_{i=1}^{n} i\, E\left(X_i\right) \\ &= \frac{2}{n\left(n+1\right)}\sum_{i=1}^{n} i\,\mu \\ &= \frac{2}{n\left(n+1\right)}\,\mu\,\frac{n\left(n+1\right)}{2} \\ &= \mu. \end{aligned}$$

Hence, $\overline{X}$ and $Y$ are both unbiased estimator of the population mean irrespective of the distribution of the population. The variance of $\overline{X}$ is given by

$$\begin{aligned} Var\left[\overline{X}\right] &= Var\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n^2} Var\left[X_1 + X_2 + \cdots + X_n\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Similarly, the variance of $Y$ can be calculated as follows:

$$\begin{aligned}
Var\,[Y] &= Var\left[\frac{X_1 + 2X_2 + \cdots + nX_n}{\frac{n\,(n+1)}{2}}\right] \\
&= \frac{4}{n^2\,(n+1)^2}\,Var\,[1\,X_1 + 2\,X_2 + \cdots + n\,X_n] \\
&= \frac{4}{n^2\,(n+1)^2}\,\sum_{i=1}^{n} Var\,[i\,X_i] \\
&= \frac{4}{n^2\,(n+1)^2}\,\sum_{i=1}^{n} i^2\,Var\,[X_i] \\
&= \frac{4}{n^2\,(n+1)^2}\,\sigma^2\,\sum_{i=1}^{n} i^2 \\
&= \sigma^2\,\frac{4}{n^2\,(n+1)^2}\,\frac{n\,(n+1)\,(2n+1)}{6} \\
&= \frac{2}{3}\,\frac{2n+1}{(n+1)}\,\frac{\sigma^2}{n} \\
&= \frac{2}{3}\,\frac{2n+1}{(n+1)}\,Var\left[\,\overline{X}\,\right].
\end{aligned}$$

Since $\frac{2}{3}\,\frac{2n+1}{(n+1)} > 1$ for $n \geq 2$, we see that $Var\left[\,\overline{X}\,\right] < Var[\,Y\,]$. This shows that although the estimators $\overline{X}$ and $Y$ are both unbiased estimator of $\mu$, yet the variance of the sample mean $\overline{X}$ is smaller than the variance of $Y$.

In statistics, between two unbiased estimators one prefers the estimator which has the minimum variance. This leads to our next topic. However, before we move to the next topic we complete this section with some known disadvantages with the notion of unbiasedness. The first disadvantage is that an unbiased estimator for a parameter may not exist. The second disadvantage is that the property of unbiasedness is not invariant under functional transformation, that is, if $\widehat{\theta}$ is an unbiased estimator of $\theta$ and $g$ is a function, then $g(\widehat{\theta})$ may not be an unbiased estimator of $g(\theta)$.

### 16.2. The Relatively Efficient Estimator

We have seen that in Example 16.6 that the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

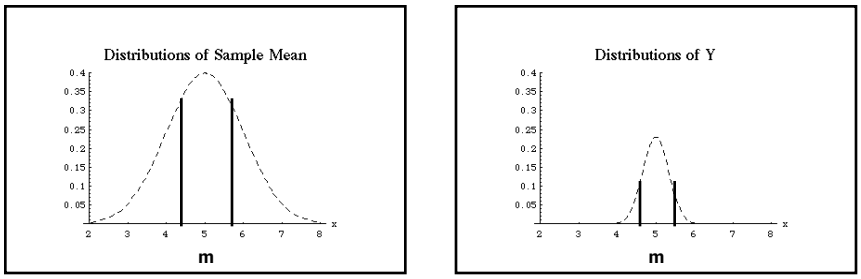
and the statistic

$$Y = \frac{X_1 + 2X_2 + \cdots + nX_n}{1 + 2 + \cdots + n}$$

are both unbiased estimators of the population mean. However, we also seen that

$$Var\left(\overline{X}\right) < Var(Y).$$

The following figure graphically illustrates the shape of the distributions of both the unbiased estimators.




If an unbiased estimator has a smaller variance or dispersion, then it has a greater chance of being close to true parameter $\theta$. Therefore when two estimators of $\theta$ are both unbiased, then one should pick the one with the smaller variance.

**Definition 16.2.** Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of $\theta$. The estimator $\widehat{\theta}_1$ is said to be more efficient than $\widehat{\theta}_2$ if

$$Var\left(\widehat{\theta}_1\right) < Var\left(\widehat{\theta}_2\right).$$

The ratio $\eta$ given by

$$\eta\left(\widehat{\theta}_1, \widehat{\theta}_2\right) = \frac{Var\left(\widehat{\theta}_2\right)}{Var\left(\widehat{\theta}_1\right)}$$

is called the relative efficiency of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$.

**Example 16.7.** Let $X_1, X_2, X_3$ be a random sample of size 3 from a population with mean $\mu$ and variance $\sigma^2 > 0$. If the statistics $\overline{X}$ and $Y$ given by

$$Y = \frac{X_1 + 2X_2 + 3X_3}{6}$$

are two unbiased estimators of the population mean $\mu$, then which one is more efficient?

**Answer:** Since $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, we get

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + X_3}{3}\right) \\ &= \frac{1}{3}\left(E(X_1) + E(X_2) + E(X_3)\right) \\ &= \frac{1}{3}\, 3\mu \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} E(Y) &= E\left(\frac{X_1 + 2X_2 + 3X_3}{6}\right) \\ &= \frac{1}{6}\left(E(X_1) + 2E(X_2) + 3E(X_3)\right) \\ &= \frac{1}{6}\, 6\mu \\ &= \mu. \end{aligned}$$

Therefore both $\overline{X}$ and $Y$ are unbiased. Next we determine the variance of both the estimators. The variances of these estimators are given by

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1 + X_2 + X_3}{3}\right) \\ &= \frac{1}{9}\left[Var(X_1) + Var(X_2) + Var(X_3)\right] \\ &= \frac{1}{9}\, 3\sigma^2 \\ &= \frac{12}{36}\, \sigma^2 \end{aligned}$$

and

$$\begin{aligned} Var(Y) &= Var\left(\frac{X_1 + 2X_2 + 3X_3}{6}\right) \\ &= \frac{1}{36}\left[Var(X_1) + 4Var(X_2) + 9Var(X_3)\right] \\ &= \frac{1}{36}\, 14\sigma^2 \\ &= \frac{14}{36}\, \sigma^2. \end{aligned}$$

Therefore

$$\frac{12}{36}\, \sigma^2 = Var\left(\overline{X}\right) < Var(Y) = \frac{14}{36}\, \sigma^2.$$

Hence, $\overline{X}$ is more efficient than the estimator $Y$. Further, the relative efficiency of $\overline{X}$ with respect to $Y$ is given by

$$\eta\left(\overline{X}, Y\right) = \frac{14}{12} = \frac{7}{6}.$$

**Example 16.8.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population with density

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

where $\theta > 0$ is a parameter. Are the estimators $X_1$ and $\overline{X}$ unbiased? Given, $X_1$ and $\overline{X}$, which one is more efficient estimator of $\theta$ ?

**Answer:** Since the population $X$ is exponential with parameter $\theta$, that is $X \sim EXP(\theta)$, the mean and variance of it are given by

$$E(X) = \theta \qquad \text{and} \qquad Var(X) = \theta^2.$$

Since $X_1, X_2, ..., X_n$ is a random sample from $X$, we see that the statistic $X_1 \sim EXP(\theta)$. Hence, the expected value of $X_1$ is $\theta$ and thus it is an unbiased estimator of the parameter $\theta$. Also, the sample mean is an unbiased estimator of $\theta$ since

$$\begin{aligned} E\left(\overline{X}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}\; n\theta \\ &= \theta. \end{aligned}$$

Next, we compute the variances of the unbiased estimators $X_1$ and $\overline{X}$. It is easy to see that

$$Var\left(X_1\right) = \theta^2$$

and

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2}\; \sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{1}{n^2}\; n\theta^2 \\ &= \frac{\theta^2}{n}. \end{aligned}$$

Hence

$$\frac{\theta^2}{n} = Var\left(\overline{X}\right) < Var\left(X_1\right) = \theta^2.$$

Thus $\overline{X}$ is more efficient than $X_1$ and the relative efficiency of $\overline{X}$ with respect to $X_1$ is

$$\eta(\overline{X},\, X_1) = \frac{\theta^2}{\frac{\theta^2}{n}} = n.$$

**Example 16.9.** Let $X_1, X_2, X_3$ be a random sample of size 3 from a population with density

$$f(x;\lambda) = \begin{cases} \frac{\lambda^x\, e^{-\lambda}}{x!} & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda$ is a parameter. Are the estimators given by

$$\widehat{\lambda_1} = \frac{1}{4}\,(X_1 + 2X_2 + X_3) \qquad \text{and} \qquad \widehat{\lambda_2} = \frac{1}{9}\,(4X_1 + 3X_2 + 2X_3)$$

unbiased? Given, $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$, which one is more efficient estimator of $\lambda$ ? Find an unbiased estimator of $\lambda$ whose variance is smaller than the variances of $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$.

**Answer:** Since each $X_i \sim POI(\lambda)$, we get

$$E\left(X_i\right) = \lambda \qquad \text{and} \qquad Var\left(X_i\right) = \lambda.$$

It is easy to see that

$$\begin{aligned} E\left(\widehat{\lambda_1}\right) &= \frac{1}{4}\left(E\left(X_1\right) + 2E\left(X_2\right) + E\left(X_3\right)\right) \\ &= \frac{1}{4}\ 4\lambda \\ &= \lambda, \end{aligned}$$

and

$$\begin{aligned} E\left(\widehat{\lambda_2}\right) &= \frac{1}{9}\left(4E\left(X_1\right) + 3E\left(X_2\right) + 2E\left(X_3\right)\right) \\ &= \frac{1}{9}\ 9\lambda \\ &= \lambda. \end{aligned}$$

Thus, both $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$ are unbiased estimators of $\lambda$. Now we compute their variances to find out which one is more efficient. It is easy to note that

$$\begin{aligned} Var\left(\widehat{\lambda_1}\right) &= \frac{1}{16}\left(Var\left(X_1\right) + 4Var\left(X_2\right) + Var\left(X_3\right)\right) \\ &= \frac{1}{16}\ 6\lambda \\ &= \frac{6}{16}\lambda \\ &= \frac{486}{1296}\lambda, \end{aligned}$$

and

$$\begin{aligned} Var\left(\widehat{\lambda_2}\right) &= \frac{1}{81}\left(16Var\left(X_1\right) + 9Var\left(X_2\right) + 4Var\left(X_3\right)\right) \\ &= \frac{1}{81}\ 29\lambda \\ &= \frac{29}{81}\lambda \\ &= \frac{464}{1296}\lambda, \end{aligned}$$

Since,

$$Var\left(\widehat{\lambda_2}\right) < Var\left(\widehat{\lambda_1}\right),$$

the estimator $\widehat{\lambda_2}$ is efficient than the estimator $\widehat{\lambda_1}$. We have seen in section 16.1 that the sample mean is always an unbiased estimator of the population mean irrespective of the population distribution. The variance of the sample mean is always equals to $\frac{1}{n}$ times the population variance, where $n$ denotes the sample size. Hence, we get

$$Var\left(\overline{X}\right) = \frac{\lambda}{3} = \frac{432}{1296}\lambda.$$

Therefore, we get

$$Var\left(\overline{X}\right) < Var\left(\widehat{\lambda_2}\right) < Var\left(\widehat{\lambda_1}\right).$$

Thus, the sample mean has even smaller variance than the two unbiased estimators given in this example.

In view of this example, now we have encountered a new problem. That is how to find an unbiased estimator which has the smallest variance among all unbiased estimators of a given parameter. We resolve this issue in the next section.

### 16.3. The Uniform Minimum Variance Unbiased Estimator

Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population with probability density function $f(x;\theta)$. Recall that an estimator $\widehat{\theta}$ of $\theta$ is a function of the random variables $X_1, X_2, ..., X_n$ which does depend on $\theta$.

**Definition 16.3.** An unbiased estimator $\widehat{\theta}$ of $\theta$ is said to be a uniform minimum variance unbiased estimator of $\theta$ if and only if

$$Var\left(\widehat{\theta}\right) \leq Var\left(\widehat{T}\right)$$

for any unbiased estimator $\widehat{T}$ of $\theta$.

If an estimator $\widehat{\theta}$ is unbiased then the mean of this estimator is equal to the parameter $\theta$, that is

$$E\left(\widehat{\theta}\right) = \theta$$

and the variance of $\widehat{\theta}$ is

$$\begin{aligned} Var\left(\widehat{\theta}\right) &= E\left[\left(\widehat{\theta} - E\left(\widehat{\theta}\right)\right)^2\right] \\ &= E\left[\left(\widehat{\theta} - \theta\right)^2\right]. \end{aligned}$$

This variance, if exists, is a function of the unbiased estimator $\widehat{\theta}$ and it has a minimum in the class of all unbiased estimators of $\theta$. Therefore we have an alternative definition of the uniform minimum variance unbiased estimator.

**Definition 16.4.** An unbiased estimator $\widehat{\theta}$ of $\theta$ is said to be a uniform minimum variance unbiased estimator of $\theta$ if it minimizes the variance $E\left[\left(\widehat{\theta} - \theta\right)^2\right]$.

**Example 16.10.** Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be unbiased estimators of $\theta$. Suppose $Var\left(\widehat{\theta_1}\right) = 1$, $Var\left(\widehat{\theta_2}\right) = 2$ and $Cov\left(\widehat{\theta_1}, \widehat{\theta_2}\right) = \frac{1}{2}$. What are the values of $c_1$ and $c_2$ for which $c_1\widehat{\theta_1} + c_2\widehat{\theta_2}$ is an unbiased estimator of $\theta$ with minimum variance among unbiased estimators of this type?

**Answer:** We want $c_1\widehat{\theta_1} + c_2\widehat{\theta_2}$ to be a minimum variance unbiased estimator of $\theta$. Then

$$\begin{aligned} & E\left[c_1\widehat{\theta_1} + c_2\widehat{\theta_2}\right] = \theta \\ \Rightarrow \quad & c_1\, E\left[\widehat{\theta_1}\right] + c_2\, E\left[\widehat{\theta_2}\right] = \theta \\ \Rightarrow \quad & c_1\,\theta + c_2\,\theta = \theta \\ \Rightarrow \quad & c_1 + c_2 = 1 \\ \Rightarrow \quad & c_2 = 1 - c_1. \end{aligned}$$

Therefore

$$\begin{aligned} Var\left[c_1\widehat{\theta_1}+c_2\widehat{\theta_2}\right] &= c_1^2\,Var\left[\widehat{\theta_1}\right]+c_2^2\,Var\left[\widehat{\theta_2}\right]+2\,c_1\,c_2\,Cov\left(\widehat{\theta_1},\,\widehat{\theta_1}\right)\\ &= c_1^2+2c_2^2+c_1c_2\\ &= c_1^2+2(1-c_1)^2+c_1(1-c_1)\\ &= 2(1-c_1)^2+c_1\\ &= 2+2c_1^2-3c_1. \end{aligned}$$

Hence, the variance $Var\left[c_1\widehat{\theta_1}+c_2\widehat{\theta_2}\right]$ is a function of $c_1$. Let us denote this function by $\phi(c_1)$, that is

$$\phi(c_1) := Var\left[c_1\widehat{\theta_1}+c_2\widehat{\theta_2}\right] = 2+2c_1^2-3c_1.$$

Taking the derivative of $\phi(c_1)$ with respect to $c_1$, we get

$$\frac{d}{dc_1}\phi(c_1) = 4c_1-3.$$

Setting this derivative to zero and solving for $c_1$, we obtain

$$4c_1-3=0 \qquad \Rightarrow \quad c_1=\frac{3}{4}.$$

Therefore

$$c_2 = 1-c_1 = 1-\frac{3}{4}=\frac{1}{4}.$$

In Example 16.10, we saw that if $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are any two unbiased estimators of $\theta$, then $c\,\widehat{\theta}_1+(1-c)\,\widehat{\theta}_2$ is also an unbiased estimator of $\theta$ for any $c\in\mathbb{R}$. Hence given two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$,

$$\mathcal{C}=\left\{\widehat{\theta}\mid \widehat{\theta}=c\,\widehat{\theta}_1+(1-c)\,\widehat{\theta}_2,\;\; c\in\mathbb{R}\right\}$$

forms an uncountable class of unbiased estimators of $\theta$. When the variances of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are known along with the their covariance, then in Example 16.10 we were able to determine the minimum variance unbiased estimator in the class $\mathcal{C}$. If the variances of the estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are not known, then it is very difficult to find the minimum variance estimator even in the class of estimators $\mathcal{C}$. Notice that $\mathcal{C}$ is a subset of the class of all unbiased estimators and finding a minimum variance unbiased estimator in this class is a difficult task.

One way to find a uniform minimum variance unbiased estimator for a parameter is to use the Cramér-Rao lower bound or the Fisher information inequality.

**Theorem 16.1.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population $X$ with probability density $f(x;\theta)$, where $\theta$ is a scalar parameter. Let $\widehat{\theta}$ be any unbiased estimator of $\theta$. Suppose the likelihood function $L(\theta)$ is a differentiable function of $\theta$ and satisfies

$$\begin{aligned}\frac{d}{d\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,...,x_n)\,L(\theta)\,dx_1\cdots dx_n \\ = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,...,x_n)\,\frac{d}{d\theta}L(\theta)\,dx_1\cdots dx_n\end{aligned} \tag{1}$$

for any $h(x_1,...,x_n)$ with $E(h(X_1,...,X_n)) < \infty$. Then

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right]}. \tag{CR1}$$

**Proof:** Since $L(\theta)$ is the joint probability density function of the sample $X_1, X_2, ..., X_n$,

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(\theta)\,dx_1\cdots dx_n = 1. \tag{2}$$

Differentiating (2) with respect to $\theta$ we have

$$\frac{d}{d\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(\theta)\,dx_1\cdots dx_n = 0$$

and use of (1) with $h(x_1,...,x_n) = 1$ yields

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{d}{d\theta}L(\theta)\;dx_1\cdots dx_n = 0. \tag{3}$$

Rewriting (3) as

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{dL(\theta)}{d\theta}\,\frac{1}{L(\theta)}\,L(\theta)\;dx_1\cdots dx_n = 0$$

we see that

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{d\ln L(\theta)}{d\theta}\,L(\theta)\;dx_1\cdots dx_n = 0.$$

Hence

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta \, \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 0. \tag{4}$$

Since $\widehat{\theta}$ is an unbiased estimator of $\theta$, we see that

$$E\left(\widehat{\theta}\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, L(\theta) \, dx_1 \cdots dx_n = \theta. \tag{5}$$

Differentiating (5) with respect to $\theta$, we have

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, L(\theta) \, dx_1 \cdots dx_n = 1.$$

Again using (1) with $h(X_1, ..., X_n) = \widehat{\theta}$, we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \; \frac{d}{d\theta} L(\theta) \; dx_1 \cdots dx_n = 1. \tag{6}$$

Rewriting (6) as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \; \frac{dL(\theta)}{d\theta} \; \frac{1}{L(\theta)} \, L(\theta) \; dx_1 \cdots dx_n = 1$$

we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \; \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 1. \tag{7}$$

From (4) and (7), we obtain

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \; \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 1. \tag{8}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 &= \left( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \; \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n \right)^2 \\ &\leq \left( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right)^2 \; L(\theta) \; dx_1 \cdots dx_n \right) \\ &\quad \cdot \left( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left( \frac{d \ln L(\theta)}{d\theta} \right)^2 \; L(\theta) \; dx_1 \cdots dx_n \right) \\ &= Var\left(\widehat{\theta}\right) \; E\left[ \left( \frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right]. \end{aligned}$$

Therefore

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}$$

and the proof of theorem is now complete.

If $L(\theta)$ is twice differentiable with respect to $\theta$, the inequality (CR1) can be stated equivalently as

$$Var\left(\widehat{\theta}\right) \geq \frac{-1}{E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]}. \tag{CR2}$$

The inequalities (CR1) and (CR2) are known as Cramér-Rao lower bound for the variance of $\widehat{\theta}$ or the Fisher information inequality. The condition (1) interchanges the order on integration and differentiation. Therefore any distribution whose range depend on the value of the parameter is not covered by this theorem. Hence distribution like the uniform distribution may not be analyzed using the Cramér-Rao lower bound.

If the estimator $\widehat{\theta}$ is minimum variance in addition to being unbiased, then equality holds. We state this as a theorem without giving a proof.

**Theorem 16.2.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population $X$ with probability density $f(x;\theta)$, where $\theta$ is a parameter. If $\widehat{\theta}$ is an unbiased estimator of $\theta$ and

$$Var\left(\widehat{\theta}\right) = \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]},$$

then $\widehat{\theta}$ is a uniform minimum variance unbiased estimator of $\theta$. The converse of this is not true.

**Definition 16.5.** An unbiased estimator $\widehat{\theta}$ is called an efficient estimator if it satisfies Cramér-Rao lower bound, that is

$$Var\left(\widehat{\theta}\right) = \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}.$$

In view of the above theorem it is easy to note that an efficient estimator of a parameter is always a uniform minimum variance unbiased estimator of

a parameter. However, not every uniform minimum variance unbiased estimator of a parameter is efficient. In other words not every uniform minimum variance unbiased estimators of a parameter satisfy the Cramér-Rao lower bound

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}.$$

**Example 16.11.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with density function

$$f(x;\theta) = \begin{cases} 3\theta\, x^2 e^{-\theta x^3} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the Cramér-Rao lower bound for the variance of unbiased estimator of the parameter $\theta$ ?

**Answer:** Let $\widehat{\theta}$ be an unbiased estimator of $\theta$. Cramér-Rao lower bound for the variance of $\widehat{\theta}$ is given by

$$Var\left(\widehat{\theta}\right) \geq \frac{-1}{E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]},$$

where $L(\theta)$ denotes the likelihood function of the given random sample $X_1, X_2, ..., X_n$. Since, the likelihood function of the sample is

$$L(\theta) = \prod_{i=1}^{n} 3\theta\, x_i^2 e^{-\theta x_i^3}$$

we get

$$\ln L(\theta) = n\, \ln \theta + \sum_{i=1}^{n} \ln\left(3x_i^2\right) - \theta \sum_{i=1}^{n} x_i^3.$$

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i^3,$$

and

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2}.$$

Hence, using this in the Cramér-Rao inequality, we get

$$Var\left(\widehat{\theta}\right) \geq \frac{\theta^2}{n}.$$

Thus the Cramér-Rao lower bound for the variance of the unbiased estimator of $\theta$ is $\frac{\theta^2}{n}$.

**Example 16.12.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with unknown mean $\mu$ and known variance $\sigma^2 > 0$. What is the maximum likelihood estimator of $\mu$? Is this maximum likelihood estimator an efficient estimator of $\mu$?

**Answer:** The probability density function of the population is

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\,(x-\mu)^2}.$$

Thus

$$\ln f(x;\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\,(x-\mu)^2$$

and hence

$$\ln L(\mu) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the derivative of $\ln L(\mu)$ with respect to $\mu$, we get

$$\frac{d\ln L(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu).$$

Setting this derivative to zero and solving for $\mu$, we see that $\widehat{\mu} = \overline{X}$.

The variance of $\overline{X}$ is given by

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Next we determine the Cramér-Rao lower bound for the estimator $\overline{X}$. We already know that

$$\frac{d\ln L(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

and hence

$$\frac{d^2\ln L(\mu)}{d\mu^2} = -\frac{n}{\sigma^2}.$$

Therefore

$$E\left(\frac{d^2\ln L(\mu)}{d\mu^2}\right) = -\frac{n}{\sigma^2}$$

and

$$-\frac{1}{E\left(\frac{d^2 \ln L(\mu)}{d\mu^2}\right)} = \frac{\sigma^2}{n}.$$

Thus

$$Var\left(\overline{X}\right) = -\frac{1}{E\left(\frac{d^2 \ln L(\mu)}{d\mu^2}\right)}$$

and $\overline{X}$ is an efficient estimator of $\mu$. Since every efficient estimator is a uniform minimum variance unbiased estimator, therefore $\overline{X}$ is a uniform minimum variance unbiased estimator of $\mu$.

**Example 16.13.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with known mean $\mu$ and unknown variance $\sigma^2 > 0$. What is the maximum likelihood estimator of $\sigma^2$? Is this maximum likelihood estimator a uniform minimum variance unbiased estimator of $\sigma^2$?

**Answer:** Let us write $\theta = \sigma^2$. Then

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-\frac{1}{2\theta}(x-\mu)^2}$$

and

$$\ln L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Differentiating $\ln L(\theta)$ with respect to $\theta$, we have

$$\frac{d}{d\theta}\ln L(\theta) = -\frac{n}{2}\,\frac{1}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Setting this derivative to zero and solving for $\theta$, we see that

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Next we show that this estimator is unbiased. For this we consider

$$\begin{aligned} E\left(\widehat{\theta}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) \\ &= \frac{\sigma^2}{n}\, E\left(\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2\right) \\ &= \frac{\theta}{n}\, E(\,\chi^2(n)\,) \\ &= \frac{\theta}{n}\, n = \theta. \end{aligned}$$

Hence $\widehat{\theta}$ is an unbiased estimator of $\theta$. The variance of $\widehat{\theta}$ can be obtained as follows:

$$\begin{aligned} Var\left(\widehat{\theta}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\ &= \frac{\sigma^4}{n}\,Var\left(\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right) \\ &= \frac{\theta^2}{n^2}\,Var(\,\chi^2(n)\,) \\ &= \frac{\theta^2}{n^2}\;4\;\frac{n}{2} \\ &= \frac{2\theta^2}{n} = \frac{2\sigma^4}{n}. \end{aligned}$$

Finally we determine the Cramér-Rao lower bound for the variance of $\widehat{\theta}$. The second derivative of $\ln L(\theta)$ with respect to $\theta$ is

$$\frac{d^2\ln L(\theta)}{d\theta^2} = \frac{n}{2\theta^2} - \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Hence

$$\begin{aligned} E\left(\frac{d^2\ln L(\theta)}{d\theta^2}\right) &= \frac{n}{2\theta^2} - \frac{1}{\theta^3}E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) \\ &= \frac{n}{2\theta^2} - \frac{\theta}{\theta^3}E\left(\,\chi^2(n)\,\right) \\ &= \frac{n}{2\theta^2} - \frac{n}{\theta^2} \\ &= -\frac{n}{2\theta^2} \end{aligned}$$

Thus

$$-\frac{1}{E\left(\frac{d^2\ln L(\theta)}{d\theta^2}\right)} = \frac{\theta^2}{n} = \frac{2\sigma^4}{n}.$$

Therefore

$$Var\left(\widehat{\theta}\right) = -\frac{1}{E\left(\frac{d^2\ln L(\theta)}{d\theta^2}\right)}.$$

Hence $\widehat{\theta}$ is an efficient estimator of $\theta$. Since every efficient estimator is a uniform minimum variance unbiased estimator, therefore $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ is a uniform minimum variance unbiased estimator of $\sigma^2$.

**Example 16.14.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a normal population known mean $\mu$ and variance $\sigma^2 > 0$. Show that $S^2 =$

$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ is an unbiased estimator of $\sigma^2$. Further, show that $S^2$ can not attain the Cramér-Rao lower bound.

**Answer:** From Example 16.2, we know that $S^2$ is an unbiased estimator of $\sigma^2$. The variance of $S^2$ can be computed as follows:

$$\begin{aligned}
Var\left(S^2\right) &= Var\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)\\
&= \frac{\sigma^4}{(n-1)^2}\,Var\left(\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2\right)\\
&= \frac{\sigma^4}{(n-1)^2}\,Var(\,\chi^2(n-1)\,)\\
&= \frac{\sigma^4}{(n-1)^2}\;2\;(n-1)\\
&= \frac{2\sigma^4}{n-1}.
\end{aligned}$$

Next we let $\theta=\sigma^2$ and determine the Cramér-Rao lower bound for the variance of $S^2$. The second derivative of $\ln L(\theta)$ with respect to $\theta$ is

$$\frac{d^2\ln L(\theta)}{d\theta^2} = \frac{n}{2\theta^2}-\frac{1}{\theta^3}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Hence

$$\begin{aligned}
E\left(\frac{d^2\ln L(\theta)}{d\theta^2}\right) &= \frac{n}{2\theta^2}-\frac{1}{\theta^3}E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)\\
&= \frac{n}{2\theta^2}-\frac{\theta}{\theta^3}E\left(\,\chi^2(n)\,\right)\\
&= \frac{n}{2\theta^2}-\frac{n}{\theta^2}\\
&= -\frac{n}{2\theta^2}
\end{aligned}$$

Thus

$$-\frac{1}{E\left(\frac{d^2\ln L(\theta)}{d\theta^2}\right)} = \frac{\theta^2}{n} = \frac{2\sigma^4}{n}.$$

Hence

$$\frac{2\sigma^4}{n-1} = Var\left(S^2\right) > -\frac{1}{E\left(\frac{d^2\ln L(\theta)}{d\theta^2}\right)} = \frac{2\sigma^4}{n}.$$

This shows that $S^2$ can not attain the Cramér-Rao lower bound.

The disadvantages of Cramér-Rao lower bound approach are the followings: (1) Not every density function $f(x;\theta)$ satisfies the assumptions of Cramér-Rao theorem and (2) not every allowable estimator attains the Cramér-Rao lower bound. Hence in any one of these situations, one does not know whether an estimator is a uniform minimum variance unbiased estimator or not.

### 16.4. Sufficient Estimator

In many situations, we can not easily find the distribution of the estimator $\widehat{\theta}$ of a parameter $\theta$ even though we know the distribution of the population. Therefore, we have no way to know whether our estimator $\widehat{\theta}$ is unbiased or biased. Hence, we need some other criteria to judge the quality of an estimator. Sufficiency is one such criteria for judging the quality of an estimator.

Recall that an estimator of a population parameter is a function of the sample values that does not contain the parameter. An estimator summarizes the information found in the sample about the parameter. If an estimator summarizes just as much information about the parameter being estimated as the sample does, then the estimator is called a sufficient estimator.

**Definition 16.6.** Let $X \sim f(x;\theta)$ be a population and let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from this population $X$. An estimator $\widehat{\theta}$ of the parameter $\theta$ is said to be a sufficient estimator of $\theta$ if the conditional distribution of the sample given the estimator $\widehat{\theta}$ does not depend on the parameter $\theta$.

**Example 16.15.** If $X_1, X_2, ..., X_n$ is a random sample from the distribution with probability density function

$$f(x;\theta) = \begin{cases} \theta^x (1-\theta)^{1-x} & \text{if } x = 0, 1 \\ \\ 0 & \text{elsewhere }, \end{cases}$$

where $0 < \theta < 1$. Show that $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic of $\theta$.

**Answer:** First, we find the distribution of the sample. This is given by

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y (1-\theta)^{n-y}.$$

Since, each $X_i \sim BER(\theta)$, we have

$$Y = \sum_{i=1}^{n} X_i \sim BIN(n, \theta).$$

Therefore, the probability density function of $Y$ is given by

$$g(y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Further, since each $X_i \sim BER(\theta)$, the space of each $X_i$ is given by

$$R_{X_i} = \{ 0, 1 \}.$$

Therefore, the space of the random variable $Y = \sum_{i=1}^{n} X_i$ is given by

$$R_Y = \{ 0, 1, 2, 3, 4, ..., n \}.$$

Let $A$ be the event $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ and $B$ denotes the event $(Y = y)$. Then $A \subset B$ and therefore $A \bigcap B = A$. Now, we find the conditional density of the sample given the estimator $Y$, that is

$$\begin{aligned} f(x_1, x_2, ..., x_n / Y = y) &= P\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \,/ Y = y\right) \\ &= P(A/B) \\ &= \frac{P(A \bigcap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \\ &= \frac{f(x_1, x_2, ..., x_n)}{g(y)} \\ &= \frac{\theta^y (1-\theta)^{n-y}}{\binom{n}{y} \theta^y (1-\theta)^{n-y}} \\ &= \frac{1}{\binom{n}{y}}. \end{aligned}$$

Hence, the conditional density of the sample given the statistic $Y$ is independent of the parameter $\theta$. Therefore, by definition $Y$ is a sufficient statistic.

**Example 16.16.** If $X_1, X_2, ..., X_n$ is a random sample from the distribution with probability density function

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{elsewhere ,} \end{cases}$$

where $-\infty < \theta < \infty$. What is the maximum likelihood estimator of $\theta$ ? Is this maximum likelihood estimator sufficient estimator of $\theta$ ?

**Answer:** We have seen in Chapter 15 that the maximum likelihood estimator of $\theta$ is $Y = X_{(1)}$, that is the first order statistic of the sample. Let us find the probability density of this statistic, which is given by

$$g(y) = \frac{n!}{(n-1)!}\,[F(y)]^0\; f(y)\;[1-F(y)]^{n-1}$$

$$= n\, f(y)\;[1-F(y)]^{n-1}$$

$$= n\, e^{-(y-\theta)}\;\left[1-\left\{1-e^{-(y-\theta)}\right\}\right]^{n-1}$$

$$= n\, e^{n\theta}\, e^{-ny}.$$

The probability density of the random sample is

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} e^{-(x_i-\theta)}$$
$$= e^{n\theta}\, e^{-n\,\overline{x}},$$

where $n\overline{x} = \sum_{i=1}^{n} x_i$. Let $A$ be the event $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ and $B$ denotes the event $(Y = y)$. Then $A \subset B$ and therefore $A \bigcap B = A$. Now, we find the conditional density of the sample given the estimator $Y$, that is

$$\begin{aligned} f(x_1, x_2, ..., x_n / Y = y) &= P\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n\,/Y = y\right) \\ &= P(A/B) \\ &= \frac{P(A \bigcap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \\ &= \frac{f(x_1, x_2, ..., x_n)}{g(y)} \\ &= \frac{e^{n\theta}\, e^{-n\,\overline{x}}}{n\, e^{n\theta}\, e^{-n\,y}} \\ &= \frac{e^{-n\,\overline{x}}}{n\, e^{-n\,y}}. \end{aligned}$$

Hence, the conditional density of the sample given the statistic $Y$ is independent of the parameter $\theta$. Therefore, by definition $Y$ is a sufficient statistic.

We have seen that to verify whether an estimator is sufficient or not one has to examine the conditional density of the sample given the estimator. To compute this conditional density one has to use the density of the estimator. The density of the estimator is not always easy to find. Therefore, verifying the sufficiency of an estimator using this definition is not always easy. The following *factorization theorem* of Fisher and Neyman helps to decide when an estimator is sufficient.

**Theorem 16.3.** Let $X_1, X_2, ..., X_n$ denote a random sample with probability density function $f(x_1, x_2, ..., x_n; \theta)$, which depends on the population parameter $\theta$. The estimator $\widehat{\theta}$ is sufficient for $\theta$ if and only if

$$f(x_1, x_2, ..., x_n; \theta) = \phi(\widehat{\theta}, \theta)\, h(x_1, x_2, ..., x_n)$$

where $\phi$ depends on $x_1, x_2, ..., x_n$ only through $\widehat{\theta}$ and $h(x_1, x_2, ..., x_n)$ does not depend on $\theta$.

Now we give two examples to illustrate the factorization theorem.

**Example 16.17.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x; \lambda) = \begin{cases} \frac{\lambda^x\, e^{-\lambda}}{x!} & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\lambda > 0$ is a parameter. Find the maximum likelihood estimator of $\lambda$ and show that the maximum likelihood estimator of $\lambda$ is sufficient estimator of the parameter $\lambda$.

**Answer:** First, we find the density of the sample or the likelihood function of the sample. The likelihood function of the sample is given by

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} f(x_i; \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{n\overline{X}} e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)}. \end{aligned}$$

Taking the logarithm of the likelihood function, we get

$$\ln L(\lambda) = n\overline{x}\ \ln\lambda - n\lambda - \ln\prod_{i=1}^{n}(x_i!).$$

Therefore

$$\frac{d}{d\lambda}\ln L(\lambda) = \frac{1}{\lambda}\,n\overline{x} - n.$$

Setting this derivative to zero and solving for $\lambda$, we get

$$\lambda = \overline{x}.$$

The second derivative test assures us that the above $\lambda$ is a maximum. Hence, the maximum likelihood estimator of $\lambda$ is the sample mean $\overline{X}$. Next, we show that $\overline{X}$ is sufficient, by using the Factorization Theorem of Fisher and Neyman. We factor the joint density of the sample as

$$\begin{aligned} L(\lambda) &= \frac{\lambda^{n\overline{x}}e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)} \\ &= \left[\lambda^{n\overline{x}}\,e^{-n\lambda}\right]\ \frac{1}{\prod_{i=1}^{n}(x_i!)} \\ &= \phi(\overline{X},\lambda)\,h\left(x_1,x_2,...,x_n\right). \end{aligned}$$

Therefore, the estimator $\overline{X}$ is a sufficient estimator of $\lambda$.

**Example 16.18.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution with density function

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(x-\mu)^2},$$

where $-\infty < \mu < \infty$ is a parameter. Find the maximum likelihood estimator of $\mu$ and show that the maximum likelihood estimator of $\mu$ is a sufficient estimator.

**Answer:** We know that the maximum likelihood estimator of $\mu$ is the sample mean $\overline{X}$. Next, we show that this maximum likelihood estimator $\overline{X}$ is a

sufficient estimator of $\mu$. The joint density of the sample is given by

$$
\begin{aligned}
& f(x_1, x_2, ..., x_n; \mu) \\
&= \prod_{i=1}^{n} f(x_i; \mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\overline{x})+(\overline{x}-\mu)]^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\overline{x})^2+2(x_i-\overline{x})(\overline{x}-\mu)+(\overline{x}-\mu)^2\right]} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\overline{x})^2+(\overline{x}-\mu)^2\right]} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{n}{2}(\overline{x}-\mu)^2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2}
\end{aligned}
$$

Hence, by the Factorization Theorem, $\overline{X}$ is a sufficient estimator of the population mean.

Note that the probability density function of the Example 16.17 which is

$$
f(x;\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x = 0, 1, 2, ..., \infty \\ 0 & \text{elsewhere ,} \end{cases}
$$

can be written as

$$
f(x;\lambda) = e^{\{x \ln \lambda - \ln x! - \lambda\}}
$$

for $x = 0, 1, 2, ...$ This density function is of the form

$$
f(x;\lambda) = e^{\{K(x)A(\lambda)+S(x)+B(\lambda)\}}.
$$

Similarly, the probability density function of the Example 16.12, which is

$$
f(x;\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}
$$

can also be written as

$$f(x;\mu) = e^{\{x\mu - \frac{x^2}{2} - \frac{\mu^2}{2} - \frac{1}{2}\ln(2\pi)\}}.$$

This probability density function is of the form

$$f(x;\mu) = e^{\{K(x)A(\mu)+S(x)+B(\mu)\}}.$$

We have also seen that in both the examples, the sufficient estimators were the sample mean $\overline{X}$, which can be written as $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Our next theorem gives a general result in this direction. The following theorem is known as the Pitman-Koopman theorem.

**Theorem 16.4.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density function of the exponential form

$$f(x;\theta) = e^{\{K(x)A(\theta)+S(x)+B(\theta)\}}$$

on a support free of $\theta$. Then the statistic $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for the parameter $\theta$.

**Proof:** The joint density of the sample is

$$\begin{aligned} f(x_1, x_2, ..., x_n;\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} e^{\{K(x_i)A(\theta)+S(x_i)+B(\theta)\}} \\ &= e^{\left\{\sum_{i=1}^{n} K(x_i)A(\theta) + \sum_{i=1}^{n} S(x_i) + n\,B(\theta)\right\}} \\ &= e^{\left\{\sum_{i=1}^{n} K(x_i)A(\theta) + n\,B(\theta)\right\}} e^{\left[\sum_{i=1}^{n} S(x_i)\right]}. \end{aligned}$$

Hence by the Factorization Theorem the estimator $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for the parameter $\theta$. This completes the proof.

**Example 16.19.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Using the Pitman-Koopman Theorem find a sufficient estimator of $\theta$.

**Answer:** The Pitman-Koopman Theorem says that if the probability density function can be expressed in the form of

$$f(x;\theta) = e^{\{K(x)A(\theta)+S(x)+B(\theta)\}}$$

then $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for $\theta$. The given population density can be written as

$$\begin{aligned} f(x;\theta) &= \theta\, x^{\theta-1} \\ &= e^{\{\ln\left[\theta\, x^{\theta-1}\right]} \\ &= e^{\{\ln\theta+(\theta-1)\ln x\}}. \end{aligned}$$

Thus,

$$\begin{aligned} K(x) &= \ln x & A(\theta) &= \theta \\ S(x) &= -\ln x & B(\theta) &= \ln\theta. \end{aligned}$$

Hence by Pitman-Koopman Theorem,

$$\begin{aligned} \sum_{i=1}^{n} K(X_i) &= \sum_{i=1}^{n} \ln X_i \\ &= \ln \prod_{i=1}^{n} X_i. \end{aligned}$$

Thus $\ln \prod_{i=1}^{n} X_i$ is a sufficient statistic for $\theta$.

**Remark 16.1.** Notice that $\prod_{i=1}^{n} X_i$ is also a sufficient statistic of $\theta$, since knowing $\ln\left(\prod_{i=1}^{n} X_i\right)$, we also know $\prod_{i=1}^{n} X_i$.

**Example 16.20.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

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

where $0 < \theta < \infty$ is a parameter. Find a sufficient estimator of $\theta$.

**Answer:** First, we rewrite the population density in the exponential form. That is

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta}\, e^{-\frac{x}{\theta}} \\ &= e^{\ln\left[\frac{1}{\theta}\, e^{-\frac{x}{\theta}}\right]} \\ &= e^{-\ln\theta - \frac{x}{\theta}}. \end{aligned}$$

Hence

$$\begin{aligned} K(x) &= x & A(\theta) &= -\frac{1}{\theta} \\ S(x) &= 0 & B(\theta) &= -\ln\theta. \end{aligned}$$

Hence by Pitman-Koopman Theorem,

$$\sum_{i=1}^{n} K(X_i) = \sum_{i=1}^{n} X_i = n\,\overline{X}.$$

Thus, $n\overline{X}$ is a sufficient statistic for $\theta$. Since knowing $n\overline{X}$, we also know $\overline{X}$, the estimator $\overline{X}$ is also a sufficient estimator of $\theta$.

**Example 16.21.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < \theta < \infty$ is a parameter. Can Pitman-Koopman Theorem be used to find a sufficient statistic for $\theta$?

**Answer:** No. We can not use Pitman-Koopman Theorem to find a sufficient statistic for $\theta$ since the domain where the population density is nonzero is not free of $\theta$.

Next, we present the connection between the maximum likelihood estimator and the sufficient estimator. If there is a sufficient estimator for the parameter $\theta$ and if the maximum likelihood estimator of this $\theta$ is unique, then the maximum likelihood estimator is a function of the sufficient estimator. That is

$$\widehat{\theta}_{\mathrm{ML}} = \psi(\widehat{\theta}_{\mathrm{S}}),$$

where $\psi$ is a real valued function, $\widehat{\theta}_{\mathrm{ML}}$ is the maximum likelihood estimator of $\theta$, and $\widehat{\theta}_{\mathrm{S}}$ is the sufficient estimator of $\theta$.

Similarly, a connection can be established between the uniform minimum variance unbiased estimator and the sufficient estimator of a parameter $\theta$. If there is a sufficient estimator for the parameter $\theta$ and if the uniform minimum variance unbiased estimator of this $\theta$ is unique, then the uniform minimum variance unbiased estimator is a function of the sufficient estimator. That is

$$\widehat{\theta}_{\text{MVUE}} = \eta(\widehat{\theta}_{\text{S}}),$$

where $\eta$ is a real valued function, $\widehat{\theta}_{\text{MVUE}}$ is the uniform minimum variance unbiased estimator of $\theta$, and $\widehat{\theta}_{\text{S}}$ is the sufficient estimator of $\theta$.

Finally, we may ask "If there are sufficient estimators, why are not there necessary estimators?" In fact, there are. Dynkin (1951) gave the following definition.

**Definition 16.7.** An estimator is said to be a necessary estimator if it can be written as a function of every sufficient estimators.

### 16.5. Consistent Estimator

Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density $f(x;\theta)$. Let $\widehat{\theta}$ be an estimator of $\theta$ based on the sample of size $n$. Obviously the estimator depends on the sample size $n$. In order to reflect the dependency of $\widehat{\theta}$ on $n$, we denote $\widehat{\theta}$ as $\widehat{\theta}_n$.

**Definition 16.7.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density $f(x;\theta)$. A sequence of estimators $\{\widehat{\theta}_n\}$ of $\theta$ is said to be consistent for $\theta$ if and only if the sequence $\{\widehat{\theta}_n\}$ converges in probability to $\theta$, that is, for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

Note that consistency is actually a concept relating to a sequence of estimators $\{\widehat{\theta}_n\}_{n=n_o}^{\infty}$ but we usually say "consistency of $\widehat{\theta}_n$" for simplicity. Further, consistency is a large sample property of an estimator.

The following theorem states that if the mean squared error goes to zero as $n$ goes to infinity, then $\{\widehat{\theta}_n\}$ converges in probability to $\theta$.

**Theorem 16.5.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density $f(x;\theta)$ and $\{\widehat{\theta}_n\}$ be a sequence of estimators of $\theta$ based on the sample. If the variance of $\widehat{\theta}_n$ exists for each $n$ and is finite and

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = 0$$

then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

**Proof:** By Markov Inequality (see Theorem 13.8) we have

$$P\left(\left(\widehat{\theta_n} - \theta\right)^2 \geq \epsilon^2\right) \leq \frac{E\left(\left(\widehat{\theta_n} - \theta\right)^2\right)}{\epsilon^2}$$

for all $\epsilon > 0$. Since the events

$$\left(\widehat{\theta_n} - \theta\right)^2 \geq \epsilon^2 \qquad \text{and} \qquad |\widehat{\theta_n} - \theta| \geq \epsilon$$

are same, we see that

$$P\left(\left(\widehat{\theta_n} - \theta\right)^2 \geq \epsilon^2\right) = P\left(|\widehat{\theta_n} - \theta| \geq \epsilon\right) \leq \frac{E\left(\left(\widehat{\theta_n} - \theta\right)^2\right)}{\epsilon^2}$$

for all $n \in \mathbb{N}$. Hence if

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta_n} - \theta\right)^2\right) = 0$$

then

$$\lim_{n\to\infty} P\left(|\widehat{\theta_n} - \theta| \geq \epsilon\right) = 0$$

and the proof of the theorem is complete.

Let

$$B\left(\widehat{\theta}, \theta\right) = E\left(\widehat{\theta}\right) - \theta$$

be the biased. If an estimator is unbiased, then $B\left(\widehat{\theta}, \theta\right) = 0$. Next we show that

$$E\left(\left(\widehat{\theta} - \theta\right)^2\right) = Var\left(\widehat{\theta}\right) + \left[B\left(\widehat{\theta}, \theta\right)\right]^2. \tag{1}$$

To see this consider

$$\begin{aligned}
E\left(\left(\widehat{\theta} - \theta\right)^2\right) &= E\left(\left(\widehat{\theta}^2 - 2\,\widehat{\theta}\,\theta + \theta^2\right)^2\right) \\
&= E\left(\widehat{\theta}^2\right) - 2E\left(\widehat{\theta}\right)\,\theta + \theta^2 \\
&= E\left(\widehat{\theta}^2\right) - E\left(\widehat{\theta}\right)^2 + E\left(\widehat{\theta}\right)^2 - 2E\left(\widehat{\theta}\right)\,\theta + \theta^2 \\
&= Var\left(\widehat{\theta}\right) + E\left(\widehat{\theta}\right)^2 - 2E\left(\widehat{\theta}\right)\,\theta + \theta^2 \\
&= Var\left(\widehat{\theta}\right) + \left[E\left(\widehat{\theta}\right) - \theta\right]^2 \\
&= Var\left(\widehat{\theta}\right) + \left[B\left(\widehat{\theta}, \theta\right)\right]^2.
\end{aligned}$$

In view of (1), we can say that if

$$\lim_{n\to\infty} Var\left(\widehat{\theta}_n\right) = 0 \tag{2}$$

and

$$\lim_{n\to\infty} B\left(\widehat{\theta}_n, \theta\right) = 0 \tag{3}$$

then

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = 0.$$

In other words, to show a sequence of estimators is consistent we have to verify the limits (2) and (3).

**Example 16.22.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population $X$ with mean $\mu$ and variance $\sigma^2 > 0$. Is the likelihood estimator

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

of $\sigma^2$ a consistent estimator of $\sigma^2$?

**Answer:** Since $\widehat{\sigma^2}$ depends on the sample size $n$, we denote $\widehat{\sigma^2}$ as $\widehat{\sigma^2}_n$. Hence

$$\widehat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

The variance of $\widehat{\sigma^2}_n$ is given by

$$\begin{aligned}
Var\left(\widehat{\sigma^2}_n\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\
&= \frac{1}{n^2} Var\left(\sigma^2 \frac{(n-1)S^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{n^2} Var\left(\frac{(n-1)S^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{n^2} Var\left(\chi^2(n-1)\right) \\
&= \frac{2(n-1)\sigma^4}{n^2} \\
&= \left[\frac{1}{n} - \frac{1}{n^2}\right] 2\,\sigma^4.
\end{aligned}$$

Hence

$$\lim_{n\to\infty} Var\left(\widehat{\theta}_n\right) = \lim_{n\to\infty}\left[\frac{1}{n} - \frac{1}{n^2}\right] 2\,\sigma^4 = 0.$$

The biased $B\left(\widehat{\theta}_n, \theta\right)$ is given by

$$\begin{aligned} B\left(\widehat{\theta}_n, \theta\right) &= E\left(\widehat{\theta}_n\right) - \sigma^2 \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) - \sigma^2 \\ &= \frac{1}{n}\,E\left(\sigma^2\,\frac{(n-1)S^2}{\sigma^2}\right) - \sigma^2 \\ &= \frac{\sigma^2}{n}E\left(\chi^2(n-1)\right) - \sigma^2 \\ &= \frac{(n-1)\sigma^2}{n} - \sigma^2 \\ &= -\frac{\sigma^2}{n}. \end{aligned}$$

Thus

$$\lim_{n\to\infty} B\left(\widehat{\theta}_n, \theta\right) = -\lim_{n\to\infty}\frac{\sigma^2}{n} = 0.$$

Hence $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is a consistent estimator of $\sigma^2$.

In the last example we saw that the likelihood estimator of variance is a consistent estimator. In general, if the density function $f(x;\theta)$ of a population satisfies some mild conditions, then the maximum likelihood estimator of $\theta$ is consistent. Similarly, if the density function $f(x;\theta)$ of a population satisfies some mild conditions, then the estimator obtained by moment method is also consistent.

Since consistency is a large sample property of an estimator, some statisticians suggest that consistency should not be used alone for judging the goodness of an estimator; rather it should be used along with other criteria.

### 16.6. Review Exercises

**1.** Let $T_1$ and $T_2$ be estimators of a population parameter $\theta$ based upon the same random sample. If $T_i \sim N\left(\theta, \sigma_i^2\right)$ $i = 1, 2$ and if $T = bT_1 + (1-b)T_2$, then for what value of $b$, $T$ is a minimum variance unbiased estimator of $\theta$ ?

**2.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \frac{1}{2\theta}\, e^{-\frac{|x|}{\theta}} \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. What is the expected value of the maximum likelihood estimator of $\theta$ ? Is this estimator unbiased?

**3.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \frac{1}{2\theta}\, e^{-\frac{|x|}{\theta}} \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. Is the maximum likelihood estimator an efficient estimator of $\theta$?

**4.** A random sample $X_1, X_2, ..., X_n$ of size $n$ is selected from a normal distribution with variance $\sigma^2$. Let $S^2$ be the unbiased estimator of $\sigma^2$, and $T$ be the maximum likelihood estimator of $\sigma^2$. If $20T - 19S^2 = 0$, then what is the sample size?

**5.** Suppose $X$ and $Y$ are independent random variables each with density function

$$f(x) = \begin{cases} 2\,x\,\theta^2 & \text{for } 0 < x < \frac{1}{\theta} \\ \\ 0 & \text{otherwise.} \end{cases}$$

If $k\,(X + 2Y)$ is an unbiased estimator of $\theta^{-1}$, then what is the value of $k$?

**6.** An object of length $c$ is measured by two persons using the same instrument. The instrument error has a normal distribution with mean 0 and variance 1. The first person measures the object 25 times, and the average of the measurements is $\bar{X} = 12$. The second person measures the objects 36 times, and the average of the measurements is $\bar{Y} = 12.8$. To estimate $c$ we use the weighted average $a\,\bar{X} + b\,\bar{Y}$ as an estimator. Determine the constants $a$ and $b$ such that $a\,\bar{X} + b\,\bar{Y}$ is the minimum variance unbiased estimator of $c$ and then calculate the minimum variance unbiased estimate of $c$.

**7.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density function

$$f(x) = \begin{cases} 3\,\theta\,x^2\,e^{-\theta\,x^3} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Find a sufficient statistics for $\theta$.

**8.** Let $X_1, X_2, ..., X_n$ be a random sample from a Weibull distribution with probability density function

$$f(x) = \begin{cases} \frac{\beta}{\theta^\beta}\, x^{\beta-1}\, e^{-(\frac{x}{\theta})^\beta} & \text{if } x > 0 \\ \\ 0 & \text{otherwise }, \end{cases}$$

where $\theta > 0$ and $\beta > 0$ are parameters. Find a sufficient statistics for $\theta$ with $\beta$ known, say $\beta = 2$. If $\beta$ is unknown, can you find a single sufficient statistics for $\theta$?

**9.** Let $X_1, X_2$ be a random sample of size 2 from population with probability density

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

where $\theta > 0$ is an unknown parameter. If $Y = \sqrt{X_1 X_2}$, then what should be the value of the constant $k$ such that $kY$ is an unbiased estimator of the parameter $\theta$ ?

**10.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise }, \end{cases}$$

where $\theta > 0$ is an unknown parameter. If $\overline{X}$ denotes the sample mean, then what should be value of the constant $k$ such that $k\overline{X}$ is an unbiased estimator of $\theta$ ?

**11.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise }, \end{cases}$$

where $\theta > 0$ is an unknown parameter. If $X_{med}$ denotes the sample median, then what should be value of the constant $k$ such that $kX_{med}$ is an unbiased estimator of $\theta$ ?

**12.** What do you understand by an unbiased estimator of a parameter $\theta$? What is the basic principle of the maximum likelihood estimation of a parameter $\theta$? What is the basic principle of the Bayesian estimation of a parameter $\theta$? What is the main difference between Bayesian method and likelihood method.

**13.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \frac{\theta}{(1+x)^{\theta+1}} & \text{for } 0 \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. What is a sufficient statistic for the parameter $\theta$?

**14.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \frac{x}{\theta^2}\, e^{-\frac{x^2}{2\theta^2}} & \text{for } 0 \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is an unknown parameter. What is a sufficient statistic for the parameter $\theta$?

**15.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < \theta < \infty$ is a parameter. What is the maximum likelihood estimator of $\theta$? Find a sufficient statistics of the parameter $\theta$.

**16.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < \theta < \infty$ is a parameter. Are the estimators $X_{(1)}$ and $\overline{X} - 1$ are unbiased estimators of $\theta$? Which one is more efficient than the other?

**17.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{for } 0 \leq x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 1$ is an unknown parameter. What is a sufficient statistic for the parameter $\theta$?

**18.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

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

where $\theta > 0$ and $\alpha > 0$ are parameters. What is a sufficient statistic for the parameter $\theta$ for a fixed $\alpha$?

**19.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \frac{\theta\,\alpha^{\theta}}{x^{(\theta+1)}} & \text{for } \alpha < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ and $\alpha > 0$ are parameters. What is a sufficient statistic for the parameter $\theta$ for a fixed $\alpha$?

**20.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \binom{m}{x}\theta^{x}(1-\theta)^{m-x} & \text{for } x = 0, 1, 2, ..., m \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is parameter. Show that $\frac{\overline{X}}{m}$ is a uniform minimum variance unbiased estimator of $\theta$ for a fixed $m$.

**21.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density function

$$f(x;\theta) = \begin{cases} \theta\,x^{\theta-1} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 1$ is parameter. Show that $-\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ is a uniform minimum variance unbiased estimator of $\frac{1}{\theta}$.

**22.** Let $X_1, X_2, ..., X_n$ be a random sample from a uniform population $X$ on the interval $[0, \theta]$, where $\theta > 0$ is a parameter. Is the likelihood estimator $\widehat{\theta} = X_{(n)}$ of $\theta$ a consistent estimator of $\theta$?

**23.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X \sim POI(\lambda)$, where $\lambda > 0$ is a parameter. Is the estimator $\overline{X}$ of $\lambda$ a consistent estimator of $\lambda$?

**24.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ having the probability density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1}, & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Is the estimator $\widehat{\theta} = \frac{\overline{X}}{1-\overline{X}}$ of $\theta$, obtained by the moment method, a consistent estimator of $\theta$?

**25.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ having the probability density function

$$f(x;p) = \begin{cases} \binom{n}{x} p^x\, (1-p)^{n-x}, & \text{if } x = 0, 1, 2, ..., n \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p < 1$ is a parameter and $m$ is a fixed positive integer. What is the maximum likelihood estimator for $p$. Is this maximum likelihood estimator for $p$ is an efficient estimator?

**26.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ having the probability density function

$$f(x;\theta) = \begin{cases} \frac{2}{\theta^2}\, \theta - x, & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Find an estimator for $\theta$ using the moment method.

**27.** A box contains 50 red and blue balls out of which $\theta$ are red. A sample of 30 balls is to be selected without replacement. If $X$ denotes the number of red balls in the sample, then find an estimator for $\theta$ using the moment method.

# Chapter 17

# SOME TECHNIQUES FOR FINDING INTERVAL ESTIMATORS FOR PARAMETERS

In point estimation we find a value for the parameter $\theta$ given a sample data. For example, if $X_1, X_2, ..., X_n$ is a random sample of size $n$ from a population with probability density function

$$f(x;\theta) = \begin{cases} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x-\theta)^2} & \text{for } x \geq \theta \\ 0 & \text{otherwise,} \end{cases}$$

then the likelihood function of $\theta$ is

$$L(\theta) = \prod_{i=1}^{n} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x_i-\theta)^2},$$

where $x_1 \geq \theta$, $x_2 \geq \theta, ..., x_n \geq \theta$. This likelihood function simplifies to

$$L(\theta) = \left[\frac{2}{\pi}\right]^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2},$$

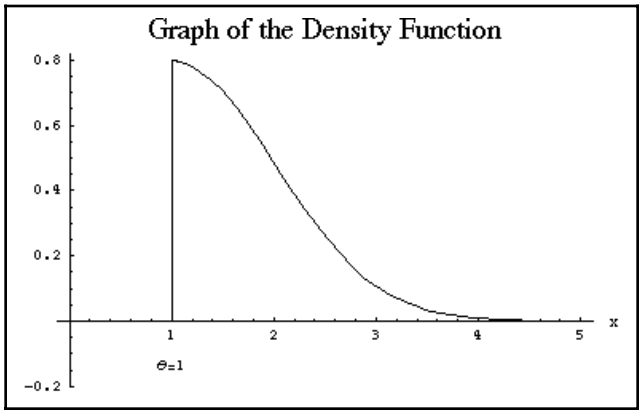
where $\min\{x_1, x_2, ..., x_n\} \geq \theta$. Taking the natural logarithm of $L(\theta)$ and maximizing, we obtain the maximum likelihood estimator of $\theta$ as the first order statistic of the sample $X_1, X_2, ..., X_n$, that is

$$\widehat{\theta} = X_{(1)},$$

where $X_{(1)} = \min\{X_1, X_2, ..., X_n\}$. Suppose the true value of $\theta = 1$. Using the maximum likelihood estimator of $\theta$, we are trying to guess this value of $\theta$ based on a random sample. Suppose $X_1 = 1.5, X_2 = 1.1, X_3 = 1.7, X_4 = 2.1, X_5 = 3.1$ is a set of sample data from the above population. Then based on this random sample, we will get

$$\widehat{\theta}_{\text{ML}} = X_{(1)} = \min\{1.5, 1.1, 1.7, 2.1, 3.1\} = 1.1.$$

If we take another random sample, say $X_1 = 1.8, X_2 = 2.1, X_3 = 2.5, X_4 = 3.1, X_5 = 2.6$ then the maximum likelihood estimator of this $\theta$ will be $\widehat{\theta} = 1.8$ based on this sample. The graph of the density function $f(x;\theta)$ for $\theta = 1$ is shown below.




From the graph, it is clear that a number close to 1 has higher chance of getting randomly picked by the sampling process, then the numbers that are substantially bigger than 1. Hence, it makes sense that $\theta$ should be estimated by the smallest sample value. However, from this example we see that the point estimate of $\theta$ is not equal to the true value of $\theta$. Even if we take many random samples, yet the estimate of $\theta$ will rarely equal the actual value of the parameter. Hence, instead of finding a single value for $\theta$, we should report a range of probable values for the parameter $\theta$ with certain degree of confidence. This brings us to the notion of confidence interval of a parameter.

### 17.1. Interval Estimators and Confidence Intervals for Parameters

The interval estimation problem can be stated as follow: Given a random sample $X_1, X_2, ..., X_n$ and a probability value $1 - \alpha$, find a pair of statistics $L = L(X_1, X_2, ..., X_n)$ and $U = U(X_1, X_2, ..., X_n)$ with $L \leq U$ such that the

probability of $\theta$ being on the random interval $[L, U]$ is $1-\alpha$. That is

$$P\left(L \leq \theta \leq U\right) = 1-\alpha.$$

Recall that a sample is a portion of the population usually chosen by method of random sampling and as such it is a set of random variables $X_1, X_2, ..., X_n$ with the same probability density function $f(x;\theta)$ as the population. Once the sampling is done, we get

$$X_1 = x_1,\ X_2 = x_2,\ \cdots,\ X_n = x_n$$

where $x_1, x_2, ..., x_n$ are the *sample data*.

**Definition 17.1.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population $X$ with density $f(x;\theta)$, where $\theta$ is an unknown parameter. The *interval estimator* of $\theta$ is a pair of statistics $L = L(X_1, X_2, ..., X_n)$ and $U = U(X_1, X_2, ..., X_n)$ with $L \leq U$ such that if $x_1, x_2, ..., x_n$ is a set of sample data, then $\theta$ belongs to the interval $[L(x_1, x_2, ...x_n),\ U(x_1, x_2, ...x_n)]$.

The interval $[l, u]$ will be denoted as an interval estimate of $\theta$ whereas the random interval $[L, U]$ will denote the interval estimator of $\theta$. Notice that the interval estimator of $\theta$ is the random interval $[L,\ U]$. Next, we define the $100(1-\alpha)\%$ confidence interval for the unknown parameter $\theta$.

**Definition 17.2.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population $X$ with density $f(x;\theta)$, where $\theta$ is an unknown parameter. The interval estimator of $\theta$ is called a $100(1-\alpha)\%$ *confidence interval* for $\theta$ if

$$P(L \leq \theta \leq U) = 1-\alpha.$$

The random variable $L$ is called the *lower confidence limit* and $U$ is called the *upper confidence limit*. The number $(1-\alpha)$ is called the *confidence coefficient* or *degree of confidence*.

There are several methods for constructing confidence intervals for an unknown parameter $\theta$. Some well known methods are: (1) Pivotal Quantity Method, (2) Maximum Likelihood Estimator (MLE) Method, (3) Bayesian Method, (4) Invariant Methods, (5) Inversion of Test Statistic Method, and (6) The Statistical or General Method.

In this chapter, we only focus on the pivotal quantity method and the MLE method. We also briefly examine the the statistical or general method. The pivotal quantity method is mainly due to George Bernard and David Fraser of the University of Waterloo, and this method is perhaps one of the most elegant methods of constructing confidence intervals for unknown parameters.

## 17.2. Pivotal Quantity Method

In this section, we explain how the notion of pivotal quantity can be used to construct confidence interval for a unknown parameter. We will also examine how to find pivotal quantities for parameters associated with certain probability density functions. We begin with the formal definition of the pivotal quantity.

**Definition 17.3.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a population $X$ with probability density function $f(x;\theta)$, where $\theta$ is an unknown parameter. A *pivotal quantity* $Q$ is a function of $X_1, X_2, ..., X_n$ and $\theta$ whose probability distribution is independent of the parameter $\theta$.

Notice that the pivotal quantity $Q(X_1, X_2, ..., X_n, \theta)$ will usually contain both the parameter $\theta$ and an estimator (that is, a statistic) of $\theta$. Now we give an example of a pivotal quantity.

**Example 17.1.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population $X$ with mean $\mu$ and a known variance $\sigma^2$. Find a pivotal quantity for the unknown parameter $\mu$.

**Answer:** Since each $X_i \sim N(\mu, \sigma^2)$,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Standardizing $\overline{X}$, we see that

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

The statistics $Q$ given by

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is a pivotal quantity since it is a function of $X_1, X_2, ..., X_n$ and $\mu$ and its probability density function is free of the parameter $\mu$.

There is no general rule for finding a pivotal quantity (or pivot) for a parameter $\theta$ of an arbitrarily given density function $f(x;\theta)$. Hence to some extents, finding pivots relies on guesswork. However, if the probability density function $f(x;\theta)$ belongs to the location-scale family, then there is a systematic way to find pivots.

**Definition 17.4.** Let $g : \mathbb{R} \to \mathbb{R}$ be a probability density function. Then for any $\mu$ and any $\sigma > 0$, the family of functions

$$\mathcal{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) \mid \mu \in (-\infty, \infty),\ \sigma \in (0, \infty) \right\}$$

is called the *location-scale family* with standard probability density $f(x; \theta)$. The parameter $\mu$ is called the *location parameter* and the parameter $\sigma$ is called the *scale parameter*. If $\sigma = 1$, then $\mathcal{F}$ is called the *location family*. If $\mu = 0$, then $\mathcal{F}$ is called the *scale family*

It should be noted that each member $f(x; \mu, \sigma)$ of the location-scale family is a probability density function. If we take $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, then the normal density function

$$f(x; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

belongs to the location-scale family. The density function

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

belongs to the scale family. However, the density function

$$f(x; \theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

does not belong to the location-scale family.

It is relatively easy to find pivotal quantities for location or scale parameter when the density function of the population belongs to the location-scale family $\mathcal{F}$. When the density function belongs to location family, the pivot for the location parameter $\mu$ is $\widehat{\mu} - \mu$, where $\widehat{\mu}$ is the maximum likelihood estimator of $\mu$. If $\widehat{\sigma}$ is the maximum likelihood estimator of $\sigma$, then the pivot for the scale parameter $\sigma$ is $\frac{\widehat{\sigma}}{\sigma}$ when the density function belongs to the scale family. The pivot for location parameter $\mu$ is $\frac{\widehat{\mu}-\mu}{\widehat{\sigma}}$ and the pivot for the scale parameter $\sigma$ is $\frac{\widehat{\sigma}}{\sigma}$ when the density function belongs to location-scale family. Sometime it is appropriate to make a minor modification to the pivot obtained in this way, such as multiplying by a constant, so that the modified pivot will have a known distribution.

**Remark 17.1.** Pivotal quantity can also be constructed using a sufficient statistic for the parameter. Suppose $T = T(X_1, X_2, ..., X_n)$ is a sufficient statistic based on a random sample $X_1, X_2, ..., X_n$ from a population $X$ with probability density function $f(x;\theta)$. Let the probability density function of $T$ be $g(t;\theta)$. If $g(t;\theta)$ belongs to the location family, then an appropriate constant multiple of $T - a(\theta)$ is a pivotal quantity for the location parameter $\theta$ for some suitable expression $a(\theta)$. If $g(t;\theta)$ belongs to the scale family, then an appropriate constant multiple of $\frac{T}{b(\theta)}$ is a pivotal quantity for the scale parameter $\theta$ for some suitable expression $b(\theta)$. Similarly, if $g(t;\theta)$ belongs to the location-scale family, then an appropriate constant multiple of $\frac{T-a(\theta)}{b(\theta)}$ is a pivotal quantity for the location parameter $\theta$ for some suitable expressions $a(\theta)$ and $b(\theta)$.

Algebraic manipulations of pivots are key factors in finding confidence intervals. If $Q = Q(X_1, X_2, ..., X_n, \theta)$ is a pivot, then a $100(1-\alpha)\%$ confidence interval for $\theta$ may be constructed as follows: First, find two values $a$ and $b$ such that

$$P(a \leq Q \leq b) = 1 - \alpha,$$

then convert the inequality $a \leq Q \leq b$ into the form $L \leq \theta \leq U$.

For example, if $X$ is normal population with unknown mean $\mu$ and known variance $\sigma^2$, then its pdf belongs to the location-scale family. A pivot for $\mu$ is $\frac{\overline{X}-\mu}{S}$. However, since the variance $\sigma^2$ is known, there is no need to take $S$. So we consider the pivot $\frac{\overline{X}-\mu}{\sigma}$ to construct the $100(1-2\alpha)\%$ confidence interval for $\mu$. Since our population $X \sim N(\mu, \sigma^2)$, the sample mean $\overline{X}$ is also a normal with the same mean $\mu$ and the variance equals to $\frac{\sigma}{\sqrt{n}}$. Hence

$$\begin{aligned}
1 - 2\alpha &= P\left(-z_\alpha \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha\right) \\
&= P\left(\mu - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

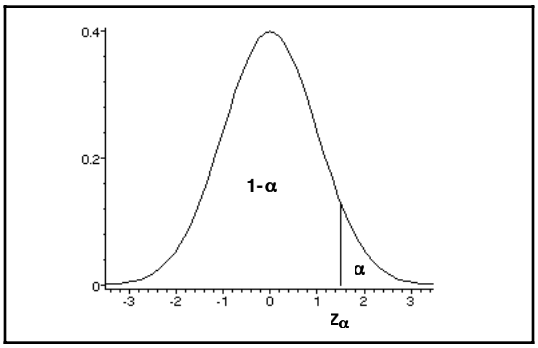
Therefore, the $100(1-2\alpha)\%$ confidence interval for $\mu$ is

$$\left[\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

Here $z_\alpha$ denotes the $100(1-\alpha)$-percentile (or $(1-\alpha)$-quartile) of a standard normal random variable $Z$, that is

$$P(Z \leq z_\alpha) = 1 - \alpha,$$

where $\alpha \leq 0.5$ (see figure below). Note that $\alpha = P(Z \leq -z_\alpha)$ if $\alpha \leq 0.5$.




A $100(1-\alpha)\%$ confidence interval for a parameter $\theta$ has the following interpretation. If $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ is a sample of size $n$, then based on this sample we construct a $100(1-\alpha)\%$ confidence interval $[l, u]$ which is a subinterval of the real line $\mathbb{R}$. Suppose we take large number of samples from the underlying population and construct all the corresponding $100(1-\alpha)\%$ confidence intervals, then approximately $100(1-\alpha)\%$ of these intervals would include the unknown value of the parameter $\theta$.

In the next several sections, we illustrate how pivotal quantity method can be used to determine confidence intervals for various parameters.

### 17.3. Confidence Interval for Population Mean

At the outset, we use the pivotal quantity method to construct a confidence interval for the mean of a normal population. Here we assume first the population variance is known and then variance is unknown. Next, we construct the confidence interval for the mean of a population with continuous, symmetric and unimodal probability distribution by applying the central limit theorem.

Let $X_1, X_2, ..., X_n$ be a random sample from a population $X \sim N(\mu, \sigma^2)$, where $\mu$ is an unknown parameter and $\sigma^2$ is a known parameter. First of all, we need a pivotal quantity $Q(X_1, X_2, ..., X_n, \mu)$. To construct this pivotal

quantity, we find the likelihood estimator of the parameter $\mu$. We know that $\widehat{\mu} = \overline{X}$. Since, each $X_i \sim N(\mu, \sigma^2)$, the distribution of the sample mean is given by

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

It is easy to see that the distribution of the estimator of $\mu$ is not independent of the parameter $\mu$. If we standardize $\overline{X}$, then we get

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

The distribution of the standardized $\overline{X}$ is independent of the parameter $\mu$. This standardized $\overline{X}$ is the pivotal quantity since it is a function of the sample $X_1, X_2, ..., X_n$ and the parameter $\mu$, and its probability distribution is independent of the parameter $\mu$. Using this pivotal quantity, we construct the confidence interval as follows:

$$\begin{aligned} 1 - \alpha &= P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) \\ &= P\left(\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right) \end{aligned}$$

Hence, the $(1 - \alpha)\%$ confidence interval for $\mu$ when the population $X$ is normal with the known variance $\sigma^2$ is given by

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}, \ \ \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

This says that if samples of size $n$ are taken from a normal population with mean $\mu$ and known variance $\sigma^2$ and if the interval

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}, \ \ \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right]$$

is constructed for every sample, then in the long-run $100(1 - \alpha)\%$ of the intervals will cover the unknown parameter $\mu$ and hence with a confidence of $(1 - \alpha)100\%$ we can say that $\mu$ lies on the interval

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}, \ \ \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

The interval estimate of $\mu$ is found by taking a good (here maximum likelihood) estimator $\overline{X}$ of $\mu$ and adding and subtracting $z_{\frac{\alpha}{2}}$ times the standard deviation of $\overline{X}$.

**Remark 17.2.** By definition a $100(1-\alpha)\%$ confidence interval for a parameter $\theta$ is an interval $[L, U]$ such that the probability of $\theta$ being in the interval $[L, U]$ is $1-\alpha$. That is

$$1-\alpha = P(L \leq \theta \leq U).$$

One can find infinitely many pairs $L, U$ such that

$$1-\alpha = P(L \leq \theta \leq U).$$

Hence, there are infinitely many confidence intervals for a given parameter. However, we only consider the confidence interval of shortest length. If a confidence interval is constructed by omitting equal tail areas then we obtain what is known as the central confidence interval. In a symmetric distribution, it can be shown that the central confidence interval is of the shortest length.

**Example 17.2.** Let $X_1, X_2, ..., X_{11}$ be a random sample of size 11 from a normal distribution with unknown mean $\mu$ and variance $\sigma^2 = 9.9$. If $\sum_{i=1}^{11} x_i = 132$, then what is the 95% confidence interval for $\mu$ ?

**Answer:** Since each $X_i \sim N(\mu, 9.9)$, the confidence interval for $\mu$ is given by

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}, \;\; \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

Since $\sum_{i=1}^{11} x_i = 132$, the sample mean $\overline{x} = \frac{132}{11} = 12$. Also, we see that

$$\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{9.9}{11}} = \sqrt{0.9}.$$

Further, since $1-\alpha = 0.95$, $\alpha = 0.05$. Thus

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96 \qquad \text{(from normal table)}.$$

Using these information in the expression of the confidence interval for $\mu$, we get

$$\left[12 - 1.96\sqrt{0.9}, \;\; 12 + 1.96\sqrt{0.9}\right]$$

that is

$$[10.141, \;\; 13.859].$$

**Example 17.3.** Let $X_1, X_2, ..., X_{11}$ be a random sample of size 11 from a normal distribution with unknown mean $\mu$ and variance $\sigma^2 = 9.9$. If $\sum_{i=1}^{11} x_i = 132$, then for what value of the constant $k$ is

$$\left[12 - k\sqrt{0.9},\ 12 + k\sqrt{0.9}\right]$$

a 90% confidence interval for $\mu$ ?

**Answer:** The 90% confidence interval for $\mu$ when the variance is given is

$$\left[\overline{x} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\ ,\ \ \overline{x} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

Thus we need to find $\overline{x}$, $\sqrt{\frac{\sigma^2}{n}}$ and $z_{\frac{\alpha}{2}}$ corresponding to $1 - \alpha = 0.9$. Hence

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{11} x_i}{11} \\
&= \frac{132}{11} \\
&= 12. \\
\sqrt{\frac{\sigma^2}{n}} &= \sqrt{\frac{9.9}{11}} \\
&= \sqrt{0.9}. \\
z_{0.05} &= 1.64 \qquad \text{(from normal table).}
\end{aligned}$$

Hence, the confidence interval for $\mu$ at 90% confidence level is

$$\left[12 - (1.64)\sqrt{0.9},\ 12 + (1.64)\ \sqrt{0.9}\right].$$

Comparing this interval with the given interval, we get

$$k = 1.64.$$

and the corresponding 90% confidence interval is [10.444, 13.556].

**Remark 17.3.** Notice that the length of the 90% confidence interval for $\mu$ is 3.112. However, the length of the 95% confidence interval is 3.718. Thus higher the confidence level bigger is the length of the confidence interval. Hence, the confidence level is directly proportional to the length of the confidence interval. In view of this fact, we see that if the confidence level is zero,

then the length is also zero. That is when the confidence level is zero, the confidence interval of $\mu$ degenerates into a point $\overline{X}$.

Until now we have considered the case when the population is normal with unknown mean $\mu$ and known variance $\sigma^2$. Now we consider the case when the population is non-normal but its probability density function is continuous, symmetric and unimodal. If the sample size is large, then by the central limit theorem

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \qquad \text{as} \quad n \to \infty.$$

Thus, in this case we can take the pivotal quantity to be

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}},$$

if the sample size is large (generally $n \geq 32$). Since the pivotal quantity is same as before, we get the sample expression for the $(1-\alpha)100\%$ confidence interval, that is

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}, \;\; \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

**Example 17.4.** Let $X_1, X_2, ..., X_{40}$ be a random sample of size 40 from a distribution with known variance and unknown mean $\mu$. If $\sum_{i=1}^{40} x_i = 286.56$ and $\sigma^2 = 10$, then what is the 90 percent confidence interval for the population mean $\mu$ ?

**Answer:** Since $1-\alpha = 0.90$, we get $\frac{\alpha}{2} = 0.05$. Hence, $z_{0.05} = 1.64$ (from the standard normal table). Next, we find the sample mean

$$\overline{x} = \frac{286.56}{40} = 7.164.$$

Hence, the confidence interval for $\mu$ is given by

$$\left[7.164 - (1.64)\left(\sqrt{\frac{10}{40}}\right), \;\; 7.164 + (1.64)\left(\sqrt{\frac{10}{40}}\right)\right]$$

that is

$$[6.344, \;\; 7.984].$$

**Example 17.5.** In sampling from a nonnormal distribution with a variance of 25, how large must the sample size be so that the length of a 95% confidence interval for the mean is 1.96 ?

**Answer:** The confidence interval when the sample is taken from a normal population with a variance of 25 is

$$\left[\overline{x} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\,,\ \ \overline{x} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

Thus the length of the confidence interval is

$$\begin{aligned}\ell &= 2\, z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \\ &= 2\, z_{0.025} \sqrt{\frac{25}{n}} \\ &= 2\,(1.96) \sqrt{\frac{25}{n}}.\end{aligned}$$

But we are given that the length of the confidence interval is $\ell = 1.96$. Thus

$$\begin{aligned}1.96 &= 2\,(1.96) \sqrt{\frac{25}{n}} \\ \sqrt{n} &= 10 \\ n &= 100.\end{aligned}$$

Hence, the sample size must be 100 so that the length of the 95% confidence interval will be 1.96.

So far, we have discussed the method of construction of confidence interval for the parameter population mean when the variance is known. It is very unlikely that one will know the variance without knowing the population mean, and thus what we have treated so far in this section is not very realistic. Now we treat case of constructing the confidence interval for population mean when the population variance is also unknown. First of all, we begin with the construction of confidence interval assuming the population $X$ is normal.

Suppose $X_1, X_2, ..., X_n$ is random sample from a normal population $X$ with mean $\mu$ and variance $\sigma^2 > 0$. Let the sample mean and sample variances be $\overline{X}$ and $S^2$ respectively. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

and

$$\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1).$$

Therefore, the random variable defined by the ratio of $\frac{(n-1)S^2}{\sigma^2}$ to $\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ has a $t$-distribution with $(n-1)$ degrees of freedom, that is

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1),$$

where $Q$ is the pivotal quantity to be used for the construction of the confidence interval for $\mu$. Using this pivotal quantity, we construct the confidence interval as follows:

$$\begin{aligned} 1-\alpha &= P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}(n-1)\right) \\ &= P\left(\overline{X} - \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1) \leq \mu \leq \overline{X} + \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\right) \end{aligned}$$

Hence, the $100(1-\alpha)\%$ confidence interval for $\mu$ when the population $X$ is normal with the unknown variance $\sigma^2$ is given by

$$\left[\overline{X} - \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\,, \quad \overline{X} + \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\right].$$

**Example 17.6.** A random sample of 9 observations from a normal population yields the observed statistics $\overline{x} = 5$ and $\frac{1}{8}\sum_{i=1}^{9}(x_i - \overline{x})^2 = 36$. What is the 95% confidence interval for $\mu$ ?

**Answer:** Since

$$\begin{aligned} n &= 9 \qquad\qquad \overline{x} = 5 \\ s^2 &= 36 \qquad \text{and} \qquad 1-\alpha = 0.95, \end{aligned}$$

the 95% confidence interval for $\mu$ is given by

$$\left[\overline{x} - \left(\frac{s}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\,, \quad \overline{x} + \left(\frac{s}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\right],$$

that is

$$\left[5 - \left(\frac{6}{\sqrt{9}}\right) t_{0.025}(8)\,, \quad 5 + \left(\frac{6}{\sqrt{9}}\right) t_{0.025}(8)\right],$$

which is

$$\left[5 - \left(\frac{6}{\sqrt{9}}\right)(2.306), \quad 5 + \left(\frac{6}{\sqrt{9}}\right)(2.306)\right].$$

Hence, the 95% confidence interval for $\mu$ is given by $[0.388,\ 9.612]$.

**Example 17.7.** Which of the following is true of a 95% confidence interval for the mean of a population?

(a) The interval includes 95% of the population values on the average.
(b) The interval includes 95% of the sample values on the average.
(c) The interval has 95% chance of including the sample mean.

**Answer:** None of the statements is correct since the 95% confidence interval for the population mean $\mu$ means that the interval has 95% chance of including the population mean $\mu$.

Finally, we consider the case when the population is non-normal but it probability density function is continuous, symmetric and unimodal. If some weak conditions are satisfied, then the sample variance $S^2$ of a random sample of size $n \geq 2$, converges stochastically to $\sigma^2$. Therefore, in

$$\frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}$$

the numerator of the left-hand member converges to $N(0,1)$ and the denominator of that member converges to 1. Hence

$$\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1) \qquad \text{as} \quad n \to \infty.$$

This fact can be used for the construction of a confidence interval for population mean when variance is unknown and the population distribution is nonnormal. We let the pivotal quantity to be

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}$$

and obtain the following confidence interval

$$\left[\overline{X} - \left(\frac{S}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}, \quad \overline{X} + \left(\frac{S}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

We summarize the results of this section by the following table.

| Population | Variance $\sigma^2$ | Sample Size $n$ | Confidence Limits |
|---|---|---|---|
| normal | known | $n \geq 2$ | $\overline{x} \mp z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ |
| normal | not known | $n \geq 2$ | $\overline{x} \mp t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}$ |
| not normal | known | $n \geq 32$ | $\overline{x} \mp z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ |
| not normal | known | $n < 32$ | no formula exists |
| not normal | not known | $n \geq 32$ | $\overline{x} \mp t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}$ |
| not normal | not known | $n < 32$ | no formula exists |

### 17.4. Confidence Interval for Population Variance

In this section, we will first describe the method for constructing the confidence interval for variance when the population is normal with a known population mean $\mu$. Then we treat the case when the population mean is also unknown.

Let $X_1, X_2, ..., X_n$ be a random sample from a normal population $X$ with known mean $\mu$ and unknown variance $\sigma^2$. We would like to construct a $100(1-\alpha)\%$ confidence interval for the variance $\sigma^2$, that is, we would like to find the estimate of $L$ and $U$ such that

$$P\left(L \leq \sigma^2 \leq U\right) = 1 - \alpha.$$

To find these estimate of $L$ and $U$, we first construct a pivotal quantity. Thus

$$\begin{aligned}
&X_i \sim N\left(\mu, \sigma^2\right), \\
&\left(\frac{X_i - \mu}{\sigma}\right) \sim N(0,1), \\
&\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1). \\
&\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).
\end{aligned}$$

We define the pivotal quantity $Q(X_1, X_2, ..., X_n, \sigma^2)$ as

$$Q(X_1, X_2, ..., X_n, \sigma^2) = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

which has a chi-square distribution with $n$ degrees of freedom. Hence

$$\begin{aligned}
1-\alpha &= P\left(a \le Q \le b\right) \\
&= P\left(a \le \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \le b\right) \\
&= P\left(\frac{1}{a} \ge \sum_{i=1}^{n}\frac{\sigma^2}{(X_i-\mu)^2} \ge \frac{1}{b}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{a} \ge \sigma^2 \ge \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{b}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{b} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{a}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right)
\end{aligned}$$

Therefore, the $(1-\alpha)\%$ confidence interval for $\sigma^2$ when mean is known is given by

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)},\ \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right].$$

**Example 17.8.** A random sample of 9 observations from a normal population with $\mu = 5$ yields the observed statistics $\frac{1}{8}\sum_{i=1}^{9} x_i^2 = 39.125$ and $\sum_{i=1}^{9} x_i = 45$. What is the 95% confidence interval for $\sigma^2$ ?

**Answer:** We have been given that

$$n = 9 \qquad \text{and} \qquad \mu = 5.$$

Further we know that

$$\sum_{i=1}^{9} x_i = 45 \qquad \text{and} \qquad \frac{1}{8}\sum_{i=1}^{9} x_i^2 = 39.125.$$

Hence

$$\sum_{i=1}^{9} x_i^2 = 313,$$

and

$$\begin{aligned}
\sum_{i=1}^{9}(x_i-\mu)^2 &= \sum_{i=1}^{9} x_i^2 - 2\mu\sum_{i=1}^{9} x_i + 9\mu^2 \\
&= 313 - 450 + 225 \\
&= 88.
\end{aligned}$$

Since $1-\alpha = 0.95$, we get $\frac{\alpha}{2} = 0.025$ and $1-\frac{\alpha}{2} = 0.975$. Using chi-square table we have

$$\chi^2_{0.025}(9) = 2.700 \qquad \text{and} \qquad \chi^2_{0.975}(9) = 19.02.$$

Hence, the 95% confidence interval for $\sigma^2$ is given by

$$\left[\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}, \ \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right],$$

that is

$$\left[\frac{88}{19.02}, \ \frac{88}{2.7}\right]$$

which is

$$[4.63, \ 32.59].$$

**Remark 17.4.** Since the $\chi^2$ distribution is not symmetric, the above confidence interval is not necessarily the shortest. Later, in the next section, we describe how one construct a confidence interval of shortest length.

Consider a random sample $X_1, X_2, ..., X_n$ from a normal population $X \sim N(\mu, \sigma^2)$, where the population mean $\mu$ and population variance $\sigma^2$ are unknown. We want to construct a $100(1-\alpha)\%$ confidence interval for the population variance. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\Rightarrow \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2(n-1).$$

We take $\frac{\sum_{i=1}^n \left(X_i-\overline{X}\right)^2}{\sigma^2}$ as the pivotal quantity $Q$ to construct the confidence interval for $\sigma^2$. Hence, we have

$$\begin{aligned}
1-\alpha &= P\left(\frac{1}{\chi^2_{\frac{\alpha}{2}}(n-1)} \le Q \le \frac{1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) \\
&= P\left(\frac{1}{\chi^2_{\frac{\alpha}{2}}(n-1)} \le \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \le \frac{1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) \\
&= P\left(\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right).
\end{aligned}$$

Hence, the $100(1-\alpha)\%$ confidence interval for variance $\sigma^2$ when the population mean is unknown is given by

$$\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)},\ \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

**Example 17.9.** Let $X_1, X_2, ..., X_n$ be a random sample of size 13 from a normal distribution $N(\mu, \sigma^2)$. If $\sum_{i=1}^{13} x_i = 246.61$ and $\sum_{i=1}^{13} x_i^2 = 4806.61$. Find the 90% confidence interval for $\sigma^2$ ?

**Answer:**

$$\begin{aligned}
\overline{x} &= 18.97\\
s^2 &= \frac{1}{n-1}\sum_{i=1}^{13}(x_i-\overline{x})^2\\
&= \frac{1}{n-1}\sum_{i=1}^{13}\left[x_i^2-n\overline{x}^2\right]^2\\
&= \frac{1}{12}\left[4806.61-4678.2\right]\\
&= \frac{1}{12}\ 128.41.
\end{aligned}$$

Hence, $12s^2 = 128.41$. Further, since $1-\alpha = 0.90$, we get $\frac{\alpha}{2} = 0.05$ and $1-\frac{\alpha}{2} = 0.95$. Therefore, from chi-square table, we get

$$\chi^2_{0.95}(12) = 21.03, \qquad \chi^2_{0.05}(12) = 5.23.$$

Hence, the 95% confidence interval for $\sigma^2$ is

$$\left[\frac{128.41}{21.03},\ \frac{128.41}{5.23}\right],$$

that is

$$[6.11,\ 24.55].$$

**Example 17.10.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution $N\left(\mu, \sigma^2\right)$, where $\mu$ and $\sigma^2$ are unknown parameters. What is the shortest 90% confidence interval for the standard deviation $\sigma$ ?

**Answer:** Let $S^2$ be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Using this random variable as a pivot, we can construct a $100(1-\alpha)\%$ confidence interval for $\sigma$ from

$$1-\alpha = P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right)$$

by suitably choosing the constants $a$ and $b$. Hence, the confidence interval for $\sigma$ is given by

$$\left[\sqrt{\frac{(n-1)S^2}{b}},\ \sqrt{\frac{(n-1)S^2}{a}}\ \right].$$

The length of this confidence interval is given by

$$L(a,b) = S\sqrt{n-1}\ \left[\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}}\right].$$

In order to find the shortest confidence interval, we should find a pair of constants $a$ and $b$ such that $L(a,b)$ is minimum. Thus, we have a constraint minimization problem. That is

$$\left.\begin{array}{l} \text{Minimize } L(a,b) \\ \text{Subject to the condition} \\ \displaystyle\int_a^b f(u)du = 1-\alpha, \end{array}\right\} \qquad \text{(MP)}$$

where

$$f(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\ 2^{\frac{n-1}{2}}}\ x^{\frac{n-1}{2}-1}\, e^{-\frac{x}{2}}.$$

Differentiating $L$ with respect to $a$, we get

$$\frac{dL}{da} = S\sqrt{n-1}\ \left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\ \frac{db}{da}\right).$$

From

$$\int_a^b f(u)\,du = 1-\alpha,$$

we find the derivative of $b$ with respect to $a$ as follows:

$$\frac{d}{da}\int_a^b f(u)\,du = \frac{d}{da}(1-\alpha)$$

that is

$$f(b)\,\frac{db}{da} - f(a) = 0.$$

Thus, we have

$$\frac{db}{da} = \frac{f(a)}{f(b)}.$$

Letting this into the expression for the derivative of $L$, we get

$$\frac{dL}{da} = S\sqrt{n-1}\ \left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\ \frac{f(a)}{f(b)}\right).$$

Setting this derivative to zero, we get

$$S\sqrt{n-1}\ \left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\ \frac{f(a)}{f(b)}\right) = 0$$

which yields

$$a^{\frac{3}{2}}\, f(a) = b^{\frac{3}{2}}\, f(b).$$

Using the form of $f$, we get from the above expression

$$a^{\frac{3}{2}} a^{\frac{n-3}{2}} e^{-\frac{a}{2}} = b^{\frac{3}{2}} b^{\frac{n-3}{2}} e^{-\frac{b}{2}}$$

that is

$$a^{\frac{n}{2}}\, e^{-\frac{a}{2}} = b^{\frac{n}{2}}\, e^{-\frac{b}{2}}.$$

From this we get

$$\ln\left(\frac{a}{b}\right) = \left(\frac{a-b}{n}\right).$$

Hence to obtain the pair of constants $a$ and $b$ that will produce the shortest confidence interval for $\sigma$, we have to solve the following system of nonlinear equations

$$\left.\begin{aligned} \int_a^b f(u)\, du = 1-\alpha \\ \ln\left(\frac{a}{b}\right) = \frac{a-b}{n}. \end{aligned}\right\} \qquad (\star)$$

If $a_o$ and $b_o$ are solutions of $(\star)$, then the shortest confidence interval for $\sigma$ is given by

$$\left[\sqrt{\frac{(n-1)S^2}{b_o}},\ \ \sqrt{\frac{(n-1)S^2}{a_o}}\ \right].$$

Since this system of nonlinear equations is hard to solve analytically, numerical solutions are given in statistical literature in the form of a table for finding the shortest interval for the variance.

## 17.5. Confidence Interval for Parameter of some Distributions not belonging to the Location-Scale Family

In this section, we illustrate the pivotal quantity method for finding confidence intervals for a parameter $\theta$ when the density function does not belong to the location-scale family. The following density functions does not belong to the location-scale family:

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

or

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

We will construct interval estimators for the parameters in these density functions. The same idea for finding the interval estimators can be used to find interval estimators for parameters of density functions that belong to the location-scale family such as

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

To find the pivotal quantities for the above mentioned distributions and others we need the following three results. The first result is Theorem 6.2 while the proof of the second result is easy and we leave it to the reader.

**Theorem 17.1.** Let $F(x;\theta)$ be the cumulative distribution function of a continuous random variable $X$. Then

$$F(X;\theta) \sim UNIF(0,1).$$

**Theorem 17.2.** If $X \sim UNIF(0,1)$, then

$$-\ln X \sim EXP(1).$$

**Theorem 17.3.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

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

where $\theta > 0$ is a parameter. Then the random variable

$$\frac{2}{\theta}\sum_{i=1}^{n} X_i \sim \chi^2(2n)$$

**Proof:** Let $Y = \frac{2}{\theta}\sum_{i=1}^{n} X_i$. Now we show that the sampling distribution of $Y$ is chi-square with $2n$ degrees of freedom. We use the moment generating method to show this. The moment generating function of $Y$ is given by

$$\begin{aligned} M_Y(t) &= M_{\frac{2}{\theta}\sum_{i=1}^{n} X_i}(t) \\ &= \prod_{i=1}^{n} M_{X_i}\left(\frac{2}{\theta}t\right) \\ &= \prod_{i=1}^{n}\left(1-\theta\,\frac{2}{\theta}t\right)^{-1} \\ &= (1-2t)^{-n} \\ &= (1-2t)^{-\frac{2n}{2}}. \end{aligned}$$

Since $(1-2t)^{-\frac{2n}{2}}$ corresponds to the moment generating function of a chi-square random variable with $2n$ degrees of freedom, we conclude that

$$\frac{2}{\theta}\sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

**Theorem 17.4.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Then the random variable $-2\theta\sum_{i=1}^{n} \ln X_i$ has a chi-square distribution with $2n$ degree of freedoms.

**Proof:** We are given that

$$X_i \sim \theta\, x^{\theta-1}, \qquad 0 < x < 1.$$

Hence, the cdf of $f$ is

$$F(x;\theta) = \int_0^x \theta\, x^{\theta-1} dx = x^\theta.$$

Thus by Theorem 17.1, each

$$F\left(X_i;\theta\right) \sim UNIF(0,1),$$

that is

$$X_i^\theta \sim UNIF(0,1).$$

By Theorem 17.2, each

$$-\ln X_i^\theta \sim EXP(1),$$

that is

$$-\theta\, \ln X_i \sim EXP(1).$$

By Theorem 17.3 (with $\theta = 1$), we obtain

$$-2\,\theta \sum_{i=1}^{n} \ln X_i \sim \chi^2(2n).$$

Hence, the sampling distribution of $-2\,\theta \sum_{i=1}^{n} \ln X_i$ is chi-square with $2n$ degree of freedoms.

The following theorem whose proof follows from Theorems 17.1, 17.2 and 17.3 is the key to finding pivotal quantity of many distributions that do not belong to the location-scale family. Further, this theorem can also be used for finding the pivotal quantities for parameters of some distributions that belong the location-scale family.

**Theorem 17.5.** Let $X_1, X_2, ..., X_n$ be a random sample from a continuous population $X$ with a distribution function $F(x;\theta)$. If $F(x;\theta)$ is monotone in $\theta$, then the statistic $Q = -2\sum_{i=1}^{n} \ln F\left(X_i;\theta\right)$ is a pivotal quantity and has a chi-square distribution with $2n$ degrees of freedom (that is, $Q \sim \chi^2(2n)$).

It should be noted that the condition $F(x;\theta)$ is monotone in $\theta$ is needed to ensure an interval. Otherwise we may get a confidence region instead of a confidence interval. Further note that the statistic $-2\sum_{i=1}^{n} \ln\left(1 - F\left(X_i;\theta\right)\right)$ is also has a chi-square distribution with $2n$ degrees of freedom, that is

$$-2\sum_{i=1}^{n} \ln\left(1 - F\left(X_i;\theta\right)\right) \sim \chi^2(2n).$$

**Example 17.11.** If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter, what is a $100(1-\alpha)\%$ confidence interval for $\theta$?

**Answer:** To construct a confidence interval for $\theta$, we need a pivotal quantity. That is, we need a random variable which is a function of the sample and the parameter, and whose probability distribution is known but does not involve $\theta$. We use the random variable

$$Q = -2\,\theta \sum_{i=1}^{n} \ln X_i \sim \chi^2(2n)$$

as the pivotal quantity. The $100(1-\alpha)\%$ confidence interval for $\theta$ can be constructed from

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \le Q \le \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\ &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \le -2\,\theta \sum_{i=1}^{n} \ln X_i \le \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\ &= P\left(\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i} \le \theta \le \frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i}\right). \end{aligned}$$

Hence, $100(1-\alpha)\%$ confidence interval for $\theta$ is given by

$$\left[\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i}, \quad \frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i}\right].$$

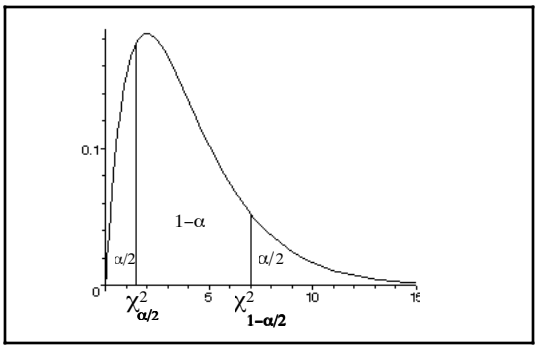
Here $\chi^2_{1-\frac{\alpha}{2}}(2n)$ denotes the $\left(1-\frac{\alpha}{2}\right)$-quantile of a chi-square random variable $Y$, that is

$$P(Y \le \chi^2_{1-\frac{\alpha}{2}}(2n)) = 1 - \frac{\alpha}{2}$$

and $\chi^2_{\frac{\alpha}{2}}(2n)$ similarly denotes $\frac{\alpha}{2}$-quantile of $Y$, that is

$$P\left(Y \le \chi^2_{\frac{\alpha}{2}}(2n)\right) = \frac{\alpha}{2}$$

for $\alpha \leq 0.5$ (see figure below).




**Example 17.12.** If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter, then what is the $100(1-\alpha)\%$ confidence interval for $\theta$?

**Answer:** The cumulation density function of $f(x;\theta)$ is

$$F(x;\theta) = \begin{cases} \frac{x}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$\begin{aligned} -2\sum_{i=1}^{n} \ln F(X_i;\theta) &= -2\sum_{i=1}^{n} \ln\left(\frac{X_i}{\theta}\right) \\ &= 2n \ln\theta - 2\sum_{i=1}^{n} \ln X_i \end{aligned}$$

by Theorem 17.5, the quantity $2n \ln\theta - 2\sum_{i=1}^{n} \ln X_i \sim \chi^2(2n)$. Since $2n \ln\theta - 2\sum_{i=1}^{n} \ln X_i$ is a function of the sample and the parameter and its distribution is independent of $\theta$, it is a pivot for $\theta$. Hence, we take

$$Q(X_1, X_2, ..., X_n, \theta) = 2n \ln\theta - 2\sum_{i=1}^{n} \ln X_i.$$

The $100(1-\alpha)\%$ confidence interval for $\theta$ can be constructed from

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq Q \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\
&= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq 2n\ \ln\theta - 2\sum_{i=1}^{n}\ln X_i \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\
&= P\left(\chi^2_{\frac{\alpha}{2}}(2n) + 2\sum_{i=1}^{n}\ln X_i \leq 2n\ \ln\theta \leq \chi^2_{1-\frac{\alpha}{2}}(2n) + 2\sum_{i=1}^{n}\ln X_i\right) \\
&= P\left(e^{\frac{1}{2n}\left\{\chi^2_{\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}} \leq \theta \leq e^{\frac{1}{2n}\left\{\chi^2_{1-\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}}\right).
\end{aligned}$$

Hence, $100(1-\alpha)\%$ confidence interval for $\theta$ is given by

$$\left[e^{\frac{1}{2n}\left\{\chi^2_{\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}},\quad e^{\frac{1}{2n}\left\{\chi^2_{1-\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}}\right].$$

The density function of the following example belongs to the scale family. However, one can use Theorem 17.5 to find a pivot for the parameter and determine the interval estimators for the parameter.

**Example 17.13.** If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

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

where $\theta > 0$ is a parameter, then what is the $100(1-\alpha)\%$ confidence interval for $\theta$?

**Answer:** The cumulative density function $F(x;\theta)$ of the density function

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

is given by

$$F(x;\theta) = 1 - e^{-\frac{x}{\theta}}.$$

Hence

$$-2\sum_{i=1}^{n}\ln\left(1 - F\left(X_i;\theta\right)\right) = \frac{2}{\theta}\sum_{i=1}^{n}X_i.$$

Thus

$$\frac{2}{\theta}\sum_{i=1}^{n}X_i \sim \chi^2(2n).$$

We take $Q = \frac{2}{\theta}\sum_{i=1}^{n}X_i$ as the pivotal quantity. The $100(1-\alpha)\%$ confidence interval for $\theta$ can be constructed using

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \le Q \le \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\
&= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \le \frac{2}{\theta}\sum_{i=1}^{n}X_i \le \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\
&= P\left(\frac{2\sum_{i=1}^{n}X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n)} \le \theta \le \frac{2\sum_{i=1}^{n}X_i}{\chi^2_{\frac{\alpha}{2}}(2n)}\right).
\end{aligned}$$

Hence, $100(1-\alpha)\%$ confidence interval for $\theta$ is given by

$$\left[\frac{2\sum_{i=1}^{n}X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n)}, \frac{2\sum_{i=1}^{n}X_i}{\chi^2_{\frac{\alpha}{2}}(2n)}\right].$$

In this section, we have seen that $100(1-\alpha)\%$ confidence interval for the parameter $\theta$ can be constructed by taking the pivotal quantity $Q$ to be either

$$Q = -2\sum_{i=1}^{n}\ln F(X_i;\theta)$$

or

$$Q = -2\sum_{i=1}^{n}\ln\left(1-F(X_i;\theta)\right).$$

In either case, the distribution of $Q$ is chi-squared with $2n$ degrees of freedom, that is $Q \sim \chi^2(2n)$. Since chi-squared distribution is not symmetric about the $y$-axis, the confidence intervals constructed in this section do not have the shortest length. In order to have a shortest confidence interval one has to solve the following minimization problem:

$$\left.\begin{array}{l}\text{Minimize } L(a,b) \\ \text{Subject to the condition } \displaystyle\int_a^b f(u)du = 1-\alpha,\end{array}\right\} \qquad \text{(MP)}$$

where

$$f(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\ 2^{\frac{n-1}{2}}}\ x^{\frac{n-1}{2}-1}e^{-\frac{x}{2}}.$$

In the case of Example 17.13, the minimization process leads to the following system of nonlinear equations

$$\left.\begin{aligned} \int_a^b f(u)\,du &= 1-\alpha \\ \ln\left(\frac{a}{b}\right) &= \frac{a-b}{2(n+1)}. \end{aligned}\right\} \qquad \text{(NE)}$$

If $a_o$ and $b_o$ are solutions of (NE), then the shortest confidence interval for $\theta$ is given by

$$\left[\ \frac{2\sum_{i=1}^n X_i}{b_o}\ ,\quad \frac{2\sum_{i=1}^n X_i}{a_o}\ \right].$$

### 17.6. Approximate Confidence Interval for Parameter with MLE

In this section, we discuss how to construct an approximate $(1-\alpha)100\%$ confidence interval for a population parameter $\theta$ using its maximum likelihood estimator $\widehat{\theta}$. Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with density $f(x;\theta)$. Let $\widehat{\theta}$ be the maximum likelihood estimator of $\theta$. If the sample size $n$ is large, then using asymptotic property of the maximum likelihood estimator, we have

$$\frac{\widehat{\theta} - E\left(\widehat{\theta}\right)}{\sqrt{Var\left(\widehat{\theta}\right)}} \sim N(0,1) \qquad \text{as} \quad n \to \infty,$$

where $Var\left(\widehat{\theta}\right)$ denotes the variance of the estimator $\widehat{\theta}$. Since, for large $n$, the maximum likelihood estimator of $\theta$ is unbiased, we get

$$\frac{\widehat{\theta} - \theta}{\sqrt{Var\left(\widehat{\theta}\right)}} \sim N(0,1) \qquad \text{as} \quad n \to \infty.$$

The variance $Var\left(\widehat{\theta}\right)$ can be computed directly whenever possible or using the Cramér-Rao lower bound

$$Var\left(\widehat{\theta}\right) \geq \frac{-1}{E\left[\frac{d^2 \ln L(\theta)}{d\theta^2}\right]}.$$

Now using $Q = \frac{\widehat{\theta}-\theta}{\sqrt{Var\left(\widehat{\theta}\right)}}$ as the pivotal quantity, we construct an approximate $(1-\alpha)100\%$ confidence interval for $\theta$ as

$$\begin{aligned} 1-\alpha &= P\left(-z_{\frac{\alpha}{2}} \leq Q \leq z_{\frac{\alpha}{2}}\right) \\ &= P\left(-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\theta}-\theta}{\sqrt{Var\left(\widehat{\theta}\right)}} \leq z_{\frac{\alpha}{2}}\right). \end{aligned}$$

If $Var\left(\widehat{\theta}\right)$ is free of $\theta$, then have

$$1-\alpha = P\left(\widehat{\theta} - z_{\frac{\alpha}{2}}\sqrt{Var\left(\widehat{\theta}\right)} \leq \theta \leq \widehat{\theta} + z_{\frac{\alpha}{2}}\sqrt{Var\left(\widehat{\theta}\right)}\right).$$

Thus $100(1-\alpha)\%$ approximate confidence interval for $\theta$ is

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\sqrt{Var\left(\widehat{\theta}\right)},\ \ \widehat{\theta} + z_{\frac{\alpha}{2}}\sqrt{Var\left(\widehat{\theta}\right)}\right]$$

provided $Var\left(\widehat{\theta}\right)$ is free of $\theta$.

**Remark 17.5.** In many situations $Var\left(\widehat{\theta}\right)$ is not free of the parameter $\theta$. In those situations we still use the above form of the confidence interval by replacing the parameter $\theta$ by $\widehat{\theta}$ in the expression of $Var\left(\widehat{\theta}\right)$.

Next, we give some examples to illustrate this method.

**Example 17.14.** Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with probability density function

$$f(x;p) = \begin{cases} p^x\,(1-p)^{(1-x)} & \text{if } x = 0, 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is a $100(1-\alpha)\%$ approximate confidence interval for the parameter $p$?

**Answer:** The likelihood function of the sample is given by

$$L(p) = \prod_{i=1}^{n} p^{x_i}\,(1-p)^{(1-x_i)}.$$

Taking the logarithm of the likelihood function, we get

$$\ln L(p) = \sum_{i=1}^{n} \left[ x_i \ln p + (1 - x_i) \ln(1 - p) \right].$$

Differentiating, the above expression, we get

$$\frac{d \ln L(p)}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - x_i).$$

Setting this equals to zero and solving for $p$, we get

$$\frac{n\overline{x}}{p} - \frac{n - n\overline{x}}{1-p} = 0,$$

that is

$$(1-p)\, n\, \overline{x} = p\, (n - n\, \overline{x}),$$

which is

$$n\, \overline{x} - p\, n\, \overline{x} = p\, n - p\, n\, \overline{x}.$$

Hence

$$p = \overline{x}.$$

Therefore, the maximum likelihood estimator of $p$ is given by

$$\widehat{p} = \overline{X}.$$

The variance of $\overline{X}$ is

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Since $X \sim Ber(p)$, the variance $\sigma^2 = p(1-p)$, and

$$Var\left(\widehat{p}\right) = Var\left(\overline{X}\right) = \frac{p(1-p)}{n}.$$

Since $Var\left(\widehat{p}\right)$ is not free of the parameter $p$, we replave $p$ by $\widehat{p}$ in the expression of $Var\left(\widehat{p}\right)$ to get

$$Var\left(\widehat{p}\right) \simeq \frac{\widehat{p}\,(1-\widehat{p})}{n}.$$

The $100(1-\alpha)\%$ approximate confidence interval for the parameter $p$ is given by

$$\left[ \widehat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}},\ \ \widehat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \right]$$

which is

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}},\ \ \overline{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}}\right].$$

The above confidence interval is a $100(1-\alpha)\%$ approximate confidence interval for proportion.

**Example 17.15.** A poll was taken of university students before a student election. Of 78 students contacted, 33 said they would vote for Mr. Smith. The population may be taken as 2200. Obtain 95% confidence limits for the proportion of voters in the population in favor of Mr. Smith.

**Answer:** The sample proportion $\widehat{p}$ is given by

$$\widehat{p} = \frac{33}{78} = 0.4231.$$

Hence

$$\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} = \sqrt{\frac{(0.4231)\,(0.5769)}{78}} = 0.0559.$$

The $2.5^{\text{th}}$ percentile of normal distribution is given by

$$z_{0.025} = 1.96 \qquad \text{(From table)}.$$

Hence, the lower confidence limit of 95% confidence interval is

$$\begin{aligned}
&\widehat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \\
&= 0.4231 - (1.96)\,(0.0559) \\
&= 0.4231 - 0.1096 \\
&= 0.3135.
\end{aligned}$$

Similarly, the upper confidence limit of 95% confidence interval is

$$\begin{aligned}
&\widehat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \\
&= 0.4231 + (1.96)\,(0.0559) \\
&= 0.4231 + 0.1096 \\
&= 0.5327.
\end{aligned}$$

Hence, the 95% confidence limits for the proportion of voters in the population in favor of Smith are 0.3135 and 0.5327.

**Remark 17.6.** In Example 17.15, the 95% percent approximate confidence interval for the parameter $p$ was $[0.3135, 0.5327]$. This confidence interval can be improved to a shorter interval by means of a quadratic inequality. Now we explain how the interval can be improved. First note that in Example 17.14, which we are using for Example 17.15, the approximate value of the variance of the ML estimator $\widehat{p}$ was obtained to be $\sqrt{\frac{p(1-p)}{n}}$. However, this is the exact variance of $\widehat{p}$. Now the pivotal quantity $Q = \frac{\widehat{p}-p}{\sqrt{Var(\widehat{p})}}$ becomes

$$Q = \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Using this pivotal quantity, we can construct a 95% confidence interval as

$$0.05 = P\left(-z_{0.025} \leq \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0.025}\right)$$

$$= P\left(\left|\frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq 1.96\right).$$

Using $\widehat{p} = 0.4231$ and $n = 78$, we solve the inequality

$$\left|\frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq 1.96$$

which is

$$\left|\frac{0.4231-p}{\sqrt{\frac{p(1-p)}{78}}}\right| \leq 1.96.$$

Squaring both sides of the above inequality and simplifying, we get

$$78\,(0.4231-p)^2 \leq (1.96)^2\,(p-p^2).$$

The last inequality is equivalent to

$$13.96306158 - 69.84520000\,p + 81.84160000\,p^2 \leq 0.$$

Solving this quadratic inequality, we obtain [0.3196, 0.5338] as a 95% confidence interval for $p$. This interval is an improvement since its length is 0.2142 where as the length of the interval [0.3135, 0.5327] is 0.2192.

**Example 17.16.** If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**Answer:** The likelihood function $L(\theta)$ of the sample is

$$L(\theta) = \prod_{i=1}^{n} \theta\, x_i^{\theta-1}.$$

Hence

$$\ln L(\theta) = n\, \ln\theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i.$$

The first derivative of the logarithm of the likelihood function is

$$\frac{d}{d\theta} \ln L(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i.$$

Setting this derivative to zero and solving for $\theta$, we obtain

$$\theta = -\frac{n}{\sum_{i=1}^{n} \ln x_i}.$$

Hence, the maximum likelihood estimator of $\theta$ is given by

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln X_i}.$$

Finding the variance of this estimator is difficult. We compute its variance by computing the Cramér-Rao bound for this estimator. The second derivative of the logarithm of the likelihood function is given by

$$\begin{aligned} \frac{d^2}{d\theta^2} \ln L(\theta) &= \frac{d}{d\theta} \left( \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i \right) \\ &= -\frac{n}{\theta^2}. \end{aligned}$$

Hence

$$E\left( \frac{d^2}{d\theta^2} \ln L(\theta) \right) = -\frac{n}{\theta^2}.$$

Therefore

$$Var\left(\widehat{\theta}\right) \geq \frac{\theta}{n}.$$

Thus we take

$$Var\left(\widehat{\theta}\right) \simeq \frac{\theta}{n}.$$

Since $Var\left(\widehat{\theta}\right)$ has $\theta$ in its expression, we replace the unknown $\theta$ by its estimate $\widehat{\theta}$ so that

$$Var\left(\widehat{\theta}\right) \simeq \frac{\widehat{\theta}^2}{n}.$$

The $100(1-\alpha)\%$ approximate confidence interval for $\theta$ is given by

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\,\frac{\widehat{\theta}}{\sqrt{n}},\ \ \widehat{\theta} + z_{\frac{\alpha}{2}}\,\frac{\widehat{\theta}}{\sqrt{n}}\right],$$

which is

$$\left[-\frac{n}{\sum_{i=1}^{n}\ln X_i} + z_{\frac{\alpha}{2}}\,\left(\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i}\right),\ \ -\frac{n}{\sum_{i=1}^{n}\ln X_i} - z_{\frac{\alpha}{2}}\,\left(\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i}\right)\right].$$

**Remark 17.7.** In the next section 17.2, we derived the exact confidence interval for $\theta$ when the population distribution in exponential. The exact $100(1-\alpha)\%$ confidence interval for $\theta$ was given by

$$\left[-\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i},\ \ -\frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i}\right].$$

Note that this exact confidence interval is not the shortest confidence interval for the parameter $\theta$.

**Example 17.17.** If $X_1, X_2, ..., X_{49}$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter, what are 90% *approximate* and *exact* confidence intervals for $\theta$ if $\sum_{i=1}^{49}\ln X_i = -0.7567$?

**Answer:** We are given the followings:

$$n = 49$$

$$\sum_{i=1}^{49}\ln X_i = -0.7576$$

$$1-\alpha = 0.90.$$

Hence, we get

$$z_{0.05} = 1.64,$$

$$\frac{n}{\sum_{i=1}^{n} \ln X_i} = \frac{49}{-0.7567} = -64.75$$

and

$$\frac{\sqrt{n}}{\sum_{i=1}^{n} \ln X_i} = \frac{7}{-0.7567} = -9.25.$$

Hence, the approximate confidence interval is given by

$$[64.75 - (1.64)(9.25), \quad 64.75 + (1.64)(9.25)]$$

that is $[49.58,\ 79.92]$.

Next, we compute the exact 90% confidence interval for $\theta$ using the formula

$$\left[ -\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln X_i}, \quad -\frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln X_i} \right].$$

From chi-square table, we get

$$\chi^2_{0.05}(98) = 77.93 \qquad \text{and} \qquad \chi^2_{0.95}(98) = 124.34.$$

Hence, the exact 90% confidence interval is

$$\left[ \frac{77.93}{(2)(0.7567)}, \quad \frac{124.34}{(2)(0.7567)} \right]$$

that is $[51.49,\ 82.16]$.

**Example 17.18.** If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} (1-\theta)\,\theta^x & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**Answer:** The logarithm of the likelihood function of the sample is

$$\ln L(\theta) = \ln \theta \sum_{i=1}^{n} x_i + n \ln(1-\theta).$$

Differentiating we see obtain

$$\frac{d}{d\theta}\ln L(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n}{1-\theta}.$$

Equating this derivative to zero and solving for $\theta$, we get $\theta = \frac{\overline{x}}{1+\overline{x}}$. Thus, the maximum likelihood estimator of $\theta$ is given by

$$\widehat{\theta} = \frac{\overline{X}}{1+\overline{X}}.$$

Next, we find the variance of this estimator using the Cramér-Rao lower bound. For this, we need the second derivative of $\ln L(\theta)$. Hence

$$\frac{d^2}{d\theta^2}\ln L(\theta) = -\frac{n\overline{x}}{\theta^2} - \frac{n}{(1-\theta)^2}.$$

Therefore

$$\begin{aligned}
&E\left(\frac{d^2}{d\theta^2}\ln L(\theta)\right) \\
&\qquad = E\left(-\frac{n\overline{X}}{\theta^2} - \frac{n}{(1-\theta)^2}\right) \\
&\qquad = \frac{n}{\theta^2}E\left(\overline{X}\right) - \frac{n}{(1-\theta)^2} \\
&\qquad = \frac{n}{\theta^2}\frac{1}{(1-\theta)} - \frac{n}{(1-\theta)^2} \qquad (\text{since each } \mathrm{X_i} \sim \mathrm{GEO}(1-\theta)) \\
&\qquad = -\frac{n}{\theta(1-\theta)}\left[\frac{1}{\theta} + \frac{\theta}{1-\theta}\right] \\
&\qquad = -\frac{n\left(1-\theta+\theta^2\right)}{\theta^2\left(1-\theta\right)^2}.
\end{aligned}$$

Therefore

$$Var\left(\widehat{\theta}\right) \simeq \frac{\widehat{\theta}^2\left(1-\widehat{\theta}\right)^2}{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}.$$

The $100(1-\alpha)\%$ approximate confidence interval for $\theta$ is given by

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{\sqrt{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}},\ \widehat{\theta} + z_{\frac{\alpha}{2}}\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{\sqrt{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}}\right],$$

where

$$\widehat{\theta} = \frac{\overline{X}}{1+\overline{X}}.$$

### 17.7. The Statistical or General Method

Now we briefly describe the statistical or general method for constructing a confidence interval. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density $f(x;\theta)$, where $\theta$ is a unknown parameter. We want to determine an interval estimator for $\theta$. Let $T(X_1, X_2, ..., X_n)$ be some statistics having the density function $g(t;\theta)$. Let $p_1$ and $p_2$ be two fixed positive number in the open interval $(0,1)$ with $p_1 + p_2 < 1$. Now we define two functions $h_1(\theta)$ and $h_2(\theta)$ as follows:

$$p_1 = \int_{-\infty}^{h_1(\theta)} g(t;\theta)\,dt \qquad \text{and} \qquad p_2 = \int_{-\infty}^{h_2(\theta)} g(t;\theta)\,dt$$

such that

$$P\left(h_1(\theta) < T(X_1, X_2, ..., X_n) < h_2(\theta)\right) = 1 - p_1 - p_2.$$

If $h_1(\theta)$ and $h_2(\theta)$ are monotone functions in $\theta$, then we can find a confidence interval

$$P\left(u_1 < \theta < u_2\right) = 1 - p_1 - p_2$$

where $u_1 = u_1(t)$ and $u_2 = u_2(t)$. The statistics $T(X_1, X_2, ..., X_n)$ may be a sufficient statistics, or a maximum likelihood estimator. If we minimize the length $u_2 - u_1$ of the confidence interval, subject to the condition $1 - p_1 - p_2 = 1 - \alpha$ for $0 < \alpha < 1$, we obtain the shortest confidence interval based on the statistics $T$.

### 17.8. Criteria for Evaluating Confidence Intervals

In many situations, one can have more than one confidence intervals for the same parameter $\theta$. Thus it necessary to have a set of criteria to decide whether a particular interval is better than the other intervals. Some well known criteria are: (1) Shortest Length and (2) Unbiasedness. Now we only briefly describe these criteria.

The criterion of shortest length demands that a good $100(1-\alpha)\%$ confidence interval $[L, U]$ of a parameter $\theta$ should have the shortest length $\ell = U - L$. In the pivotal quantity method one finds a pivot $Q$ for a parameter $\theta$ and then converting the probability statement

$$P(a < Q < b) = 1 - \alpha$$

to

$$P(L < \theta < U) = 1 - \alpha$$

obtains a $100(1-\alpha)\%$ confidence interval for $\theta$. If the constants $a$ and $b$ can be found such that the difference $U - L$ depending on the sample $X_1, X_2, ..., X_n$ is minimum for every realization of the sample, then the random interval $[L, U]$ is said to be the shortest confidence interval based on $Q$.

If the pivotal quantity $Q$ has certain type of density functions, then one can easily construct confidence interval of shortest length. The following result is important in this regard.

**Theorem 17.6.** Let the density function of the pivot $Q \sim h(q;\theta)$ be continuous and unimodal. If in some interval $[a, b]$ the density function $h$ has a mode, and satisfies conditions (i) $\int_a^b h(q;\theta)dq = 1-\alpha$ and (ii) $h(a) = h(b) > 0$, then the interval $[a, b]$ is of the shortest length among all intervals that satisfy condition (i).

If the density function is not unimodal, then minimization of $\ell$ is necessary to construct a shortest confidence interval. One of the weakness of this shortest length criterion is that in some cases, $\ell$ could be a random variable. Often, the expected length of the interval $E(\ell) = E(U - L)$ is also used as a criterion for evaluating the goodness of an interval. However, this too has weaknesses. A weakness of this criterion is that minimization of $E(\ell)$ depends on the unknown true value of the parameter $\theta$. If the sample size is very large, then every approximate confidence interval constructed using MLE method has minimum expected length.

A confidence interval is only shortest based on a particular pivot $Q$. It is possible to find another pivot $Q^\star$ which may yield even a shorter interval than the shortest interval found based on $Q$. The question naturally arises is how to find the pivot that gives the shortest confidence interval among all other pivots. It has been pointed out that a pivotal quantity $Q$ which is a some function of the complete and sufficient statistics gives shortest confidence interval.

Unbiasedness, is yet another criterion for judging the goodness of an interval estimator. The unbiasedness is defined as follow. A $100(1 - \alpha)\%$ confidence interval $[L, U]$ of the parameter $\theta$ is said to be unbiased if

$$P(L \leq \theta^\star \leq U) \begin{cases} \geq 1 - \alpha & \text{if } \theta^\star = \theta \\ \\ \leq 1 - \alpha & \text{if } \theta^\star \neq \theta. \end{cases}$$

### 17.9. Review Exercises

**1.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with gamma density function

$$f(x;\theta,\beta) = \begin{cases} \frac{1}{\Gamma(\beta)\,\theta^{\beta}}\, x^{\beta-1}\, e^{-\frac{x}{\theta}} & \text{for } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[ \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n\beta)}, \quad \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\frac{\alpha}{2}}(2n\beta)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for the parameter $\theta$.

**2.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with Weibull density function

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

where $\theta$ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[ \frac{2\sum_{i=1}^{n} X_i^{\beta}}{\chi^2_{1-\frac{\alpha}{2}}(2n)}, \quad \frac{2\sum_{i=1}^{n} X_i^{\beta}}{\chi^2_{\frac{\alpha}{2}}(2n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for the parameter $\theta$.

**3.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with Pareto density function

$$f(x;\theta,\beta) = \begin{cases} \theta\, \beta^{\theta}\, x^{-(\theta+1)} & \text{for } \beta \le x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[ \frac{2\sum_{i=1}^{n} \ln\left(\frac{X_i}{\beta}\right)}{\chi^2_{1-\frac{\alpha}{2}}(2n)}, \quad \frac{2\sum_{i=1}^{n} \ln\left(\frac{X_i}{\beta}\right)}{\chi^2_{\frac{\alpha}{2}}(2n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for $\frac{1}{\theta}$.

**4.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with Laplace density function

$$f(x;\theta) = \frac{1}{2\theta}\, e^{-\frac{|x|}{\theta}}, \qquad -\infty < x < \infty$$

where $\theta$ is an unknown parameter. Show that

$$\left[ \frac{2\sum_{i=1}^{n}|X_i|}{\chi^2_{1-\frac{\alpha}{2}}(2n)}, \quad \frac{2\sum_{i=1}^{n}|X_i|}{\chi^2_{\frac{\alpha}{2}}(2n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for $\theta$.

**5.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} \frac{1}{2\theta^2}\, x^3\, e^{-\frac{x^2}{2\theta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is an unknown parameter. Show that

$$\left[ \frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{1-\frac{\alpha}{2}}(4n)}, \quad \frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{\frac{\alpha}{2}}(4n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for $\theta$.

**6.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta,\beta) = \begin{cases} \beta\,\theta\, \frac{x^{\beta-1}}{(1+x^{\beta})^{\theta+1}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[ \frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln\left(1+X_i^{\beta}\right)}, \quad \frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln\left(1+X_i^{\beta}\right)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for $\theta$.

**7.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Then show that $Q = X_{(1)} - \theta$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for $\theta$.

**8.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Then show that $Q = 2n\left(X_{(1)} - \theta\right)$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for $\theta$.

**9.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Then show that $Q = e^{-\left(X_{(1)}-\theta\right)}$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for $\theta$.

**10.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with uniform density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is an unknown parameter. Then show that $Q = \frac{X_{(n)}}{\theta}$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for $\theta$.

**11.** Let $X_1, X_2, ..., X_n$ be a random sample from a population with uniform density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is an unknown parameter. Then show that $Q = \frac{X_{(n)}-X_{(1)}}{\theta}$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for $\theta$.

**12.** If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x-\theta)^2} & \text{if } \theta \le x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta$ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**13.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} (\theta+1)\, x^{-\theta-2} & \text{if} \quad 1 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. What is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**14.** Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta^2\, x\, e^{-\theta\, x} & \text{if} \quad 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. What is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**15.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x) = \begin{cases} \frac{1}{\beta}\, e^{\frac{-(x-4)}{\beta}} & \text{for} \quad x > 4 \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta > 0$. What is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**16.** Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$. What is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**17.** A sample $X_1, X_2, ..., X_n$ of size $n$ is drawn from a gamma distribution

$$f(x;\beta) = \begin{cases} \frac{x^3 e^{-\frac{x}{\beta}}}{6\beta^4} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is a $100(1-\alpha)\%$ approximate confidence interval for $\theta$ if the sample size is large?

**18.** Let $X_1, X_2, ..., X_n$ be a random sample from a continuous population $X$ with a distribution function $F(x;\theta)$. Show that the statistic $Q = -2\sum_{i=1}^{n} \ln F(X_i;\theta)$ is a pivotal quantity and has a chi-square distribution with $2n$ degrees of freedom.

**19.** Let $X_1, X_2, ..., X_n$ be a random sample from a continuous population $X$ with a distribution function $F(x;\theta)$. Show that the statistic $Q = -2\sum_{i=1}^{n} \ln(1 - F(X_i;\theta))$ is a pivotal quantity and has a chi-square distribution with $2n$ degrees of freedom.

# Chapter 18

# TEST OF STATISTICAL HYPOTHESES FOR PARAMETERS

## 18.1. Introduction

Inferential statistics consists of estimation and hypothesis testing. We have already discussed various methods of finding point and interval estimators of parameters. We have also examined the goodness of an estimator.

Suppose $X_1, X_2, ..., X_n$ is a random sample from a population with probability density function given by

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Further, let $n = 4$ and suppose $x_1 = 0.92$, $x_2 = 0.75$, $x_3 = 0.85$, $x_4 = 0.8$ is a set of random sample data from the above distribution. If we apply the maximum likelihood method, then we will find that the estimator $\widehat{\theta}$ of $\theta$ is

$$\widehat{\theta} = -1 - \frac{4}{\ln(X_1) + \ln(X_2) + \ln(X_3) + \ln(X_2)}.$$

Hence, the maximum likelihood estimate of $\theta$ is

$$\begin{aligned} \widehat{\theta} &= -1 - \frac{4}{\ln(0.92) + \ln(0.75) + \ln(0.85) + \ln(0.80)} \\ &= -1 + \frac{4}{0.7567} = 4.2861 \end{aligned}$$

Therefore, the corresponding probability density function of the population is given by

$$f(x) = \begin{cases} 5.2861\, x^{4.2861} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Since, the point estimate will rarely equal to the true value of $\theta$, we would like to report a range of values with some degree of confidence. If we want to report an interval of values for $\theta$ with a confidence level of 90%, then we need a 90% confidence interval for $\theta$. If we use the pivotal quantity method, then we will find that the confidence interval for $\theta$ is

$$\left[ -1 - \frac{\chi^2_{\frac{\alpha}{2}}(8)}{2\sum_{i=1}^{4} \ln X_i}, \quad -1 - \frac{\chi^2_{1-\frac{\alpha}{2}}(8)}{2\sum_{i=1}^{4} \ln X_i} \right].$$

Since $\chi^2_{0.05}(8) = 2.73$, $\chi^2_{0.95}(8) = 15.51$, and $\sum_{i=1}^{4} \ln(x_i) = -0.7567$, we obtain

$$\left[ -1 + \frac{2.73}{2(0.7567)}, \quad -1 + \frac{15.51}{2(0.7567)} \right]$$

which is

$$[\, 0.803, \;\; 9.249 \,].$$

Thus we may draw inference, at a 90% confidence level, that the population $X$ has the distribution

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases} \qquad (\star)$$

where $\theta \in [0.803, 9.249]$. If we think carefully, we will notice that we have made one assumption. The assumption is that the observable quantity $X$ can be modeled by a density function as shown in $(\star)$. Since, we are concerned with the parametric statistics, our assumption is in fact about $\theta$.

Based on the sample data, we found that an interval estimate of $\theta$ at a 90% confidence level is $[0.803, 9.249]$. But, we assumed that $\theta \in [0.803, 9.249]$. However, we can not be sure that our assumption regarding the parameter is real and is not due to the chance in the random sampling process. The validation of this assumption can be done by the hypothesis test. In this chapter, we discuss testing of statistical hypotheses. Most of the ideas regarding the hypothesis test came from Jerry Neyman and Karl Pearson during 1928-1938.

**Definition 18.1.** A *statistical hypothesis* $H$ is a conjecture about the distribution $f(x;\theta)$ of a population $X$. This conjecture is usually about the

parameter $\theta$ if one is dealing with a parametric statistics; otherwise it is about the form of the distribution of $X$.

**Definition 18.2.** A hypothesis $H$ is said to be a *simple hypothesis* if $H$ completely specifies the density $f(x;\theta)$ of the population; otherwise it is called a *composite hypothesis*.

**Definition 18.3.** The hypothesis to be tested is called the null hypothesis. The negation of the null hypothesis is called the alternative hypothesis. The null and alternative hypotheses are denoted by $H_o$ and $H_a$, respectively.

If $\theta$ denotes a population parameter, then the general format of the null hypothesis and alternative hypothesis is

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \in \Omega_a \tag{$\star$}$$

where $\Omega_o$ and $\Omega_a$ are subsets of the parameter space $\Omega$ with

$$\Omega_o \cap \Omega_a = \emptyset \qquad \text{and} \qquad \Omega_o \cup \Omega_a \subseteq \Omega.$$

**Remark 18.1.** If $\Omega_o \cup \Omega_a = \Omega$, then $(\star)$ becomes

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \notin \Omega_o.$$

If $\Omega_o$ is a singleton set, then $H_o$ reduces to a simple hypothesis. For example, $\Omega_o = \{4.2861\}$, the null hypothesis becomes $H_o : \theta = 4.2861$ and the alternative hypothesis becomes $H_a : \theta \neq 4.2861$. Hence, the null hypothesis $H_o : \theta = 4.2861$ is a simple hypothesis and the alternative $H_a : \theta \neq 4.2861$ is a composite hypothesis.

**Definition 18.4.** A *hypothesis test* is an ordered sequence

$$(X_1, X_2, ..., X_n; H_o, H_a; C)$$

where $X_1, X_2, ..., X_n$ is a random sample from a population $X$ with the probability density function $f(x;\theta)$, $H_o$ and $H_a$ are hypotheses concerning the parameter $\theta$ in $f(x;\theta)$, and $C$ is a Borel set in $\mathbb{R}^n$.

**Remark 18.2.** Borel sets are defined using the notion of $\sigma$-algebra. A collection of subsets $\mathcal{A}$ of a set $S$ is called a $\sigma$-algebra if (i) $S \in \mathcal{A}$, (ii) $A^c \in \mathcal{A}$, whenever $A \in \mathcal{A}$, and (iii) $\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$, whenever $A_1, A_2, ..., A_n, ... \in \mathcal{A}$. The Borel sets are the member of the smallest $\sigma$-algebra containing all open sets

of $\mathbb{R}^n$. Two examples of Borel sets in $\mathbb{R}^n$ are the sets that arise by countable union of closed intervals in $\mathbb{R}^n$, and countable intersection of open sets in $\mathbb{R}^n$.

The set $C$ is called the critical region in the hypothesis test. The critical region is obtained using a *test statistics* $W(X_1, X_2, ..., X_n)$. If the outcome of $(X_1, X_2, ..., X_n)$ turns out to be an element of $C$, then we decide to accept $H_a$; otherwise we accept $H_o$.

Broadly speaking, a hypothesis test is a rule that tells us for which sample values we should decide to accept $H_o$ as true and for which sample values we should decide to reject $H_o$ and accept $H_a$ as true. Typically, a hypothesis test is specified in terms of a test statistics $W$. For example, a test might specify that $H_o$ is to be rejected if the sample total $\sum_{k=1}^{n} X_k$ is less than 8. In this case the critical region $C$ is the set $\{(x_1, x_2, ..., x_n) \,|\, x_1 + x_2 + \cdots + x_n < 8\}$.

### 18.2. A Method of Finding Tests

There are several methods to find test procedures and they are: (1) Likelihood Ratio Tests, (2) Invariant Tests, (3) Bayesian Tests, and (4) Union-Intersection and Intersection-Union Tests. In this section, we only examine likelihood ratio tests.

**Definition 18.5.** The *likelihood ratio test statistic* for testing the simple null hypothesis $H_o : \theta \in \Omega_o$ against the composite alternative hypothesis $H_a : \theta \notin \Omega_o$ based on a set of random sample data $x_1, x_2, ..., x_n$ is defined as

$$W(x_1, x_2, ..., x_n) = \frac{\max\limits_{\theta \in \Omega_o} L(\theta, x_1, x_2, ..., x_n)}{\max\limits_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)},$$

where $\Omega$ denotes the parameter space, and $L(\theta, x_1, x_2, ..., x_n)$ denotes the likelihood function of the random sample, that is

$$L(\theta, x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i; \theta).$$

A *likelihood ratio test* (LRT) is any test that has a critical region $C$ (that is, rejection region) of the form

$$C = \{(x_1, x_2, ..., x_n) \mid W(x_1, x_2, ..., x_n) \leq k\},$$

where $k$ is a number in the unit interval $[0, 1]$.

If $H_o : \theta = \theta_0$ and $H_a : \theta = \theta_a$ are both simple hypotheses, then the likelihood ratio test statistic is defined as

$$W(x_1, x_2, ..., x_n) = \frac{L(\theta_o, x_1, x_2, ..., x_n)}{L(\theta_a, x_1, x_2, ..., x_n)}.$$

Now we give some examples to illustrate this definition.

**Example 18.1.** Let $X_1, X_2, X_3$ denote three independent observations from a distribution with density

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^\theta & \text{if } 0 \le x \le 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the form of the LRT critical region for testing $H_o : \theta = 1$ versus $H_a : \theta = 2$?

**Answer:** In this example, $\theta_o = 1$ and $\theta_a = 2$. By the above definition, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{L\,(\theta_o, x_1, x_2, x_3)}{L\,(\theta_a, x_1, x_2, x_3)} \le k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{(1+\theta_o)^3 \prod_{i=1}^3 x_i^{\theta_o}}{(1+\theta_a)^3 \prod_{i=1}^3 x_i^{\theta_a}} \le k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{8x_1x_2x_3}{27x_1^2x_2^2x_3^2} \le k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{1}{x_1x_2x_3} \le \frac{27}{8}k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; x_1x_2x_3 \ge a, \right\} \end{aligned}$$

where $a$ is some constant. Hence the likelihood ratio test is of the form: "Reject $H_o$ if $\prod_{i=1}^{3} X_i \ge a$."

**Example 18.2.** Let $X_1, X_2, ..., X_{12}$ be a random sample from a normal population with mean zero and variance $\sigma^2$. What is the form of the LRT critical region for testing the null hypothesis $H_o : \sigma^2 = 10$ versus $H_a : \sigma^2 = 5$?

**Answer:** Here $\sigma_o^2 = 10$ and $\sigma_a^2 = 5$. By the above definition, the form of the

critical region is given by (with ${\sigma_o}^2 = 10$ and ${\sigma_a}^2 = 5$)

$$\begin{aligned} C &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \frac{L\left({\sigma_o}^2, x_1, x_2, ..., x_{12}\right)}{L\left({\sigma_a}^2, x_1, x_2, ..., x_{12}\right)} \leq k \right\} \\ &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \prod_{i=1}^{12} \frac{\frac{1}{\sqrt{2\pi\sigma_o^2}} e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_o}\right)^2}}{\frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_a}\right)^2}} \leq k \right\} \\ &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \left(\frac{1}{2}\right)^6 e^{\frac{1}{20}\sum_{i=1}^{12} x_i^2} \leq k \right\} \\ &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \sum_{i=1}^{12} x_i^2 \leq a \right\}, \end{aligned}$$

where $a$ is some constant. Hence the likelihood ratio test is of the form: "Reject $H_o$ if $\sum_{i=1}^{12} X_i^2 \leq a$."

**Example 18.3.** Suppose that $X$ is a random variable about which the hypothesis $H_o : X \sim UNIF(0, 1)$ against $H_a : X \sim N(0, 1)$ is to be tested. What is the form of the LRT critical region based on one observation of $X$?

**Answer:** In this example, $L_o(x) = 1$ and $L_a(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. By the above definition, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \;\middle|\; \frac{L_o(x)}{L_a(x)} \leq k \right\}, \qquad \text{where } k \in [0, \infty) \\ &= \left\{ x \in \mathbb{R} \;\middle|\; \sqrt{2\pi}\, e^{\frac{1}{2}x^2} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \;\middle|\; x^2 \leq 2 \ln\left(\frac{k}{\sqrt{2\pi}}\right) \right\} \\ &= \{ x \in \mathbb{R} \mid x \leq a, \} \end{aligned}$$

where $a$ is some constant. Hence the likelihood ratio test is of the form: "Reject $H_o$ if $X \leq a$."

In the above three examples, we have dealt with the case when null as well as alternative were simple. If the null hypothesis is simple (for example, $H_o : \theta = \theta_o$) and the alternative is a composite hypothesis (for example, $H_a : \theta \neq \theta_o$), then the following algorithm can be used to construct the likelihood ratio critical region:

(1) Find the likelihood function $L(\theta, x_1, x_2, ..., x_n)$ for the given sample.

(2) Find $L(\theta_o, x_1, x_2, ..., x_n)$.
(3) Find $\max_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)$.
(4) Rewrite $\frac{L(\theta_o, x_1, x_2, ..., x_n)}{\max_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)}$ in a "suitable form".
(5) Use step (4) to construct the critical region.

Now we give an example to illustrate these steps.

**Example 18.4.** Let $X$ be a single observation from a population with probability density

$$f(x;\theta) = \begin{cases} \frac{\theta^x \, e^{-\theta}}{x!} & \text{for } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. Find the likelihood ratio critical region for testing the null hypothesis $H_o : \theta = 2$ against the composite alternative $H_a : \theta \neq 2$.

**Answer:** The likelihood function based on one observation $x$ is

$$L(\theta, x) = \frac{\theta^x \, e^{-\theta}}{x!}.$$

Next, we find $L(\theta_o, x)$ which is given by

$$L(2, x) = \frac{2^x \, e^{-2}}{x!}.$$

Our next step is to evaluate $\max_{\theta \geq 0} L(\theta, x)$. For this we differentiate $L(\theta, x)$ with respect to $\theta$, and then set the derivative to 0 and solve for $\theta$. Hence

$$\frac{dL(\theta, x)}{d\theta} = \frac{1}{x!} \left[ e^{-\theta} \, x \, \theta^{x-1} - \theta^x \, e^{-\theta} \right]$$

and $\frac{dL(\theta,x)}{d\theta} = 0$ gives $\theta = x$. Hence

$$\max_{\theta \geq 0} L(\theta, x) = \frac{x^x \, e^{-x}}{x!}.$$

To do the step (4), we consider

$$\frac{L(2, x)}{\max_{\theta \in \Omega} L(\theta, x)} = \frac{\frac{2^x \, e^{-2}}{x!}}{\frac{x^x \, e^{-x}}{x!}}$$

which simplifies to

$$\frac{L(2,x)}{\max\limits_{\theta\in\Omega} L(\theta,x)} = \left(\frac{2e}{x}\right)^x e^{-2}.$$

Thus, the likelihood ratio critical region is given by

$$C = \left\{ x \in \mathbb{R} \;\middle|\; \left(\frac{2e}{x}\right)^x e^{-2} \leq k \right\} = \left\{ x \in \mathbb{R} \;\middle|\; \left(\frac{2e}{x}\right)^x \leq a \right\}$$

where $a$ is some constant. The likelihood ratio test is of the form: "Reject $H_o$ if $\left(\frac{2e}{X}\right)^X \leq a$."

So far, we have learned how to find tests for testing the null hypothesis against the alternative hypothesis. However, we have not considered the goodness of these tests. In the next section, we consider various criteria for evaluating the goodness of an hypothesis test.

### 18.3. Methods of Evaluating Tests

There are several criteria to evaluate the goodness of a test procedure. Some well known criteria are: (1) Powerfulness, (2) Unbiasedness and Invariancy, and (3) Local Powerfulness. In order to examine some of these criteria, we need some terminologies such as error probabilities, power functions, type I error, and type II error. First, we develop these terminologies.

A statistical hypothesis is a conjecture about the distribution $f(x;\theta)$ of the population $X$. This conjecture is usually about the parameter $\theta$ if one is dealing with a parametric statistics; otherwise it is about the form of the distribution of $X$. If the hypothesis completely specifies the density $f(x;\theta)$ of the population, then it is said to be a simple hypothesis; otherwise it is called a composite hypothesis. The hypothesis to be tested is called the null hypothesis. We often hope to reject the null hypothesis based on the sample information. The negation of the null hypothesis is called the alternative hypothesis. The null and alternative hypotheses are denoted by $H_o$ and $H_a$, respectively.

In hypothesis test, the basic problem is to decide, based on the sample information, whether the null hypothesis is true. There are four possible situations that determines our decision is correct or in error. These four situations are summarized below:

| | $H_o$ is true | $H_o$ is false |
|---|---|---|
| Accept $H_o$ | Correct Decision | Type II Error |
| Reject $H_o$ | Type I Error | Correct Decision |

**Definition 18.6.** Let $H_o : \theta \in \Omega_o$ and $H_a : \theta \notin \Omega_o$ be the null and alternative hypothesis to be tested based on a random sample $X_1, X_2, ..., X_n$ from a population $X$ with density $f(x;\theta)$, where $\theta$ is a parameter. The *significance level* of the hypothesis test

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \notin \Omega_o,$$

denoted by $\alpha$, is defined as

$$\alpha = P\left(\text{Type I Error}\right).$$

Thus, the significance level of a hypothesis test we mean the probability of rejecting a true null hypothesis, that is

$$\alpha = P\left(\text{Reject H}_\text{o} \text{ / H}_\text{o} \text{ is true}\right).$$

This is also equivalent to

$$\alpha = P\left(\text{Accept H}_\text{a} \text{ / H}_\text{o} \text{ is true}\right).$$

**Definition 18.7.** Let $H_o : \theta \in \Omega_o$ and $H_a : \theta \notin \Omega_o$ be the null and alternative hypothesis to be tested based on a random sample $X_1, X_2, ..., X_n$ from a population $X$ with density $f(x;\theta)$, where $\theta$ is a parameter. The *probability of type II error* of the hypothesis test

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \notin \Omega_o,$$

denoted by $\beta$, is defined as

$$\beta = P\left(\text{Accept H}_\text{o} \text{ / H}_\text{o} \text{ is false}\right).$$

Similarly, this is also equivalent to

$$\beta = P\left(\text{Accept H}_\text{o} \text{ / H}_\text{a} \text{ is true}\right).$$

**Remark 18.3.** Note that $\alpha$ can be numerically evaluated if the null hypothesis is a simple hypothesis and rejection rule is given. Similarly, $\beta$ can be

evaluated if the alternative hypothesis is simple and rejection rule is known. If null and the alternatives are composite hypotheses, then $\alpha$ and $\beta$ become functions of $\theta$.

**Example 18.5.** Let $X_1, X_2, ..., X_{20}$ be a random sample from a distribution with probability density function

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p \leq \frac{1}{2}$ is a parameter. The hypothesis $H_o : p = \frac{1}{2}$ to be tested against $H_a : p < \frac{1}{2}$. If $H_o$ is rejected when $\sum_{i=1}^{20} X_i \leq 6$, then what is the probability of type I error?

**Answer:** Since each observation $X_i \sim BER(p)$, the sum the observations $\sum_{i=1}^{20} X_i \sim BIN(20, p)$. The probability of type I error is given by

$$\begin{aligned} \alpha &= P\,(\text{Type I Error}) \\ &= P\,(\text{Reject H}_\text{o} \text{ / H}_\text{o} \text{ is true}) \\ &= P\left(\sum_{i=1}^{20} X_i \leq 6 \;\middle/\; \text{H}_\text{o} \text{ is true}\right) \\ &= P\left(\sum_{i=1}^{20} X_i \leq 6 \;\middle/\; H_o : p = \frac{1}{2}\right) \\ &= \sum_{k=0}^{6} \binom{20}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{20-k} \\ &= 0.0577 \qquad \text{(from binomial table).} \end{aligned}$$

Hence the probability of type I error is 0.0577.

**Example 18.6.** Let $p$ represent the proportion of defectives in a manufacturing process. To test $H_o : p \leq \frac{1}{4}$ versus $H_a : p > \frac{1}{4}$, a random sample of size 5 is taken from the process. If the number of defectives is 4 or more, the null hypothesis is rejected. What is the probability of rejecting $H_o$ if $p = \frac{1}{5}$?

**Answer:** Let $X$ denote the number of defectives out of a random sample of size 5. Then $X$ is a binomial random variable with $n = 5$ and $p = \frac{1}{5}$. Hence,

the probability of rejecting $H_o$ is given by

$$\begin{aligned}
\alpha &= P\,(\text{Reject } \mathrm{H_o} \text{ / } \mathrm{H_o} \text{ is true}) \\
&= P\,(X \geq 4 \text{ / } \mathrm{H_o} \text{ is true}) \\
&= P\left(X \geq 4 \;\Big/\; p = \frac{1}{5}\right) \\
&= P\left(X = 4 \;\Big/ p = \frac{1}{5}\right) + P\left(X = 5 \;\Big/ p = \frac{1}{5}\right) \\
&= \binom{5}{4} p^4 (1-p)^1 + \binom{5}{5} p^5 (1-p)^0 \\
&= 5\left(\frac{1}{5}\right)^4 \left(\frac{4}{5}\right) + \left(\frac{1}{5}\right)^5 \\
&= \left(\frac{1}{5}\right)^5 [20+1] \\
&= \frac{21}{3125}.
\end{aligned}$$

Hence the probability of rejecting the null hypothesis $H_o$ is $\frac{21}{3125}$.

**Example 18.7.** A random sample of size 4 is taken from a normal distribution with unknown mean $\mu$ and variance $\sigma^2 > 0$. To test $H_o : \mu = 0$ against $H_a : \mu < 0$ the following test is used: "Reject $H_o$ if and only if $X_1 + X_2 + X_3 + X_4 < -20$." Find the value of $\sigma$ so that the significance level of this test will be closed to 0.14.

**Answer:** Since

$$\begin{aligned}
0.14 &= \alpha \qquad\qquad\qquad\qquad (\text{significance level}) \\
&= P\,(\text{Type I Error}) \\
&= P\,(\text{Reject } \mathrm{H_o} \text{ / } \mathrm{H_o} \text{ is true}) \\
&= P\,(X_1 + X_2 + X_3 + X_4 < -20 \;/ H_o : \mu = 0) \\
&= P\left(\overline{X} < -5 \;/ H_o : \mu = 0\right) \\
&= P\left(\frac{\overline{X} - 0}{\frac{\sigma}{2}} < \frac{-5-0}{\frac{\sigma}{2}}\right) \\
&= P\left(Z < -\frac{10}{\sigma}\right),
\end{aligned}$$

we get from the standard normal table

$$1.08 = \frac{10}{\sigma}.$$

Therefore

$$\sigma = \frac{10}{1.08} = 9.26.$$

Hence, the standard deviation has to be 9.26 so that the significance level will be closed to 0.14.

**Example 18.8.** A normal population has a standard deviation of 16. The critical region for testing $H_o : \mu = 5$ versus the alternative $H_a : \mu = k$ is $\bar{X} > k - 2$. What would be the value of the constant $k$ and the sample size $n$ which would allow the probability of Type I error to be 0.0228 and the probability of Type II error to be 0.1587.

**Answer:** It is given that the population $X \sim N\left(\mu, 16^2\right)$. Since

$$\begin{aligned}
0.0228 &= \alpha \\
&= P\left(\text{Type I Error}\right) \\
&= P\left(\text{Reject } \mathrm{H_o} \text{ / } \mathrm{H_o} \text{ is true}\right) \\
&= P\left(\overline{X} > k - 2 \,/ H_o : \mu = 5\right) \\
&= P\left(\frac{\overline{X} - 5}{\sqrt{\frac{256}{n}}} > \frac{k-7}{\sqrt{\frac{256}{n}}}\right) \\
&= P\left(Z > \frac{k-7}{\sqrt{\frac{256}{n}}}\right) \\
&= 1 - P\left(Z \leq \frac{k-7}{\sqrt{\frac{256}{n}}}\right)
\end{aligned}$$

Hence, from standard normal table, we have

$$\frac{(k-7)\sqrt{n}}{16} = 2$$

which gives

$$(k-7)\sqrt{n} = 32.$$

Similarly

$$\begin{aligned}
0.1587 &= P\,(\text{Type II Error}) \\
&= P\,(\text{Accept } \mathrm{H_o} \text{ / } \mathrm{H_a} \text{ is true}) \\
&= P\left(\overline{X} \le k-2 \;/H_a : \mu = k\right) \\
&= P\left(\frac{\overline{X}-\mu}{\sqrt{\frac{256}{n}}} \le \frac{k-2-\mu}{\sqrt{\frac{256}{n}}} \;\Big/\; H_a : \mu = k\right) \\
&= P\left(\frac{\overline{X}-k}{\sqrt{\frac{256}{n}}} \le \frac{k-2-k}{\sqrt{\frac{256}{n}}}\right) \\
&= P\left(Z \le -\frac{2}{\sqrt{\frac{256}{n}}}\right) \\
&= 1 - P\left(Z \le \frac{2\sqrt{n}}{16}\right).
\end{aligned}$$

Hence $0.1587 = 1 - P\left(Z \le \frac{2\sqrt{n}}{16}\right)$ or $P\left(Z \le \frac{2\sqrt{n}}{16}\right) = 0.8413$. Thus, from the standard normal table, we have

$$\frac{2\sqrt{n}}{16} = 1$$

which yields

$$n = 64.$$

Letting this value of $n$ in

$$(k-7)\sqrt{n} = 32,$$

we see that $k = 11$.

While deciding to accept $H_o$ or $H_a$, we may make a wrong decision. The probability $\gamma$ of a wrong decision can be computed as follows:

$$\begin{aligned}
\gamma &= P\,(\mathrm{H_a} \text{ accepted and } \mathrm{H_o} \text{ is true}) + P\,(\mathrm{H_o} \text{ accepted and } \mathrm{H_a} \text{ is true}) \\
&= P\,(\mathrm{H_a} \text{ accepted / } \mathrm{H_o} \text{ is true})\,P\,(\mathrm{H_o} \text{ is true}) \\
&\quad + P\,(\mathrm{H_o} \text{ accepted / } \mathrm{H_a} \text{ is true})\,P\,(\mathrm{H_a} \text{ is true}) \\
&= \alpha\,P\,(\mathrm{H_o} \text{ is true}) + \beta\,P\,(\mathrm{H_a} \text{ is true}).
\end{aligned}$$

In most cases, the probabilities $P\,(\mathrm{H_o} \text{ is true})$ and $P\,(\mathrm{H_a} \text{ is true})$ are not known. Therefore, it is, in general, not possible to determine the exact

numerical value of the probability $\gamma$ of making a wrong decision. However, since $\gamma$ is a weighted sum of $\alpha$ and $\beta$, and $P\,(\mathrm{H_o}$ is true$)+P\,(\mathrm{H_a}$ is true$)=1$, we have

$$\gamma \leq \max\{\alpha, \beta\}.$$

A good decision rule (or a good test) is the one which yields the smallest $\gamma$. In view of the above inequality, one will have a small $\gamma$ if the probability of type I error as well as probability of type II error are small.

The alternative hypothesis is mostly a composite hypothesis. Thus, it is not possible to find a value for the probability of type II error, $\beta$. For composite alternative, $\beta$ is a function of $\theta$. That is, $\beta : \Omega_o^c :\rightarrow [0,1]$. Here $\Omega_o^c$ denotes the complement of the set $\Omega_o$ in the parameter space $\Omega$. In hypothesis test, instead of $\beta$, one usually considers the *power of the test* $1-\beta(\theta)$, and a small probability of type II error is equivalent to large power of the test.

**Definition 18.8.** Let $H_o : \theta \in \Omega_o$ and $H_a : \theta \notin \Omega_o$ be the null and alternative hypothesis to be tested based on a random sample $X_1, X_2, ..., X_n$ from a population $X$ with density $f(x;\theta)$, where $\theta$ is a parameter. The *power function* of a hypothesis test

$$H_o : \theta \in \Omega_o \qquad \text{versus} \qquad H_a : \theta \notin \Omega_o$$

is a function $\pi : \Omega \rightarrow [0,1]$ defined by

$$\pi(\theta) = \begin{cases} P\,(\text{Type I Error}) & \text{if } \mathrm{H_o} \text{ is true} \\ \\ 1 - P\,(\text{Type II Error}) & \text{if } \mathrm{H_a} \text{ is true.} \end{cases}$$

**Example 18.9.** A manufacturing firm needs to test the null hypothesis $H_o$ that the probability $p$ of a defective item is 0.1 or less, against the alternative hypothesis $H_a : p > 0.1$. The procedure is to select two items at random. If both are defective, $H_o$ is rejected; otherwise, a third is selected. If the third item is defective $H_o$ is rejected. If all other cases, $H_o$ is accepted, what is the power of the test in terms of $p$ (if $H_o$ is true)?

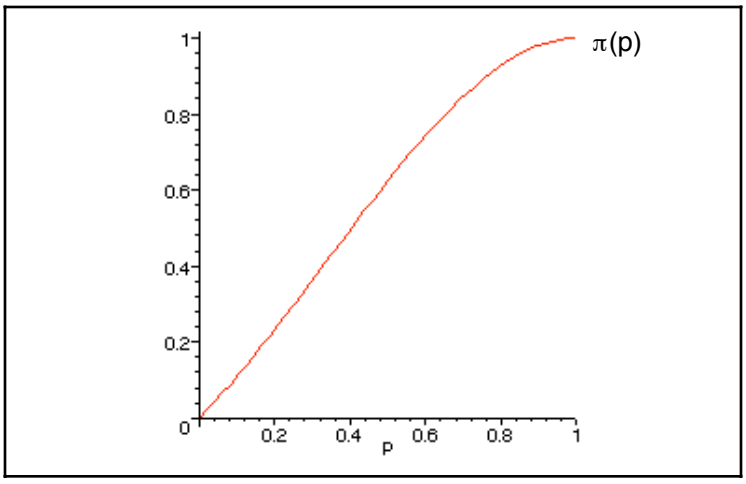
**Answer:** Let $p$ be the probability of a defective item. We want to calculate the power of the test at the null hypothesis. The power function of the test is given by

$$\pi(p) = \begin{cases} P\,(\text{Type I Error}) & \text{if } p \leq 0.1 \\ \\ 1 - P\,(\text{Type II Error}) & \text{if } p > 0.1. \end{cases}$$

Hence, we have

$$\begin{aligned}
&\pi(p)\\
&= P\,(\text{Reject } \mathrm{H_o} \;/\; \mathrm{H_o} \text{ is true})\\
&= P\,(\text{Reject } \mathrm{H_o} \;/\; \mathrm{H_o} : \mathrm{p} = \mathrm{p})\\
&= P\,(\text{first two items are both defective } /\mathrm{p}) +\\
&\quad + P\,(\text{at least one of the first two items is not defective and third is}/\mathrm{p})\\
&= p^2 + (1-p)^2\, p + \binom{2}{1} p(1-p)p\\
&= p + p^2 - p^3.
\end{aligned}$$

The graph of this power function is shown below.




**Remark 18.4.** If $X$ denotes the number of independent trials needed to obtain the first success, then $X \sim GEO(p)$, and

$$P(X = k) = (1-p)^{k-1}\, p,$$

where $k = 1, 2, 3, ..., \infty$. Further

$$P(X \leq n) = 1 - (1-p)^n$$

since

$$\begin{aligned}
\sum_{k=1}^{n} (1-p)^{k-1}\, p &= p \sum_{k=1}^{n} (1-p)^{k-1}\\
&= p\, \frac{1-(1-p)^n}{1-(1-p)}\\
&= 1 - (1-p)^n.
\end{aligned}$$

**Example 18.10.** Let $X$ be the number of independent trails required to obtain a success where $p$ is the probability of success on each trial. The hypothesis $H_o : p = 0.1$ is to be tested against the alternative $H_a : p = 0.3$. The hypothesis is rejected if $X \leq 4$. What is the power of the test if $H_a$ is true?

**Answer:** The power function is given by

$$\pi(p) = \begin{cases} P\,(\text{Type I Error}) & \text{if } p = 0.1 \\ 1 - P\,(\text{Type II Error}) & \text{if p = 0.3.} \end{cases}$$

Hence, we have

$$\begin{aligned} \alpha &= 1 - P\,(\text{Accept H}_\text{o} \text{ / H}_\text{o} \text{ is false}) \\ &= P\,(\text{Reject H}_\text{o} \text{ / H}_\text{a} \text{ is true}) \\ &= P\,(X \leq 4 \text{ / H}_\text{a} \text{ is true}) \\ &= P\,(X \leq 4 \text{ / } p = 0.3) \\ &= \sum_{k=1}^{4} P\,(X = k \text{ /} p = 0.3) \\ &= \sum_{k=1}^{4} (1-p)^{k-1}\, p \qquad (\text{where p} = 0.3) \\ &= \sum_{k=1}^{4} (0.7)^{k-1}\,(0.3) \\ &= 0.3 \sum_{k=1}^{4} (0.7)^{k-1} \\ &= 1 - (0.7)^4 \\ &= 0.7599. \end{aligned}$$

Hence, the power of the test at the alternative is 0.7599.

**Example 18.11.** Let $X_1, X_2, ..., X_{25}$ be a random sample of size 25 drawn from a normal distribution with unknown mean $\mu$ and variance $\sigma^2 = 100$. It is desired to test the null hypothesis $\mu = 4$ against the alternative $\mu = 6$. What is the power at $\mu = 6$ of the test with rejection rule: reject $\mu = 4$ if $\sum_{i=1}^{25} X_i \geq 125$?

**Answer:** The power of the test at the alternative is

$$\begin{aligned}
\pi(6) &= 1 - P\,(\text{Type II Error}) \\
&= 1 - P\,(\text{Accept } \mathrm{H_o} \ / \ \mathrm{H_o} \text{ is false}) \\
&= P\,(\text{Reject } \mathrm{H_o} \ / \ \mathrm{H_a} \text{ is true}) \\
&= P\left(\sum_{i=1}^{25} X_i \geq 125 \ / \ H_a : \mu = 6\right) \\
&= P\left(\overline{X} \geq 5 \ / \ H_a \mu = 6\right) \\
&= P\left(\frac{\overline{X} - 6}{\frac{10}{\sqrt{25}}} \geq \frac{5-6}{\frac{10}{\sqrt{25}}}\right) \\
&= P\left(Z \geq -\frac{1}{2}\right) \\
&= 0.6915.
\end{aligned}$$

**Example 18.12.** A urn contains 7 balls, $\theta$ of which are red. A sample of size 2 is drawn without replacement to test $H_o : \theta \leq 1$ against $H_a : \theta > 1$. If the null hypothesis is rejected if one or more red balls are drawn, find the power of the test when $\theta = 2$.

**Answer:** The power of the test at $\theta = 2$ is given by

$$\begin{aligned}
\pi(2) &= 1 - P\,(\text{Type II Error}) \\
&= 1 - P\,(\text{Accept } \mathrm{H_o} \ / \ \mathrm{H_o} \text{ is false}) \\
&= 1 - P\,(\text{zero red balls are drawn } /2 \text{ balls were red}) \\
&= 1 - \frac{\binom{5}{2}}{\binom{7}{2}} \\
&= 1 - \frac{10}{21} \\
&= \frac{11}{21} \\
&= 0.524.
\end{aligned}$$

In all of these examples, we have seen that if the rule for rejection of the null hypothesis $H_o$ is given, then one can compute the significance level or power function of the hypothesis test. The rejection rule is given in terms of a statistic $W\,(X_1, X_2, ..., X_n)$ of the sample $X_1, X_2, ..., X_n$. For instance, in Example 18.5, the rejection rule was: "Reject the null hypothesis $H_o$ if $\sum_{i=1}^{20} X_i \leq 6$." Similarly, in Example 18.7, the rejection rule was: "Reject $H_o$

if and only if $X_1+X_2+X_3+X_4 < -20$", and so on. The statistic $W$, used in the statement of the rejection rule, partitioned the set $S^n$ into two subsets, where $S$ denotes the support of the density function of the population $X$. One subset is called the rejection or critical region and other subset is called the acceptance region. The rejection rule is obtained in such a way that the probability of the type I error is as small as possible and the power of the test at the alternative is as large as possible.

Next, we give two definitions that will lead us to the definition of uniformly most powerful test.

**Definition 18.9.** Given $0 \leq \delta \leq 1$, a test (or test procedure) $T$ for testing the null hypothesis $H_o : \theta \in \Omega_o$ against the alternative $H_a : \theta \in \Omega_a$ is said to be a *test of level* $\delta$ if

$$\max_{\theta \in \Omega_o} \pi(\theta) \leq \delta,$$

where $\pi(\theta)$ denotes the power function of the test $T$.

**Definition 18.10.** Given $0 \leq \delta \leq 1$, a test (or test procedure) for testing the null hypothesis $H_o : \theta \in \Omega_o$ against the alternative $H_a : \theta \in \Omega_a$ is said to be a *test of size* $\delta$ if

$$\max_{\theta \in \Omega_o} \pi(\theta) = \delta.$$

**Definition 18.11.** Let $T$ be a test procedure for testing the null hypothesis $H_o : \theta \in \Omega_o$ against the alternative $H_a : \theta \in \Omega_a$. The test (or test procedure) $T$ is said to be the *uniformly most powerful* (UMP) test of level $\delta$ if $T$ is of level $\delta$ and for any other test $W$ of level $\delta$,

$$\pi_T(\theta) \geq \pi_W(\theta)$$

for all $\theta \in \Omega_a$. Here $\pi_T(\theta)$ and $\pi_W(\theta)$ denote the power functions of tests $T$ and $W$, respectively.

**Remark 18.5.** If $T$ is a test procedure for testing $H_o : \theta = \theta_o$ against $H_a : \theta = \theta_a$ based on a sample data $x_1, ..., x_n$ from a population $X$ with a continuous probability density function $f(x;\theta)$, then there is a critical region $C$ associated with the the test procedure $T$, and power function of $T$ can be computed as

$$\pi_T = \int_C L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

Similarly, the size of a critical region $C$, say $\alpha$, can be given by

$$\alpha = \int_C L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

The following famous result tells us which tests are uniformly most powerful if the null hypothesis and the alternative hypothesis are both simple.

**Theorem 18.1 (Neyman-Pearson).** Let $X_1, X_2, ..., X_n$ be a random sample from a population with probability density function $f(x;\theta)$. Let

$$L(\theta, x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

be the likelihood function of the sample. Then any critical region $C$ of the form

$$C = \left\{ (x_1, x_2, ..., x_n) \;\middle|\; \frac{L(\theta_o, x_1, ..., x_n)}{L(\theta_a, x_1, ..., x_n)} \leq k \right\}$$

for some constant $0 \leq k < \infty$ is best (or uniformly most powerful) of its size for testing $H_o : \theta = \theta_o$ against $H_a : \theta = \theta_a$.

**Proof:** We assume that the population has a continuous probability density function. If the population has a discrete distribution, the proof can be appropriately modified by replacing integration by summation.

Let $C$ be the critical region of size $\alpha$ as described in the statement of the theorem. Let $B$ be any other critical region of size $\alpha$. We want to show that the power of $C$ is greater than or equal to that of $B$. In view of Remark 18.5, we would like to show that

$$\int_C L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \geq \int_B L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n. \qquad (1)$$

Since $C$ and $B$ are both critical regions of size $\alpha$, we have

$$\int_C L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n = \int_B L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n. \qquad (2)$$

The last equality (2) can be written as

$$\begin{aligned}
&\int_{C\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C\cap B^c} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
&= \int_{C\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C^c\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n
\end{aligned}$$

since

$$C = (C \cap B) \cup (C \cap B^c) \quad \text{and} \quad B = (C \cap B) \cup (C^c \cap B). \tag{3}$$

Therefore from the last equality, we have

$$\int_{C\cap B^c} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n = \int_{C^c\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n. \tag{4}$$

Since

$$C = \left\{ (x_1, x_2, ..., x_n) \;\middle|\; \frac{L(\theta_o, x_1, ..., x_n)}{L(\theta_a, x_1, ..., x_n)} \le k \right\} \tag{5}$$

we have

$$L(\theta_a, x_1, ..., x_n) \ge \frac{L(\theta_o, x_1, ..., x_n)}{k} \tag{6}$$

on $C$, and

$$L(\theta_a, x_1, ..., x_n) < \frac{L(\theta_o, x_1, ..., x_n)}{k} \tag{7}$$

on $C^c$. Therefore from (4), (6) and (7), we have

$$\begin{aligned}
&\int_{C\cap B^c} L(\theta_a, x_1,..., x_n)\, dx_1 \cdots dx_n \\
&\qquad \ge \int_{C\cap B^c} \frac{L(\theta_o, x_1, ..., x_n)}{k}\, dx_1 \cdots dx_n \\
&\qquad = \int_{C^c\cap B} \frac{L(\theta_o, x_1, ..., x_n)}{k}\, dx_1 \cdots dx_n \\
&\qquad \ge \int_{C^c\cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n.
\end{aligned}$$

Thus, we obtain

$$\int_{C\cap B^c} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \ge \int_{C^c\cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

From (3) and the last inequality, we see that

$$\begin{aligned}
&\int_C L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
&\quad = \int_{C\cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C\cap B^c} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
&\quad \ge \int_{C\cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C^c\cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
&\quad \ge \int_B L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n
\end{aligned}$$

and hence the theorem is proved.

Now we give several examples to illustrate the use of this theorem.

**Example 18.13.** Let $X$ be a random variable with a density function $f(x)$. What is the critical region for the best test of

$$H_o : f(x) = \begin{cases} \frac{1}{2} & \text{if } -1 < x < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

against

$$H_a : f(x) = \begin{cases} 1 - |x| & \text{if } -1 < x < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

at the significance size $\alpha = 0.10$?

**Answer:** We assume that the test is performed with a sample of size 1. Using Neyman-Pearson Theorem, the best critical region for the best test at the significance size $\alpha$ is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \mid \frac{L_o(x)}{L_a(x)} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{\frac{1}{2}}{1 - |x|} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid |x| \leq 1 - \frac{1}{2k} \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{1}{2k} - 1 \leq x \leq 1 - \frac{1}{2k} \right\}. \end{aligned}$$

Since

$$\begin{aligned} 0.1 &= P(C) \\ &= P\left( \frac{L_o(X)}{L_a(X)} \leq k \ / \ H_o \text{ is true} \right) \\ &= P\left( \frac{\frac{1}{2}}{1 - |X|} \leq k \ / \ H_o \text{ is true} \right) \\ &= P\left( \frac{1}{2k} - 1 \leq X \leq 1 - \frac{1}{2k} \ / \ H_o \text{ is true} \right), \\ &= \int_{\frac{1}{2k} - 1}^{1 - \frac{1}{2k}} \frac{1}{2} \, dx \\ &= 1 - \frac{1}{2k}, \end{aligned}$$

we get the critical region $C$ to be

$$C = \{ x \in \mathbb{R} \mid -0.1 \leq x \leq 0.1 \}.$$

Thus the best critical region is $C = [-0.1,\ 0.1]$ and the best test is: "Reject $H_o$ if $-0.1 \leq X \leq 0.1$".

**Example 18.14.** Suppose $X$ has the density function

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Based on a single observed value of $X$, find the most powerful critical region of size $\alpha = 0.1$ for testing $H_o : \theta = 1$ against $H_a : \theta = 2$.

**Answer:** By Neyman-Pearson Theorem, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \mid \frac{L(\theta_o, x)}{L(\theta_a, x)} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{(1+\theta_o)\, x^{\theta_o}}{(1+\theta_a)\, x^{\theta_a}} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{2x}{3x^2} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{1}{x} \leq \frac{3}{2}k \right\} \\ &= \{ x \in \mathbb{R} \mid x \geq a, \} \end{aligned}$$

where $a$ is some constant. Hence the most powerful or best test is of the form: "Reject $H_o$ if $X \geq a$."

Since, the significance level of the test is given to be $\alpha = 0.1$, the constant $a$ can be determined. Now we proceed to find $a$. Since

$$\begin{aligned} 0.1 = \alpha &= P\,(\text{Reject } H_o \ / \ H_o \text{ is true}\} \\ &= P\,(X \geq a \ / \ \theta = 1) \\ &= \int_a^1 2x\, dx \\ &= 1 - a^2, \end{aligned}$$

hence

$$a^2 = 1 - 0.1 = 0.9.$$

Therefore

$$a = \sqrt{0.9},$$

since $k$ in Neyman-Pearson Theorem is positive. Hence, the most powerful test is given by "Reject $H_o$ if $X \geq \sqrt{0.9}$".

**Example 18.15.** Suppose that $X$ is a random variable about which the hypothesis $H_o : X \sim UNIF(0,1)$ against $H_a : X \sim N(0,1)$ is to be tested. What is the most powerful test with a significance level $\alpha = 0.05$ based on one observation of $X$?

**Answer:** By Neyman-Pearson Theorem, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \mid \frac{L_o(x)}{L_a(x)} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \sqrt{2\pi}\, e^{\frac{1}{2}x^2} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid x^2 \leq 2 \ln\left( \frac{k}{\sqrt{2\pi}} \right) \right\} \\ &= \{ x \in \mathbb{R} \mid x \leq a, \} \end{aligned}$$

where $a$ is some constant. Hence the most powerful or best test is of the form: "Reject $H_o$ if $X \leq a$."

Since, the significance level of the test is given to be $\alpha = 0.05$, the constant $a$ can be determined. Now we proceed to find $a$. Since

$$\begin{aligned} 0.05 = \alpha &= P(\text{Reject } H_o \ / \ H_o \text{ is true}\} \\ &= P(X \leq a \ / \ X \sim UNIF(0,1)) \\ &= \int_0^a dx \\ &= a, \end{aligned}$$

hence $a = 0.05$. Thus, the most powerful critical region is given by

$$C = \{ x \in \mathbb{R} \mid 0 < x \leq 0.05 \}$$

based on the support of the uniform distribution on the open interval $(0,1)$. Since the support of this uniform distribution is the interval $(0,1)$, the acceptance region (or the complement of $C$ in $(0,1)$) is

$$C^c = \{ x \in \mathbb{R} \mid 0.05 < x < 1 \}.$$

However, since the support of the standard normal distribution is $\mathbb{R}$, the actual critical region should be the complement of $C^c$ in $\mathbb{R}$. Therefore, the critical region of this hypothesis test is the set

$$\{x \in \mathbb{R} \,|\, x \leq 0.05 \text{ or } x \geq 1\}.$$

The most powerful test for $\alpha = 0.05$ is: "Reject $H_o$ if $X \leq 0.05$ or $X \geq 1$."

**Example 18.16.** Let $X_1, X_2, X_3$ denote three independent observations from a distribution with density

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^\theta & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the form of the best critical region of size 0.034 for testing $H_o : \theta = 1$ versus $H_a : \theta = 2$?

**Answer:** By Neyman-Pearson Theorem, the form of the critical region is given by (with $\theta_o = 1$ and $\theta_a = 2$)

$$\begin{aligned} C &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{L(\theta_o, x_1, x_2, x_3)}{L(\theta_a, x_1, x_2, x_3)} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{(1+\theta_o)^3 \prod_{i=1}^3 x_i^{\theta_o}}{(1+\theta_a)^3 \prod_{i=1}^3 x_i^{\theta_a}} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{8 x_1 x_2 x_3}{27 x_1^2 x_2^2 x_3^2} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{1}{x_1 x_2 x_3} \leq \frac{27}{8} k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1 x_2 x_3 \geq a, \right\} \end{aligned}$$

where $a$ is some constant. Hence the most powerful or best test is of the form: "Reject $H_o$ if $\displaystyle\prod_{i=1}^{3} X_i \geq a$."

Since, the significance level of the test is given to be $\alpha = 0.034$, the constant $a$ can be determined. To evaluate the constant $a$, we need the probability distribution of $X_1 X_2 X_3$. The distribution of $X_1 X_2 X_3$ is not easy to get. Hence, we will use Theorem 17.5. There, we have shown that

$-2(1+\theta)\sum_{i=1}^{3}\ln X_i \sim \chi^2(6)$. Now we proceed to find $a$. Since

$$\begin{aligned}
0.034 &= \alpha \\
&= P\,(\text{Reject } H_o \;/\; H_o \text{ is true}\} \\
&= P\,(X_1X_2X_3 \geq a \;/\; \theta = 1) \\
&= P\,(\ln(X_1X_2X_3) \geq \ln a \;/\; \theta = 1) \\
&= P\,(-2(1+\theta)\ln(X_1X_2X_3) \leq -2(1+\theta)\ln a \;/\; \theta = 1) \\
&= P\,(-4\ln(X_1X_2X_3) \leq -4\ln a) \\
&= P\left(\chi^2(6) \leq -4\ln a\right)
\end{aligned}$$

hence from chi-square table, we get

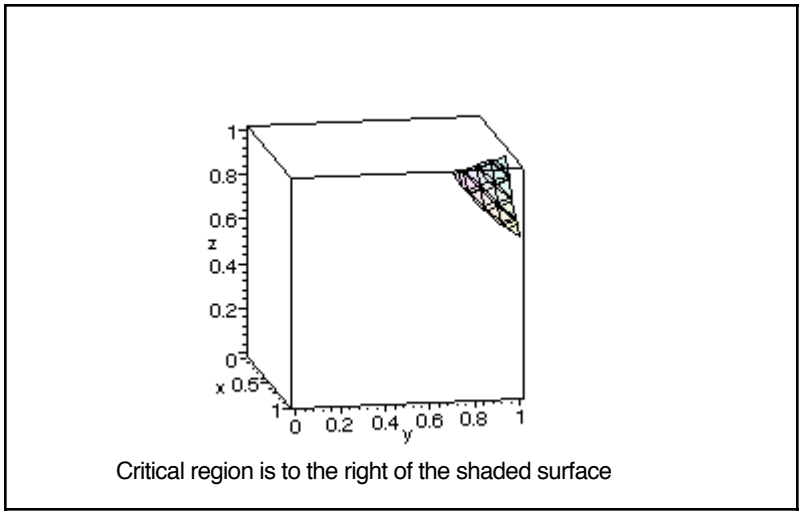
$$-4\ln a = 1.4.$$

Therefore

$$a = e^{-0.35} = 0.7047.$$

Hence, the most powerful test is given by "Reject $H_o$ if $X_1X_2X_3 \geq 0.7047$".

The critical region $C$ is the region above the surface $x_1x_2x_3 = 0.7047$ of the unit cube $[0,1]^3$. The following figure illustrates this region.




Critical region is to the right of the shaded surface

**Example 18.17.** Let $X_1, X_2, ..., X_{12}$ be a random sample from a normal population with mean zero and variance $\sigma^2$. What is the most powerful test of size 0.025 for testing the null hypothesis $H_o : \sigma^2 = 10$ versus $H_a : \sigma^2 = 5$?

**Answer:** By Neyman-Pearson Theorem, the form of the critical region is given by (with ${\sigma_o}^2 = 10$ and ${\sigma_a}^2 = 5$)

$$\begin{aligned}
C &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \frac{L\left({\sigma_o}^2, x_1, x_2, ..., x_{12}\right)}{L\left({\sigma_a}^2, x_1, x_2, ..., x_{12}\right)} \leq k \right\} \\
&= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \prod_{i=1}^{12} \frac{\frac{1}{\sqrt{2\pi\sigma_o^2}} e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_o}\right)^2}}{\frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_a}\right)^2}} \leq k \right\} \\
&= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \left(\frac{1}{2}\right)^6 e^{\frac{1}{20}\sum_{i=1}^{12} x_i^2} \leq k \right\} \\
&= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \sum_{i=1}^{12} x_i^2 \leq a \right\},
\end{aligned}$$

where $a$ is some constant. Hence the most powerful or best test is of the form: "Reject $H_o$ if $\sum_{i=1}^{12} X_i^2 \leq a$."

Since, the significance level of the test is given to be $\alpha = 0.025$, the constant $a$ can be determined. To evaluate the constant $a$, we need the probability distribution of $X_1^2 + X_2^2 + \cdots + X_{12}^2$. It can be shown that the distribution of $\sum_{i=1}^{12} \left(\frac{X_i}{\sigma}\right)^2 \sim \chi^2(12)$. Now we proceed to find $a$. Since

$$\begin{aligned}
0.025 = \alpha &= P\left(\text{Reject } H_o \;/\; H_o \text{ is true}\right\} \\
&= P\left(\sum_{i=1}^{12} \left(\frac{X_i}{\sigma}\right)^2 \leq a \;/\; \sigma^2 = 10\right) \\
&= P\left(\sum_{i=1}^{12} \left(\frac{X_i}{\sqrt{10}}\right)^2 \leq a \;/\; \sigma^2 = 10\right) \\
&= P\left(\chi^2(12) \leq \frac{a}{10}\right),
\end{aligned}$$

hence from chi-square table, we get

$$\frac{a}{10} = 4.4.$$

Therefore

$$a = 44.$$

Hence, the most powerful test is given by "Reject $H_o$ if $\sum_{i=1}^{12} X_i^2 \leq 44$." The best critical region of size 0.025 is given by

$$C = \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \mid \sum_{i=1}^{12} x_i^2 \leq 44 \right\}.$$

In last five examples, we have found the most powerful tests and corresponding critical regions when the both $H_o$ and $H_a$ are simple hypotheses. If either $H_o$ or $H_a$ is not simple, then it is not always possible to find the most powerful test and corresponding critical region. In this situation, hypothesis test is found by using the likelihood ratio. A test obtained by using likelihood ratio is called the *likelihood ratio test* and the corresponding critical region is called the *likelihood ratio critical region*.

### 18.4. Some Examples of Likelihood Ratio Tests

In this section, we illustrate, using likelihood ratio, how one can construct hypothesis test when one of the hypotheses is not simple. As pointed out earlier, the test we will construct using the likelihood ratio is not the most powerful test. However, such a test has all the desirable properties of a hypothesis test. To construct the test one has to follow a sequence of steps. These steps are outlined below:

(1) Find the likelihood function $L(\theta, x_1, x_2, ..., x_n)$ for the given sample.

(2) Evaluate $\max_{\theta \in \Omega_o} L(\theta, x_1, x_2, ..., x_n)$.

(3) Find the maximum likelihood estimator $\widehat{\theta}$ of $\theta$.

(4) Compute $\max_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)$ using $L\left(\widehat{\theta}, x_1, x_2, ..., x_n\right)$.

(5) Using steps (2) and (4), find $W(x_1, ..., x_n) = \dfrac{\max_{\theta \in \Omega_o} L(\theta, x_1, x_2, ..., x_n)}{\max_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)}$.

(6) Using step (5) determine $C = \{(x_1, x_2, ..., x_n) \,|\, W(x_1, ..., x_n) \leq k\}$, where $k \in [0, 1]$.

(7) Reduce $W(x_1, ..., x_n) \leq k$ to an equivalent inequality $\widehat{W}(x_1, ..., x_n) \leq A$.

(8) Determine the distribution of $\widehat{W}(x_1, ..., x_n)$.

(9) Find $A$ such that given $\alpha$ equals $P\left(\widehat{W}(x_1, ..., x_n) \leq A \,|\, H_o \text{ is true}\right)$.

In the remaining examples, for notational simplicity, we will denote the likelihood function $L(\theta, x_1, x_2, ..., x_n)$ simply as $L(\theta)$.

**Example 18.19.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and `known variance` $\sigma^2$. What is the likelihood ratio test of size $\alpha$ for testing the null hypothesis $H_o : \mu = \mu_o$ versus the alternative hypothesis $H_a : \mu \neq \mu_o$?

**Answer:** The likelihood function of the sample is given by

$$L(\mu) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Since $\Omega_o = \{\mu_o\}$, we obtain

$$\max_{\mu\in\Omega_o} L(\mu) = L(\mu_o)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_o)^2}.$$

We have seen in Example 15.13 that if $X \sim N(\mu, \sigma^2)$, then the maximum likelihood estimator of $\mu$ is $\overline{X}$, that is

$$\widehat{\mu} = \overline{X}.$$

Hence

$$\max_{\mu\in\Omega} L(\mu) = L(\widehat{\mu}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

Now the likelihood ratio statistics $W(x_1, x_2, ..., x_n)$ is given by

$$W(x_1, x_2, ..., x_n) = \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_o)^2}}{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

which simplifies to

$$W(x_1, x_2, ..., x_n) = e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_o)^2}.$$

Now the inequality $W(x_1, x_2, ..., x_n) \leq k$ becomes

$$e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_o)^2} \leq k$$

and which can be rewritten as

$$(\overline{x} - \mu_o)^2 \geq -\frac{2\sigma^2}{n} \ln(k)$$

or

$$|\overline{x} - \mu_o| \geq K$$

where $K = \sqrt{-\frac{2\sigma^2}{n} \ln(k)}$. In view of the above inequality, the critical region can be described as

$$C = \{(x_1, x_2, ..., x_n) \ | \ |\overline{x} - \mu_o| \geq K \}.$$

Since we are given the size of the critical region to be $\alpha$, we can determine the constant $K$. Since the size of the critical region is $\alpha$, we have

$$\alpha = P\left(\left|\overline{X} - \mu_o\right| \geq K\right).$$

For finding $K$, we need the probability density function of the statistic $\overline{X} - \mu_o$ when the population $X$ is $N(\mu, \sigma^2)$ and the null hypothesis $H_o : \mu = \mu_o$ is true. Since $\sigma^2$ is known and $X_i \sim N(\mu, \sigma^2)$,

$$\frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

and

$$\begin{aligned}
\alpha &= P\left(\left|\overline{X} - \mu_o\right| \geq K\right) \\
&= P\left(\left|\frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}}\right| \geq K\frac{\sqrt{n}}{\sigma}\right) \\
&= P\left(|Z| \geq K\frac{\sqrt{n}}{\sigma}\right) \qquad \text{where} \qquad Z = \frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \\
&= 1 - P\left(-K\,\frac{\sqrt{n}}{\sigma} \leq Z \leq K\,\frac{\sqrt{n}}{\sigma}\right)
\end{aligned}$$

we get

$$z_{\frac{\alpha}{2}} = K\,\frac{\sqrt{n}}{\sigma}$$

which is

$$K = z_{\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}},$$

where $z_{\frac{\alpha}{2}}$ is a real number such that the integral of the standard normal density from $z_{\frac{\alpha}{2}}$ to $\infty$ equals $\frac{\alpha}{2}$.

Hence, the likelihood ratio test is given by "Reject $H_o$ if

$$\left|\overline{X} - \mu_o\right| \geq z_{\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}."$$

If we denote

$$z = \frac{\overline{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$$

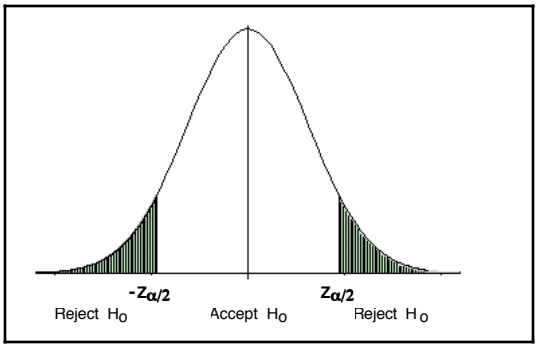
then the above inequality becomes

$$|Z| \geq z_{\frac{\alpha}{2}}.$$

Thus critical region is given by

$$C = \left\{(x_1, x_2, ..., x_n) \mid \ |z| \geq z_{\frac{\alpha}{2}}\right\}.$$

This tells us that the null hypothesis must be rejected when the absolute value of $z$ takes on a value greater than or equal to $z_{\frac{\alpha}{2}}$.




**Remark 18.6.** The hypothesis $H_a : \mu \neq \mu_o$ is called a two-sided alternative hypothesis. An alternative hypothesis of the form $H_a : \mu > \mu_o$ is called a right-sided alternative. Similarly, $H_a : \mu < \mu_o$ is called the a left-sided

alternative. In the above example, if we had a right-sided alternative, that is $H_a : \mu > \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \ z \geq z_\alpha \}.$$

Similarly, if the alternative would have been left-sided, that is $H_a : \mu < \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \ z \leq -z_\alpha \}.$$

We summarize the three cases of hypotheses test of the mean (of the normal population with known variance) in the following table.

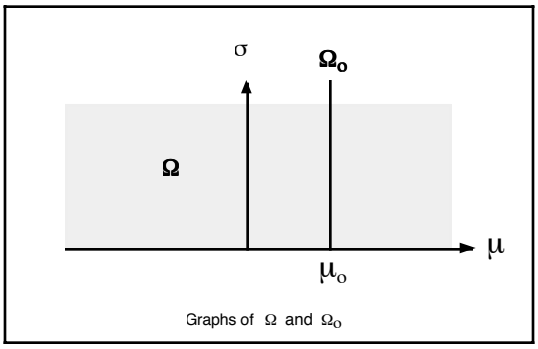
| $H_o$ | $H_a$ | Critical Region (or Test) |
|---|---|---|
| $\mu = \mu_o$ | $\mu > \mu_o$ | $z = \frac{\overline{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha$ |
| $\mu = \mu_o$ | $\mu < \mu_o$ | $z = \frac{\overline{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \leq -z_\alpha$ |
| $\mu = \mu_o$ | $\mu \neq \mu_o$ | $\lvert z \rvert = \left\lvert \frac{\overline{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \right\rvert \geq z_{\frac{\alpha}{2}}$ |

**Example 18.20.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and `unknown variance` $\sigma^2$. What is the likelihood ratio test of size $\alpha$ for testing the null hypothesis $H_o : \mu = \mu_o$ versus the alternative hypothesis $H_a : \mu \neq \mu_o$?

**Answer:** In this example,

$$\begin{aligned}
\Omega &= \left\{ \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \mid \ -\infty < \mu < \infty, \ \sigma^2 > 0 \right\}, \\
\Omega_o &= \left\{ \left(\mu_o, \sigma^2\right) \in \mathbb{R}^2 \mid \sigma^2 > 0 \right\}, \\
\Omega_a &= \left\{ \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \mid \mu \neq \mu_o, \ \sigma^2 > 0 \right\}.
\end{aligned}$$

These sets are illustrated below.



Graphs of Ω and $\Omega_o$

The likelihood function is given by

$$\begin{aligned} L\left(\mu,\sigma^2\right) &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Next, we find the maximum of $L\left(\mu,\sigma^2\right)$ on the set $\Omega_o$. Since the set $\Omega_o$ is equal to $\left\{\left(\mu_o,\sigma^2\right)\in\mathbb{R}^2 \mid 0<\sigma<\infty\right\}$, we have

$$\max_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu,\sigma^2\right) = \max_{\sigma^2>0} L\left(\mu_o,\sigma^2\right).$$

Since $L\left(\mu_o,\sigma^2\right)$ and $\ln L\left(\mu_o,\sigma^2\right)$ achieve the maximum at the same $\sigma$ value, we determine the value of $\sigma$ where $\ln L\left(\mu_o,\sigma^2\right)$ achieves the maximum. Taking the natural logarithm of the likelihood function, we get

$$\ln\left(L\left(\mu,\sigma^2\right)\right) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_o)^2.$$

Differentiating $\ln L\left(\mu_o,\sigma^2\right)$ with respect to $\sigma^2$, we get from the last equality

$$\frac{d}{d\sigma^2}\ln\left(L\left(\mu,\sigma^2\right)\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu_o)^2.$$

Setting this derivative to zero and solving for $\sigma$, we obtain

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2}.$$

Thus $\ln\left(L\left(\mu,\sigma^2\right)\right)$ attains maximum at $\sigma=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2}$. Since this value of $\sigma$ is also yield maximum value of $L\left(\mu,\sigma^2\right)$, we have

$$\max_{\sigma^2>0} L\left(\mu_o,\sigma^2\right)=\left(2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2\right)^{-\frac{n}{2}}e^{-\frac{n}{2}}.$$

Next, we determine the maximum of $L\left(\mu,\sigma^2\right)$ on the set $\Omega$. As before, we consider $\ln L\left(\mu,\sigma^2\right)$ to determine where $L\left(\mu,\sigma^2\right)$ achieves maximum. Taking the natural logarithm of $L\left(\mu,\sigma^2\right)$, we obtain

$$\ln\left(L\left(\mu,\sigma^2\right)\right)=-\frac{n}{2}\ln(\sigma^2)-\frac{n}{2}\ln(2\pi)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the partial derivatives of $\ln L\left(\mu,\sigma^2\right)$ first with respect to $\mu$ and then with respect to $\sigma^2$, we get

$$\frac{\partial}{\partial\mu}\ln L\left(\mu,\sigma^2\right)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L\left(\mu,\sigma^2\right)=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,$$

respectively. Setting these partial derivatives to zero and solving for $\mu$ and $\sigma$, we obtain

$$\mu=\overline{x}\qquad\text{and}\qquad\sigma^2=\frac{n-1}{n}\,s^2,$$

where $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$ is the sample variance.

Letting these optimal values of $\mu$ and $\sigma$ into $L\left(\mu,\sigma^2\right)$, we obtain

$$\max_{(\mu,\sigma^2)\in\Omega} L\left(\mu,\sigma^2\right)=\left(2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{-\frac{n}{2}}e^{-\frac{n}{2}}.$$

Hence

$$\frac{\max\limits_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu,\sigma^2\right)}{\max\limits_{(\mu,\sigma^2)\in\Omega} L\left(\mu,\sigma^2\right)}=\frac{\left(2\pi\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\mu_o)^2\right)^{-\frac{n}{2}}e^{-\frac{n}{2}}}{\left(2\pi\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\overline{x})^2\right)^{-\frac{n}{2}}e^{-\frac{n}{2}}}=\left(\frac{\sum\limits_{i=1}^{n}(x_i-\mu_o)^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right)^{-\frac{n}{2}}.$$

Since

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=(n-1)\,s^2$$

and

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n\,(\overline{x}-\mu_o)^2\,,$$

we get

$$W(x_1,x_2,...,x_n)=\frac{\max\limits_{(\mu,\sigma^2)\in\Omega_o}L\left(\mu,\sigma^2\right)}{\max\limits_{(\mu,\sigma^2)\in\Omega}L\left(\mu,\sigma^2\right)}=\left(1+\frac{n}{n-1}\frac{(\overline{x}-\mu_o)^2}{s^2}\right)^{-\frac{n}{2}}.$$

Now the inequality $W(x_1,x_2,...,x_n)\leq k$ becomes

$$\left(1+\frac{n}{n-1}\frac{(\overline{x}-\mu_o)^2}{s^2}\right)^{-\frac{n}{2}}\leq k$$

and which can be rewritten as

$$\left(\frac{\overline{x}-\mu_o}{s}\right)^2\geq\frac{n-1}{n}\left(k^{-\frac{2}{n}}-1\right)$$

or

$$\left|\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}\right|\geq K$$

where $K=\sqrt{(n-1)\left[k^{-\frac{2}{n}}-1\right]}$. In view of the above inequality, the critical region can be described as

$$C=\left\{(x_1,x_2,...,x_n)\mid\left|\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}\right|\geq K\right\}$$

and the best likelihood ratio test is: "Reject $H_o$ if $\left|\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}\right|\geq K$". Since we are given the size of the critical region to be $\alpha$, we can find the constant $K$. For finding $K$, we need the probability density function of the statistic $\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}$ when the population $X$ is $N(\mu,\sigma^2)$ and the null hypothesis $H_o:\mu=\mu_o$ is true.

Since the population is normal with mean $\mu$ and variance $\sigma^2$,

$$\frac{\overline{X}-\mu_o}{\frac{S}{\sqrt{n}}}\sim t(n-1),$$

where $S^2$ is the sample variance and equals to $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$. Hence

$$K = t_{\frac{\alpha}{2}}(n-1)\,\frac{s}{\sqrt{n}},$$

where $t_{\frac{\alpha}{2}}(n-1)$ is a real number such that the integral of the t-distribution with $n-1$ degrees of freedom from $t_{\frac{\alpha}{2}}(n-1)$ to $\infty$ equals $\frac{\alpha}{2}$.

Therefore, the likelihood ratio test is given by "Reject $H_o : \mu = \mu_o$ if

$$\left|\overline{X}-\mu_o\right| \geq t_{\frac{\alpha}{2}}(n-1)\,\frac{S}{\sqrt{n}}."$$

If we denote

$$t = \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}$$

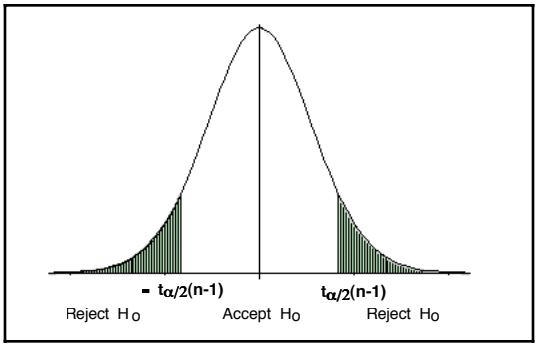
then the above inequality becomes

$$|T| \geq t_{\frac{\alpha}{2}}(n-1).$$

Thus critical region is given by

$$C = \left\{(x_1, x_2, ..., x_n) \mid \ |t| \geq t_{\frac{\alpha}{2}}(n-1)\right\}.$$

This tells us that the null hypothesis must be rejected when the absolute value of $t$ takes on a value greater than or equal to $t_{\frac{\alpha}{2}}(n-1)$.




**Remark 18.7.** In the above example, if we had a right-sided alternative, that is $H_a : \mu > \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \ t \geq t_\alpha(n-1)\}.$$

Similarly, if the alternative would have been left-sided, that is $H_a : \mu < \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \ t \leq -t_\alpha(n-1)\}.$$

We summarize the three cases of hypotheses test of the mean (of the normal population with unknown variance) in the following table.

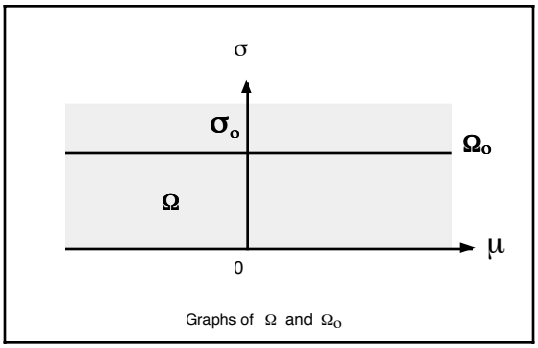
| $H_o$ | $H_a$ | Critical Region (or Test) |
|---|---|---|
| $\mu = \mu_o$ | $\mu > \mu_o$ | $t = \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \geq t_\alpha(n-1)$ |
| $\mu = \mu_o$ | $\mu < \mu_o$ | $t = \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \leq -t_\alpha(n-1)$ |
| $\mu = \mu_o$ | $\mu \neq \mu_o$ | $\lvert t\rvert = \left\lvert \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \right\rvert \geq t_{\frac{\alpha}{2}}(n-1)$ |

**Example 18.21.** Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2$. What is the likelihood ratio test of significance of size $\alpha$ for testing the null hypothesis $H_o : \sigma^2 = \sigma_o^2$ versus $H_a : \sigma^2 \neq \sigma_o^2$?

**Answer:** In this example,

$$\Omega = \left\{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \ -\infty < \mu < \infty, \ \sigma^2 > 0\right\},$$
$$\Omega_o = \left\{(\mu, \sigma_o^2) \in \mathbb{R}^2 \mid \ -\infty < \mu < \infty\right\},$$
$$\Omega_a = \left\{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \ -\infty < \mu < \infty, \ \sigma \neq \sigma_o\right\}.$$

These sets are illustrated below.



Graphs of Ω and $\Omega_o$

The likelihood function is given by

$$\begin{aligned} L\left(\mu,\sigma^2\right) &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Next, we find the maximum of $L\left(\mu,\sigma^2\right)$ on the set $\Omega_o$. Since the set $\Omega_o$ is equal to $\left\{\left(\mu,\sigma_o^2\right)\in\mathbb{R}^2 \mid -\infty<\mu<\infty\right\}$, we have

$$\max_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu,\sigma^2\right) = \max_{-\infty<\mu<\infty} L\left(\mu,\sigma_o^2\right).$$

Since $L\left(\mu,\sigma_o^2\right)$ and $\ln L\left(\mu,\sigma_o^2\right)$ achieve the maximum at the same $\mu$ value, we determine the value of $\mu$ where $\ln L\left(\mu,\sigma_o^2\right)$ achieves the maximum. Taking the natural logarithm of the likelihood function, we get

$$\ln\left(L\left(\mu,\sigma_o^2\right)\right) = -\frac{n}{2}\ln(\sigma_o^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma_o^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Differentiating $\ln L\left(\mu,\sigma_o^2\right)$ with respect to $\mu$, we get from the last equality

$$\frac{d}{d\mu}\ln\left(L\left(\mu,\sigma^2\right)\right) = \frac{1}{\sigma_o^2}\sum_{i=1}^{n}(x_i-\mu).$$

Setting this derivative to zero and solving for $\mu$, we obtain

$$\mu = \overline{x}.$$

Hence, we obtain

$$\max_{-\infty<\mu<\infty} L\left(\mu,\sigma^2\right) = \left(\frac{1}{2\pi\sigma_o^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_o^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

Next, we determine the maximum of $L\left(\mu,\sigma^2\right)$ on the set $\Omega$. As before, we consider $\ln L\left(\mu,\sigma^2\right)$ to determine where $L\left(\mu,\sigma^2\right)$ achieves maximum. Taking the natural logarithm of $L\left(\mu,\sigma^2\right)$, we obtain

$$\ln\left(L\left(\mu,\sigma^2\right)\right) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the partial derivatives of $\ln L\left(\mu,\sigma^2\right)$ first with respect to $\mu$ and then with respect to $\sigma^2$, we get

$$\frac{\partial}{\partial\mu}\ln L\left(\mu,\sigma^2\right)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L\left(\mu,\sigma^2\right)=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,$$

respectively. Setting these partial derivatives to zero and solving for $\mu$ and $\sigma$, we obtain

$$\mu=\overline{x}\qquad\text{and}\qquad\sigma^2=\frac{n-1}{n}\ s^2,$$

where $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$ is the sample variance.

Letting these optimal values of $\mu$ and $\sigma$ into $L\left(\mu,\sigma^2\right)$, we obtain

$$\max_{(\mu,\sigma^2)\in\Omega}L\left(\mu,\sigma^2\right)=\left(\frac{n}{2\pi(n-1)s^2}\right)^{\frac{n}{2}}e^{-\frac{n}{2(n-1)s^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

Therefore

$$\begin{aligned}W(x_1,x_2,...,x_n)&=\frac{\max\limits_{(\mu,\sigma^2)\in\Omega_o}L\left(\mu,\sigma^2\right)}{\max\limits_{(\mu,\sigma^2)\in\Omega}L\left(\mu,\sigma^2\right)}\\&=\frac{\left(\frac{1}{2\pi\sigma_o^2}\right)^{\frac{n}{2}}e^{-\frac{1}{2\sigma_o^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}}{\left(\frac{n}{2\pi(n-1)s^2}\right)^{\frac{n}{2}}e^{-\frac{n}{2(n-1)s^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}}\\&=n^{-\frac{n}{2}}\,e^{\frac{n}{2}}\left(\frac{(n-1)s^2}{\sigma_o^2}\right)^{\frac{n}{2}}e^{-\frac{(n-1)s^2}{2\sigma_o^2}}.\end{aligned}$$

Now the inequality $W(x_1,x_2,...,x_n)\le k$ becomes

$$n^{-\frac{n}{2}}\,e^{\frac{n}{2}}\left(\frac{(n-1)s^2}{\sigma_o^2}\right)^{\frac{n}{2}}e^{-\frac{(n-1)s^2}{2\sigma_o^2}}\le k$$

which is equivalent to

$$\left(\frac{(n-1)s^2}{\sigma_o^2}\right)^{n}e^{-\frac{(n-1)s^2}{\sigma_o^2}}\le\left(k\left(\frac{n}{e}\right)^{\frac{n}{2}}\right)^2:=K_o,$$

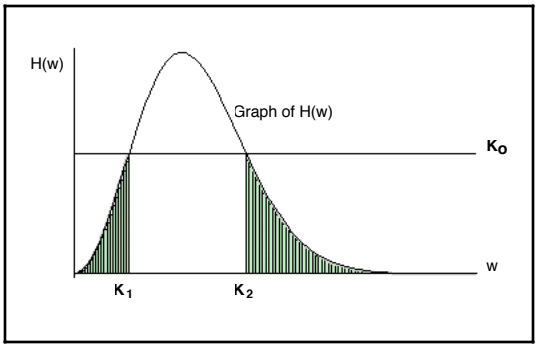
where $K_o$ is a constant. Let $H$ be a function defined by

$$H(w) = w^n\, e^{-w}.$$

Using this, we see that the above inequality becomes

$$H\left(\frac{(n-1)s^2}{\sigma_o^2}\right) \leq K_o.$$

The figure below illustrates this inequality.




From this it follows that

$$\frac{(n-1)s^2}{\sigma_o^2} \leq K_1 \qquad \text{or} \qquad \frac{(n-1)s^2}{\sigma_o^2} \geq K_2.$$

In view of these inequalities, the critical region can be described as

$$C = \left\{(x_1, x_2, ..., x_n) \;\middle|\; \frac{(n-1)s^2}{\sigma_o^2} \leq K_1 \text{ or } \frac{(n-1)s^2}{\sigma_o^2} \geq K_2\right\},$$

and the best likelihood ratio test is: "Reject $H_o$ if

$$\frac{(n-1)S^2}{\sigma_o^2} \leq K_1 \text{ or } \frac{(n-1)S^2}{\sigma_o^2} \geq K_2."$$

Since we are given the size of the critical region to be $\alpha$, we can determine the constants $K_1$ and $K_2$. As the sample $X_1, X_2, ..., X_n$ is taken from a normal distribution with mean $\mu$ and variance $\sigma^2$, we get

$$\frac{(n-1)S^2}{\sigma_o^2} \sim \chi^2(n-1)$$

when the null hypothesis $H_o : \sigma^2 = \sigma_o^2$ is true.

Therefore, the likelihood ratio critical region $C$ becomes

$$\left\{(x_1, x_2, ..., x_n) \;\middle|\; \frac{(n-1)s^2}{\sigma_o^2} \leq \chi^2_{\frac{\alpha}{2}}(n-1) \text{ or } \frac{(n-1)s^2}{\sigma_o^2} \geq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\}$$

and the likelihood ratio test is: "Reject $H_o : \sigma^2 = \sigma_o^2$ if

$$\frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{\frac{\alpha}{2}}(n-1) \text{ or } \frac{(n-1)S^2}{\sigma_o^2} \geq \chi^2_{1-\frac{\alpha}{2}}(n-1)"$$

where $\chi^2_{\frac{\alpha}{2}}(n-1)$ is a real number such that the integral of the chi-square density function with $(n-1)$ degrees of freedom from 0 to $\chi^2_{\frac{\alpha}{2}}(n-1)$ is $\frac{\alpha}{2}$. Further, $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ denotes the real number such that the integral of the chi-square density function with $(n-1)$ degrees of freedom from $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ to $\infty$ is $\frac{\alpha}{2}$.

**Remark 18.8.** We summarize the three cases of hypotheses test of the variance (of the normal population with unknown mean) in the following table.

| $H_o$ | $H_a$ | Critical Region (or Test) |
|---|---|---|
| $\sigma^2 = \sigma_o^2$ | $\sigma^2 > \sigma_o^2$ | $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \geq \chi^2_{1-\alpha}(n-1)$ |
| $\sigma^2 = \sigma_o^2$ | $\sigma^2 < \sigma_o^2$ | $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \leq \chi^2_{\alpha}(n-1)$ |
| $\sigma^2 = \sigma_o^2$ | $\sigma^2 \neq \sigma_o^2$ | $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \geq \chi^2_{1-\alpha/2}(n-1)$ or $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \leq \chi^2_{\alpha/2}(n-1)$ |

## 18.5. Review Exercises

**1.** Five trials $X_1, X_2, ..., X_5$ of a Bernoulli experiment were conducted to test $H_o : p = \frac{1}{2}$ against $H_a : p = \frac{3}{4}$. The null hypothesis $H_o$ will be rejected if $\sum_{i=1}^{5} X_i = 5$. Find the probability of Type I and Type II errors.

**2.** A manufacturer of car batteries claims that the life of his batteries is normally distributed with a standard deviation equal to 0.9 year. If a random

sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

**3.** Let $X_1, X_2, ..., X_8$ be a random sample of size 8 from a Poisson distribution with parameter $\lambda$. Reject the null hypothesis $H_o : \lambda = 0.5$ if the observed sum $\sum_{i=1}^{8} x_i \geq 8$. First, compute the significance level $\alpha$ of the test. Second, find the power function $\beta(\lambda)$ of the test as a sum of Poisson probabilities when $H_a$ is true.

**4.** Suppose $X$ has the density function

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{for } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

If one observation of $X$ is taken, what are the probabilities of Type I and Type II errors in testing the null hypothesis $H_o : \theta = 1$ against the alternative hypothesis $H_a : \theta = 2$, if $H_o$ is rejected for $X > 0.92$.

**5.** Let $X$ have the density function

$$f(x) = \begin{cases} (\theta + 1)\, x^{\theta} & \text{for } 0 < x < 1 \text{ where } \theta > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

The hypothesis $H_o : \theta = 1$ is to be rejected in favor of $H_1 : \theta = 2$ if $X > 0.90$. What is the probability of Type I error?

**6.** Let $X_1, X_2, ..., X_6$ be a random sample from a distribution with density function

$$f(x) = \begin{cases} \theta\, x^{\theta - 1} & \text{for } 0 < x < 1 \text{ where } \theta > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

The null hypothesis $H_o : \theta = 1$ is to be rejected in favor of the alternative $H_a : \theta > 1$ if and only if *at least* 5 of the sample observations are larger than 0.7. What is the significance level of the test?

**7.** A researcher wants to test $H_o : \theta = 0$ versus $H_a : \theta = 1$, where $\theta$ is a parameter of a population of interest. The statistic $W$, based on a random sample of the population, is used to test the hypothesis. Suppose that under $H_o$, $W$ has a normal distribution with mean 0 and variance 1, and under $H_a$, $W$ has a normal distribution with mean 4 and variance 1. If $H_o$ is rejected when $W > 1.50$, then what are the probabilities of a Type I or Type II error respectively?

**8.** Let $X_1$ and $X_2$ be a random sample of size 2 from a normal distribution $N(\mu,\ 1)$. Find the *likelihood ratio critical region* of size 0.005 for testing the null hypothesis $H_o : \mu = 0$ against the composite alternative $H_a : \mu \neq 0$?

**9.** Let $X_1, X_2, ..., X_{10}$ be a random sample from a Poisson distribution with mean $\theta$. What is the most powerful (or best) critical region of size 0.08 for testing the null hypothesis $H_0 : \theta = 0.1$ against $H_a : \theta = 0.5$?

**10.** Let $X$ be a random sample of size 1 from a distribution with probability density function

$$f(x,\theta) = \begin{cases} (1-\frac{\theta}{2}) + \theta\, x & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

For a significance level $\alpha = 0.1$, what is the *best (or uniformly most powerful) critical region* for testing the null hypothesis $H_o : \theta = -1$ against $H_a : \theta = 1$?

**11.** Let $X_1, X_2$ be a random sample of size 2 from a distribution with probability density function

$$f(x,\theta) = \begin{cases} \frac{\theta^x\, e^{-\theta}}{x!} & \text{if } x = 0, 1, 2, 3, .... \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. For a significance level $\alpha = 0.053$, what is the *best critical region* for testing the null hypothesis $H_o : \theta = 1$ against $H_a : \theta = 2$? Sketch the graph of the best critical region.

**12.** Let $X_1, X_2, ..., X_8$ be a random sample of size 8 from a distribution with probability density function

$$f(x,\theta) = \begin{cases} \frac{\theta^x\, e^{-\theta}}{x!} & \text{if } x = 0, 1, 2, 3, .... \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. What is the *likelihood ratio critical region* for testing the null hypothesis $H_o : \theta = 1$ against $H_a : \theta \neq 1$? If $\alpha = 0.1$ can you determine the best likelihood ratio critical region?

**13.** Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with probability density function

$$f(x,\theta) = \begin{cases} \frac{x^6\, e^{-\frac{x}{\beta}}}{\Gamma(7)\beta^7}, & \text{if } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta \geq 0$. What is the *likelihood ratio critical region* for testing the null hypothesis $H_o : \beta = 5$ against $H_a : \beta \neq 5$? What is the *most powerful test ?*

**14.** Let $X_1, X_2, ..., X_5$ denote a random sample of size 5 from a population $X$ with probability density function

$$f(x;\theta) = \begin{cases} (1-\theta)^{x-1}\,\theta & \text{if } x = 1, 2, 3, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is a parameter. What is the *likelihood ratio critical region* of size 0.05 for testing $H_o : \theta = 0.5$ versus $H_a : \theta \neq 0.5$?

**15.** Let $X_1, X_2, X_3$ denote a random sample of size 3 from a population $X$ with probability density function

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2}} \qquad\qquad -\infty < x < \infty,$$

where $-\infty < \mu < \infty$ is a parameter. What is the *likelihood ratio critical region* of size 0.05 for testing $H_o : \mu = 3$ versus $H_a : \mu \neq 3$?

**16.** Let $X_1, X_2, X_3$ denote a random sample of size 3 from a population $X$ with probability density function

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

where $0 < \theta < \infty$ is a parameter. What is the *likelihood ratio critical region* for testing $H_o : \theta = 3$ versus $H_a : \theta \neq 3$?

**17.** Let $X_1, X_2, X_3$ denote a random sample of size 3 from a population $X$ with probability density function

$$f(x;\theta) = \begin{cases} \frac{e^{-\theta}\,\theta^x}{x!} & \text{if } x = 0, 1, 2, 3, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$ is a parameter. What is the *likelihood ratio critical region* for testing $H_o : \theta = 0.1$ versus $H_a : \theta \neq 0.1$?

**18.** A box contains 4 marbles, $\theta$ of which are white and the rest are black. A sample of size 2 is drawn to test $H_o : \theta = 2$ versus $H_a : \theta \neq 2$. If the null

hypothesis is rejected if both marbles are the same color, find the significance level of the test.

**19.** Let $X_1, X_2, X_3$ denote a random sample of size 3 from a population $X$ with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$ is a parameter. What is the *likelihood ratio critical region* of size $\frac{117}{125}$ for testing $H_o : \theta = 5$ versus $H_a : \theta \neq 5$?

**20.** Let $X_1$, $X_2$ and $X_3$ denote three independent observations from a distribution with density

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \beta < \infty$ is a parameter. What is the *best (or uniformly most powerful* critical region for testing $H_o : \beta = 5$ versus $H_a : \beta = 10$?

**21.** Suppose $X$ has the density function

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{for } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

If $X_1, X_2, X_3, X_4$ is a random sample of size 4 taken from $X$, what are the probabilities of Type I and Type II errors in testing the null hypothesis $H_o : \theta = 1$ against the alternative hypothesis $H_a : \theta = 2$, if $H_o$ is rejected for $\max\{X_1, X_2, X_3, X_4\} \le \frac{1}{2}$.

**22.** Let $X_1, X_2, X_3$ denote a random sample of size 3 from a population $X$ with probability density function

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

where $0 < \theta < \infty$ is a parameter. The null hypothesis $H_o : \theta = 3$ is to be rejected in favor of the alternative $H_a : \theta \neq 3$ if and only if $\overline{X} > 6.296$. What is the significance level of the test?

# Chapter 19

# SIMPLE LINEAR REGRESSION AND CORRELATION ANALYSIS

Let $X$ and $Y$ be two random variables with joint probability density function $f(x, y)$. Then the conditional density of $Y$ given that $X = x$ is

$$f(y/x) = \frac{f(x, y)}{g(x)}$$

where

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

is the marginal density of $X$. The conditional mean of $Y$

$$E\,(Y|X = x) = \int_{-\infty}^{\infty} yf(y/x)\, dy$$

is called the regression equation of $Y$ on $X$.

**Example 19.1.** Let $X$ and $Y$ be two random variables with the joint probability density function

$$f(x, y) = \begin{cases} xe^{-x(1+y)} & \text{if } x > 0,\ y > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the regression equation of $Y$ on $X$ and then sketch the regression curve.

**Answer:** The marginal density of $X$ is given by

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} xe^{-x(1+y)}\,dy \\ &= \int_{-\infty}^{\infty} xe^{-x}\,e^{-xy}\,dy \\ &= xe^{-x}\int_{-\infty}^{\infty} e^{-xy}\,dy \\ &= xe^{-x}\left[-\frac{1}{x}\,e^{-xy}\right]_0^{\infty} \\ &= e^{-x}. \end{aligned}$$

The conditional density of $Y$ given $X = x$ is

$$f(y/x) = \frac{f(x,y)}{g(x)} = \frac{xe^{-x(1+y)}}{e^{-x}} = xe^{-xy}, \qquad y > 0.$$

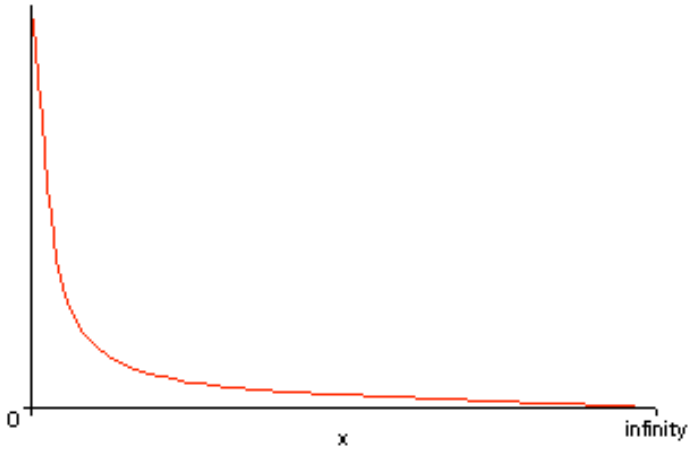
The conditional mean of $Y$ given $X = x$ is given by

$$E(Y/x) = \int_{-\infty}^{\infty} yf(y/x)\,dy = \int_{-\infty}^{\infty} y\,x\,e^{-xy}\,dy = \frac{1}{x}.$$

Thus the regression equation of $Y$ on $X$ is

$$E(Y/x) = \frac{1}{x}, \quad x > 0.$$

The graph of this equation of $Y$ on $X$ is shown below.




Graph of the regression equation E(Y/x) = 1/ x

From this example it is clear that the conditional mean $E(Y/x)$ is a function of $x$. If this function is of the form $\alpha + \beta x$, then the corresponding regression equation is called a linear regression equation; otherwise it is called a nonlinear regression equation. The term linear regression refers to a specification that is linear in the parameters. Thus $E(Y/x) = \alpha + \beta x^2$ is also a linear regression equation. The regression equation $E(Y/x) = \alpha x^\beta$ is an example of a nonlinear regression equation.

The main purpose of regression analysis is to predict $Y_i$ from the knowledge of $x_i$ using the relationship like

$$E(Y_i/x_i) = \alpha + \beta x_i.$$

The $Y_i$ is called the response or dependent variable where as $x_i$ is called the predictor or independent variable. The term regression has an interesting history, dating back to Francis Galton (1822-1911). Galton studied the heights of fathers and sons, in which he observed a regression (a "turning back") from the heights of sons to the heights of their fathers. That is tall fathers tend to have tall sons and short fathers tend to have short sons. However, he also found that very tall fathers tend to have shorter sons and very short fathers tend to have taller sons. Galton called this phenomenon regression towards the mean.

In regression analysis, that is when investigating the relationship between a predictor and response variable, there are two steps to the analysis. The first step is totally data oriented. This step is always performed. The second step is the statistical one, in which we draw conclusions about the (population) regression equation $E(Y_i/x_i)$. Normally the regression equation contains several parameters. There are two well known methods for finding the estimates of the parameters of the regression equation. These two methods are: (1) The least square method and (2) the normal regression method.

### 19.1. The Least Squares Method

Let $\{(x_i, y_i) \,|\, i = 1, 2, ..., n\}$ be a set of data. Assume that

$$E(Y_i/x_i) = \alpha + \beta x_i, \tag{1}$$

that is

$$y_i = \alpha + \beta x_i, \qquad i = 1, 2, ..., n.$$

Then the sum of the squares of the error is given by

$$\mathcal{E}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 . \tag{2}$$

The least squares estimates of $\alpha$ and $\beta$ are defined to be those values which minimize $\mathcal{E}(\alpha, \beta)$. That is,

$$\left(\widehat{\alpha}, \widehat{\beta}\right) = \arg \min_{(\alpha,\beta)} \mathcal{E}(\alpha, \beta).$$

This least squares method is due to Adrien M. Legendre (1752-1833). Note that the least squares method also works even if the regression equation is nonlinear (that is, not of the form (1)).

Next, we give several examples to illustrate the method of least squares.

**Example 19.2.** Given the five pairs of points $(x, y)$ shown in table below

| $x$ | 4 | 0 | $-2$ | 3 | 1 |
|---|---|---|---|---|---|
| $y$ | 5 | 0 | 0 | 6 | 3 |

what is the line of the form $y = x + b$ best fits the data by method of least squares?

**Answer:** Suppose the best fit line is $y = x + b$. Then for each $x_i$, $x_i + b$ is the estimated value of $y_i$. The difference between $y_i$ and the estimated value of $y_i$ is the error or the residual corresponding to the $i^{\text{th}}$ measurement. That is, the error corresponding to the $i^{\text{th}}$ measurement is given by

$$\epsilon_i = y_i - x_i - b.$$

Hence the sum of the squares of the errors is

$$\begin{aligned} \mathcal{E}(b) &= \sum_{i=1}^{5} \epsilon_i^2 \\ &= \sum_{i=1}^{5} (y_i - x_i - b)^2 . \end{aligned}$$

Differentiating $\mathcal{E}(b)$ with respect to $b$, we get

$$\frac{d}{db}\mathcal{E}(b) = 2 \sum_{i=1}^{5} (y_i - x_i - b)\,(-1).$$

Setting $\frac{d}{db}\mathcal{E}(b)$ equal to 0, we get

$$\sum_{i=1}^{5}(y_i - x_i - b) = 0$$

which is

$$5b = \sum_{i=1}^{5} y_i - \sum_{i=1}^{5} x_i.$$

Using the data, we see that

$$5b = 14 - 6$$

which yields $b = \frac{8}{5}$. Hence the best fitted line is

$$y = x + \frac{8}{5}.$$

**Example 19.3.** Suppose the line $y = bx + 1$ is fit by the method of least squares to the 3 data points

| $x$ | 1 | 2 | 4 |
|---|---|---|---|
| $y$ | 2 | 2 | 0 |

What is the value of the constant $b$?

**Answer:** The error corresponding to the $i^{\text{th}}$ measurement is given by

$$\epsilon_i = y_i - bx_i - 1.$$

Hence the sum of the squares of the errors is

$$\begin{aligned}\mathcal{E}(b) &= \sum_{i=1}^{3} \epsilon_i^2 \\ &= \sum_{i=1}^{3} (y_i - bx_i - 1)^2 .\end{aligned}$$

Differentiating $\mathcal{E}(b)$ with respect to $b$, we get

$$\frac{d}{db}\mathcal{E}(b) = 2\sum_{i=1}^{3} (y_i - bx_i - 1)(-x_i).$$

Setting $\frac{d}{db}\mathcal{E}(b)$ equal to 0, we get

$$\sum_{i=1}^{3}(y_i - bx_i - 1)\,x_i = 0$$

which in turn yields

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

Using the given data we see that

$$b = \frac{6-7}{21} = -\frac{1}{21},$$

and the best fitted line is

$$y = -\frac{1}{21}x + 1.$$

**Example 19.4.** Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with

$$E(Y_i/x_i) = \theta + 2\ln x_i$$

where $\theta$ is an unknown parameter and $x_1, x_2, ..., x_n$ are given constants. What is the least square estimate of the parameter $\theta$?

**Answer:** The sum of the squares of errors is

$$\mathcal{E}(\theta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \theta - 2\ln x_i)^2 .$$

Differentiating $\mathcal{E}(\theta)$ with respect to $\theta$, we get

$$\frac{d}{d\theta}\mathcal{E}(\theta) = 2\sum_{i=1}^{n}(y_i - \theta - 2\ln x_i)\,(-1).$$

Setting $\frac{d}{d\theta}\mathcal{E}(\theta)$ equal to 0, we get

$$\sum_{i=1}^{n}(y_i - \theta - 2\ln x_i) = 0$$

which is

$$\theta = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n}\ln x_i\right).$$

Hence the least squares estimate of $\theta$ is $\widehat{\theta} = \overline{y} - \frac{2}{n}\sum_{i=1}^{n} \ln x_i$.

**Example 19.5.** Given the three pairs of points $(x, y)$ shown below:

| $x$ | 4 | 1 | 2 |
|---|---|---|---|
| $y$ | 2 | 1 | 0 |

What is the curve of the form $y = x^{\beta}$ best fits the data by method of least squares?

**Answer:** The sum of the squares of the errors is given by

$$\begin{aligned}\mathcal{E}(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - x_i^{\beta}\right)^2.\end{aligned}$$

Differentiating $\mathcal{E}(\beta)$ with respect to $\beta$, we get

$$\frac{d}{d\beta}\mathcal{E}(\beta) = 2\sum_{i=1}^{n} \left(y_i - x_i^{\beta}\right)\left(-x_i^{\beta} \ln x_i\right)$$

Setting this derivative $\frac{d}{d\beta}\mathcal{E}(\beta)$ to 0, we get

$$\sum_{i=1}^{n} y_i x_i^{\beta} \ln x_i = \sum_{i=1}^{n} x_i^{\beta}\, x_i^{\beta} \ln x_i.$$

Using the given data we obtain

$$(2)\, 4^{\beta} \ln 4 = 4^{2\beta} \ln 4 + 2^{2\beta} \ln 2$$

which simplifies to

$$4 = (2)\, 4^{\beta} + 1$$

or

$$4^{\beta} = \frac{3}{2}.$$

Taking the natural logarithm of both sides of the above expression, we get

$$\beta = \frac{\ln 3 - \ln 2}{\ln 4} = 0.2925$$

Thus the least squares best fit model is $y = x^{0.2925}$.

**Example 19.6.** Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with $E(Y_i/x_i) = \alpha + \beta x_i$, where $\alpha$ and $\beta$ are unknown parameters, and $x_1, x_2, ..., x_n$ are given constants. What are the least squares estimate of the parameters $\alpha$ and $\beta$?

**Answer:** The sum of the squares of the errors is given by

$$\begin{aligned}\mathcal{E}(\alpha, \beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - -\alpha - \beta x_i)^2 .\end{aligned}$$

Differentiating $\mathcal{E}(\alpha, \beta)$ with respect to $\alpha$ and $\beta$ respectively, we get

$$\frac{\partial}{\partial \alpha}\mathcal{E}(\alpha, \beta) = 2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-1)$$

and

$$\frac{\partial}{\partial \beta}\mathcal{E}(\alpha, \beta) = 2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-x_i).$$

Setting these partial derivatives $\frac{\partial}{\partial \alpha}\mathcal{E}(\alpha, \beta)$ and $\frac{\partial}{\partial \beta}\mathcal{E}(\alpha, \beta)$ to 0, we get

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \tag{3}$$

and

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0. \tag{4}$$

From (3), we obtain

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$

which is

$$\overline{y} = \alpha + \beta \overline{x}. \tag{5}$$

Similarly, from (4), we have

$$\sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$$

which can be rewritten as follows

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+n\overline{x}\,\overline{y}=n\,\alpha\,\overline{x}+\beta\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})+n\beta\,\overline{x}^2 \qquad (6)$$

Defining

$$S_{xy}:=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

we see that (6) reduces to

$$S_{xy}+n\overline{x}\,\overline{y}=\alpha\,n\,\overline{x}+\beta\,\left[S_{xx}+n\overline{x}^2\right] \qquad (7)$$

Substituting (5) into (7), we have

$$S_{xy}+n\overline{x}\,\overline{y}=[\overline{y}-\beta\,\overline{x}]\,n\,\overline{x}+\beta\,\left[S_{xx}+n\overline{x}^2\right].$$

Simplifying the last equation, we get

$$S_{xy}=\beta\,S_{xx}$$

which is

$$\beta=\frac{S_{xy}}{S_{xx}}. \qquad (8)$$

In view of (8) and (5), we get

$$\alpha=\overline{y}-\frac{S_{xy}}{S_{xx}}\,\overline{x}. \qquad (9)$$

Thus the least squares estimates of $\alpha$ and $\beta$ are

$$\widehat{\alpha}=\overline{y}-\frac{S_{xy}}{S_{xx}}\,\overline{x}\quad\text{and}\quad\widehat{\beta}=\frac{S_{xy}}{S_{xx}},$$

respectively.

We need some notations. The random variable $Y$ given $X=x$ will be denoted by $Y_x$. Note that this is the variable appears in the model $E(Y/x)=\alpha+\beta x$. When one chooses in succession values $x_1,x_2,...,x_n$ for $x$, a sequence $Y_{x_1},Y_{x_2},...,Y_{x_n}$ of random variable is obtained. For the sake of convenience, we denote the random variables $Y_{x_1},Y_{x_2},...,Y_{x_n}$ simply as $Y_1,Y_2,...,Y_n$. To do some statistical analysis, we make following three assumptions:

(1) $E(Y_x)=\alpha+\beta\,x$ so that $\mu_i=E(Y_i)=\alpha+\beta\,x_i$;

(2) $Y_1, Y_2, ..., Y_n$ are independent;

(3) Each of the random variables $Y_1, Y_2, ..., Y_n$ has the same variance $\sigma^2$.

**Theorem 19.1.** Under the above three assumptions, the least squares estimators $\widehat{\alpha}$ and $\widehat{\beta}$ of a linear model $E(Y/x) = \alpha + \beta\, x$ are unbiased.

**Proof:** From the previous example, we know that the least squares estimators of $\alpha$ and $\beta$ are

$$\widehat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\ \overline{X} \quad \text{and} \quad \widehat{\beta} = \frac{S_{xY}}{S_{xx}},$$

where

$$S_{xY} := \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}).$$

First, we show $\widehat{\beta}$ is unbiased. Consider

$$\begin{aligned}
E\left(\widehat{\beta}\right) &= E\left(\frac{S_{xY}}{S_{xx}}\right) = \frac{1}{S_{xx}}\ E\left(S_{xY}\right) \\
&= \frac{1}{S_{xx}}\ E\left(\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})\right) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ E\left(Y_i - \overline{Y}\right) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ E\left(Y_i\right) - \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ E\left(\overline{Y}\right) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ E\left(Y_i\right) - \frac{1}{S_{xx}}\ E\left(\overline{Y}\right)\ \sum_{i=1}^{n}(x_i - \overline{x}) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ E\left(Y_i\right) = \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ (\alpha + \beta x_i) \\
&= \alpha\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x}) + \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ x_i \\
&= \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ x_i \\
&= \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ x_i - \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ \overline{x} \\
&= \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n}(x_i - \overline{x})\ (x_i - \overline{x}) \\
&= \beta\ \frac{1}{S_{xx}}\ S_{xx} = \beta.
\end{aligned}$$

Thus the estimator $\widehat{\beta}$ is unbiased estimator of the parameter $\beta$.

Next, we show that $\widehat{\alpha}$ is also an unbiased estimator of $\alpha$. Consider

$$\begin{aligned}
E(\widehat{\alpha}) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\,\overline{x}\right) = E\left(\overline{Y}\right) - \overline{x}\; E\left(\frac{S_{xY}}{S_{xx}}\right) \\
&= E\left(\overline{Y}\right) - \overline{x}\, E\left(\widehat{\beta}\right) = E\left(\overline{Y}\right) - \overline{x}\;\beta \\
&= \frac{1}{n}\;\left(\sum_{i=1}^{n} E\left(Y_i\right)\right) - \overline{x}\;\beta \\
&= \frac{1}{n}\;\left(\sum_{i=1}^{n} E\left(\alpha + \beta x_i\right)\right) - \overline{x}\;\beta \\
&= \frac{1}{n}\;\left(n\alpha + \beta \sum_{i=1}^{n} x_i\right) - \overline{x}\;\beta \\
&= \alpha + \beta\,\overline{x} - \overline{x}\;\beta = \alpha
\end{aligned}$$

This proves that $\widehat{\alpha}$ is an unbiased estimator of $\alpha$ and the proof of the theorem is now complete.

### 19.2. The Normal Regression Analysis

In a regression analysis, we assume that the $x_i$'s are constants while $y_i$'s are values of the random variables $Y_i$'s. A regression analysis is called a normal regression analysis if the conditional density of $Y_i$ given $X_i = x_i$ is of the form

$$f(y_i/x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{y_i-\alpha-\beta x_i}{\sigma}\right)^2},$$

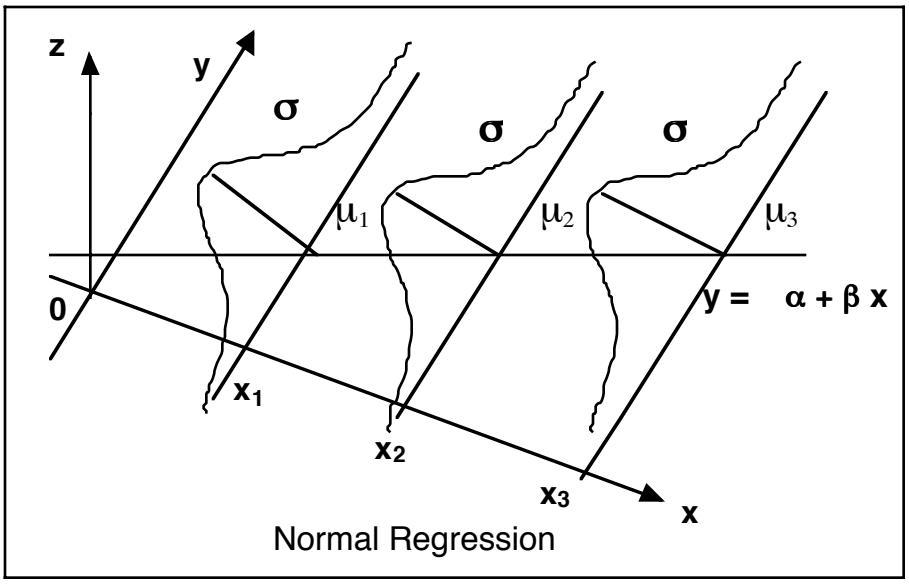
where $\sigma^2$ denotes the variance, and $\alpha$ and $\beta$ are the regression coefficients. That is $Y|_{x_i} \sim N(\alpha + \beta x, \sigma^2)$. If there is no danger of confusion, then we will write $Y_i$ for $Y|_{x_i}$. The figure on the next page shows the regression model of $Y$ with equal variances, and with means falling on the straight line $\mu_y = \alpha + \beta\, x$.

Normal regression analysis concerns with the estimation of $\sigma$, $\alpha$, and $\beta$. We use maximum likelihood method to estimate these parameters. The maximum likelihood function of the sample is given by

$$L(\sigma, \alpha, \beta) = \prod_{i=1}^{n} f(y_i/x_i)$$

and

$$\ln L(\sigma, \alpha, \beta) = \sum_{i=1}^{n} \ln f(y_i/x_i)$$
$$= -n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 .$$




Taking the partial derivatives of $\ln L(\sigma, \alpha, \beta)$ with respect to $\alpha, \beta$ and $\sigma$ respectively, we get

$$\frac{\partial}{\partial \alpha} \ln L(\sigma, \alpha, \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)$$

$$\frac{\partial}{\partial \beta} \ln L(\sigma, \alpha, \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)\, x_i$$

$$\frac{\partial}{\partial \sigma} \ln L(\sigma, \alpha, \beta) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 .$$

Equating each of these partial derivatives to zero and solving the system of three equations, we obtain the maximum likelihood estimator of $\beta, \alpha, \sigma$ as

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}, \qquad \widehat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\, \overline{x}, \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n} \left[ S_{YY} - \frac{S_{xY}}{S_{xx}}\, S_{xY} \right]},$$

where

$$S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right).$$

**Theorem 19.2.** In the normal regression analysis, the likelihood estimators $\widehat{\beta}$ and $\widehat{\alpha}$ are unbiased estimators of $\beta$ and $\alpha$, respectively.

**Proof:** Recall that

$$\begin{aligned}
\widehat{\beta} &= \frac{S_{xY}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i,
\end{aligned}$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$. Thus $\widehat{\beta}$ is a linear combination of $Y_i$'s. Since $Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$, we see that $\widehat{\beta}$ is also a normal random variable.

First we show $\widehat{\beta}$ is an unbiased estimator of $\beta$. Since

$$\begin{aligned}
E\left(\widehat{\beta}\right) &= E\left(\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) E\left(Y_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) (\alpha + \beta x_i) = \beta,
\end{aligned}$$

the maximum likelihood estimator of $\beta$ is unbiased.

Next, we show that $\widehat{\alpha}$ is also an unbiased estimator of $\alpha$. Consider

$$\begin{aligned}
E\left(\widehat{\alpha}\right) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\,\overline{x}\right) = E\left(\overline{Y}\right) - \overline{x}\; E\left(\frac{S_{xY}}{S_{xx}}\right) \\
&= E\left(\overline{Y}\right) - \overline{x}\, E\left(\widehat{\beta}\right) = E\left(\overline{Y}\right) - \overline{x}\; \beta \\
&= \frac{1}{n}\; \left(\sum_{i=1}^{n} E\left(Y_i\right)\right) - \overline{x}\; \beta \\
&= \frac{1}{n}\; \left(\sum_{i=1}^{n} E\left(\alpha + \beta x_i\right)\right) - \overline{x}\; \beta \\
&= \frac{1}{n}\; \left(n\alpha + \beta \sum_{i=1}^{n} x_i\right) - \overline{x}\; \beta \\
&= \alpha + \beta\, \overline{x} - \overline{x}\; \beta = \alpha.
\end{aligned}$$

This proves that $\widehat{\alpha}$ is an unbiased estimator of $\alpha$ and the proof of the theorem is now complete.

**Theorem 19.3.** In normal regression analysis, the distributions of the estimators $\widehat{\beta}$ and $\widehat{\alpha}$ are given by

$$\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \qquad \text{and} \qquad \widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

**Proof:** Since

$$\begin{aligned}
\widehat{\beta} &= \frac{S_{xY}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i,
\end{aligned}$$

the $\widehat{\beta}$ is a linear combination of $Y_i$'s. As $Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$, we see that $\widehat{\beta}$ is also a normal random variable. By Theorem 19.2, $\widehat{\beta}$ is an unbiased estimator of $\beta$.

The variance of $\widehat{\beta}$ is given by

$$\begin{aligned}
Var\left(\widehat{\beta}\right) &= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 Var\left(Y_i/x_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 \sigma^2 \\
&= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2\ \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

Hence $\widehat{\beta}$ is a normal random variable with mean (or expected value) $\beta$ and variance $\frac{\sigma^2}{S_{xx}}$. That is $\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$.

Now determine the distribution of $\widehat{\alpha}$. Since each $Y_i \sim N(\alpha + \beta x_i,\ \sigma^2)$, the distribution of $\overline{Y}$ is given by

$$\overline{Y} \sim N\left(\alpha + \beta\overline{x},\ \frac{\sigma^2}{n}\right).$$

Since

$$\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$$

the distribution of $\overline{x}\,\widehat{\beta}$ is given by

$$\overline{x}\,\widehat{\beta} \sim N\left(\overline{x}\,\beta,\ \overline{x}^2\,\frac{\sigma^2}{S_{xx}}\right).$$

Since $\widehat{\alpha} = \overline{Y} - \overline{x}\,\widehat{\beta}$ and $\overline{Y}$ and $\overline{x}\,\widehat{\beta}$ being two normal random variables, $\widehat{\alpha}$ is also a normal random variable with mean equal to $\alpha + \beta\,\overline{x} - \beta\,\overline{x} = \alpha$ and variance variance equal to $\frac{\sigma^2}{n} + \frac{\overline{x}^2\,\sigma^2}{S_{xx}}$. That is

$$\widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

and the proof of the theorem is now complete.

It should be noted that in the proof of the last theorem, we have assumed the fact that $\overline{Y}$ and $\overline{x}\,\widehat{\beta}$ are statistically independent.

In the next theorem, we give an unbiased estimator of the variance $\sigma^2$. For this we need the distribution of the statistic $U$ given by

$$U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}.$$

It can be shown (we will omit the proof, for a proof see Graybill (1961)) that the distribution of the statistic

$$U = \frac{n\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

**Theorem 19.4.** An unbiased estimator $S^2$ of $\sigma^2$ is given by

$$S^2 = \frac{n\,\widehat{\sigma}^2}{n-2},$$

where $\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{YY} - \frac{S_{xY}}{S_{xx}}\,S_{xY}\right]}$.

**Proof:** Since

$$\begin{aligned} E(S^2) &= E\left(\frac{n\,\widehat{\sigma}^2}{n-2}\right) \\ &= \frac{\sigma^2}{n-2}\,E\left(\frac{n\,\widehat{\sigma}^2}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n-2}\,E(\chi^2(n-2)) \\ &= \frac{\sigma^2}{n-2}\,(n-2) = \sigma^2. \end{aligned}$$

The proof of the theorem is now complete.

Note that the estimator $S^2$ can be written as $S^2 = \frac{SSE}{n-2}$, where

$$SSE = S_{YY} = \widehat{\beta}\, S_{xY} = \sum_{i=1}^{2} [y_i - \widehat{\alpha} - \widehat{\beta}\, x_i]$$

the estimator $S^2$ is unbiased estimator of $\sigma^2$. The proof of the theorem is now complete.

In the next theorem we give the distribution of two statistics that can be used for testing hypothesis and constructing confidence interval for the regression parameters $\alpha$ and $\beta$.

**Theorem 19.5.** The statistics

$$Q_\beta = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n}}$$

and

$$Q_\alpha = \frac{\widehat{\alpha} - \alpha}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n\, (\overline{x})^2 + S_{xx}}}$$

have both a $t$-distribution with $n-2$ degrees of freedom.

**Proof:** From Theorem 19.3, we know that

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right).$$

Hence by standardizing, we get

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

Further, we know that the likelihood estimator of $\sigma$ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{YY} - \frac{S_{xY}}{S_{xx}}\, S_{xY}\right]}$$

and the distribution of the statistic $U = \frac{n\widehat{\sigma^2}}{\sigma^2}$ is chi-square with $n-2$ degrees of freedom.

Since $Z = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ and $U = \frac{n\,\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$, by Theorem 14.6, the statistic $\frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2)$. Hence

$$Q_\beta = \frac{\widehat{\beta}-\beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{n}} = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\,\widehat{\sigma^2}}{(n-2)\,S_{xx}}}} = \frac{\frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{n\,\widehat{\sigma^2}}{(n-2)\,\sigma^2}}} \sim t(n-2).$$

Similarly, it can be shown that

$$Q_\alpha = \frac{\widehat{\alpha}-\alpha}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{n\,(\overline{x})^2 + S_{xx}}} \sim t(n-2).$$

This completes the proof of the theorem.

In the normal regression model, if $\beta = 0$, then $E(Y_x) = \alpha$. This implies that $E(Y_x)$ does not depend on $x$. Therefore if $\beta \neq 0$, then $E(Y_x)$ is dependent on $x$. Thus the null hypothesis $H_o : \beta = 0$ should be tested against $H_a : \beta \neq 0$. To devise a test we need the distribution of $\widehat{\beta}$. Theorem 19.3 says that $\widehat{\beta}$ is normally distributed with mean $\beta$ and variance $\frac{\sigma^2}{S_xx}$. Therefore, we have

$$Z = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

In practice the variance $Var(Y_i/x_i)$ which is $\sigma^2$ is usually unknown. Hence the above statistic $Z$ is not very useful. However, using the statistic $Q_\beta$, we can devise a hypothesis test to test the hypothesis $H_o : \beta = \beta_o$ against $H_a : \beta \neq \beta_o$ at a significance level $\gamma$. For this one has to evaluate the quantity

$$|t| = \left|\frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\,\widehat{\sigma^2}}{(n-2)\,S_{xx}}}}\right|$$

$$= \left|\frac{\widehat{\beta}-\beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{n}}\right|$$

and compare it to quantile $t_{\gamma/2}(n-2)$. The hypothesis test, at significance level $\gamma$, is then "Reject $H_o : \beta = \beta_o$ if $|t| > t_{\gamma/2}(n-2)$".

The statistic

$$Q_\beta = \frac{\widehat{\beta}-\beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{n}}$$

is a pivotal quantity for the parameter $\beta$ since the distribution of this quantity $Q_\beta$ is a $t$-distribution with $n-2$ degrees of freedom. Thus it can be used for the construction of a $(1-\gamma)100\%$ confidence interval for the parameter $\beta$ as follows:

$$
\begin{aligned}
&1-\gamma \\
&= P\left(-t_{\frac{\gamma}{2}}(n-2) \leq \frac{\widehat{\beta}-\beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} \leq t_{\frac{\gamma}{2}}(n-2)\right) \\
&= P\left(\widehat{\beta}-t_{\frac{\gamma}{2}}(n-2)\widehat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} \leq \beta \leq \widehat{\beta}+t_{\frac{\gamma}{2}}(n-2)\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\right).
\end{aligned}
$$

Hence, the $(1-\gamma)\%$ confidence interval for $\beta$ is given by

$$
\left[\widehat{\beta}-t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\,,\quad \widehat{\beta}+t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\,\right].
$$

In a similar manner one can devise hypothesis test for $\alpha$ and construct confidence interval for $\alpha$ using the statistic $Q_\alpha$. We leave these to the reader.

Now we give two examples to illustrate how to find the normal regression line and related things.

**Example 19.7.** Let the following data on the number of hours, $x$ which ten persons studied for a French test and their scores, $y$ on the test is shown below:

| $x$ | 4 | 9 | 10 | 14 | 4 | 7 | 12 | 22 | 1 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 31 | 58 | 65 | 73 | 37 | 44 | 60 | 91 | 21 | 84 |

Find the normal regression line that approximates the regression of test scores on the number of hours studied. Further test the hypothesis $H_o : \beta = 3$ versus $H_a : \beta \neq 3$ at the significance level 0.02.

**Answer:** From the above data, we have

$$
\sum_{i=1}^{10} x_i = 100, \qquad \sum_{i=1}^{10} x_i^2 = 1376
$$

$$
\sum_{i=1}^{10} y_i = 564, \qquad \sum_{i=1}^{10} y_i^2 =
$$

$$
\sum_{i=1}^{10} x_i y_i = 6945
$$

$$S_{xx} = 376, \qquad S_{xy} = 1305, \qquad S_{yy} = 4752.4.$$

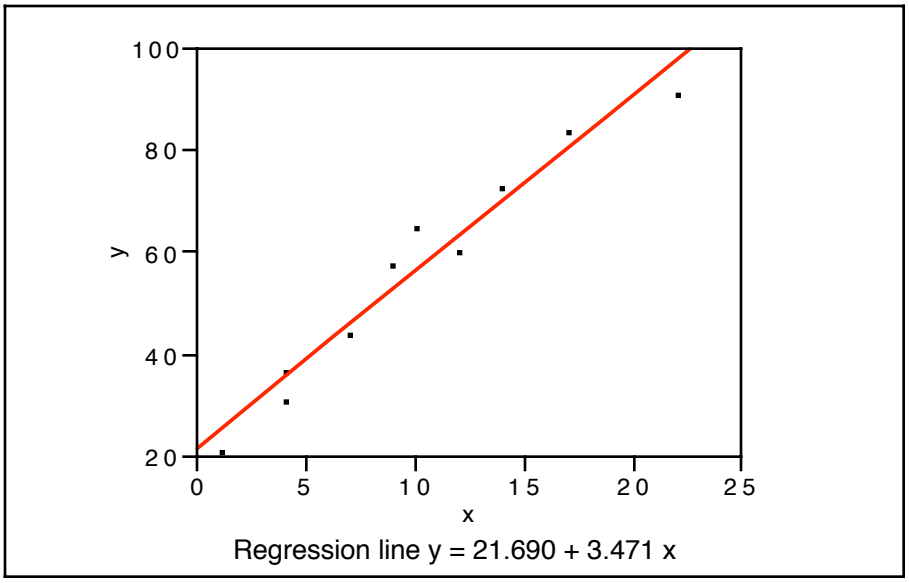
Hence

$$\widehat{\beta} = \frac{s_{xy}}{s_{xx}} = 3.471 \quad \text{and} \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 21.690.$$

Thus the normal regression line is

$$y = 21.690 + 3.471x.$$

This regression line is shown below.




Now we test the hypothesis $H_o : \beta = 3$ against $H_a : \beta \neq 3$ at 0.02 level of significance. From the data, the maximum likelihood estimate of $\sigma$ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\,S_{xy}\right]}$$

$$= \sqrt{\frac{1}{n}\left[S_{yy} - \widehat{\beta}\,S_{xy}\right]}$$

$$= \sqrt{\frac{1}{10}\left[4752.4 - (3.471)(1305)\right]}$$

$$= 4.720$$

and

$$|t| = \left| \frac{3.471 - 3}{4.720} \sqrt{\frac{(8)\,(376)}{10}} \right| = 1.73.$$

Hence

$$1.73 = |t| < t_{0.01}(8) = 2.896.$$

Thus we do not reject the null hypothesis that $H_o : \beta = 3$ at the significance level 0.02.

This means that we can not conclude that on the average an extra hour of study will increase the score by more than 3 points.

**Example 19.8.** The frequency of chirping of a cricket is thought to be related to temperature. This suggests the possibility that temperature can be estimated from the chirp frequency. Let the following data on the number chirps per second, $x$ by the striped ground cricket and the temperature, $y$ in Fahrenheit is shown below:

| $x$ | 20 | 16 | 20 | 18 | 17 | 16 | 15 | 17 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 89 | 72 | 93 | 84 | 81 | 75 | 70 | 82 | 69 | 83 |

Find the normal regression line that approximates the regression of temperature on the number chirps per second by the striped ground cricket. Further test the hypothesis $H_o : \beta = 4$ versus $H_a : \beta \neq 4$ at the significance level 0.1.

**Answer:** From the above data, we have

$$\sum_{i=1}^{10} x_i = 170, \qquad \sum_{i=1}^{10} x_i^2 = 2920$$

$$\sum_{i=1}^{10} y_i = 789, \qquad \sum_{i=1}^{10} y_i^2 = 64270$$

$$\sum_{i=1}^{10} x_i y_i = 13688$$

$$S_{xx} = 376, \qquad S_{xy} = 1305, \qquad S_{yy} = 4752.4.$$

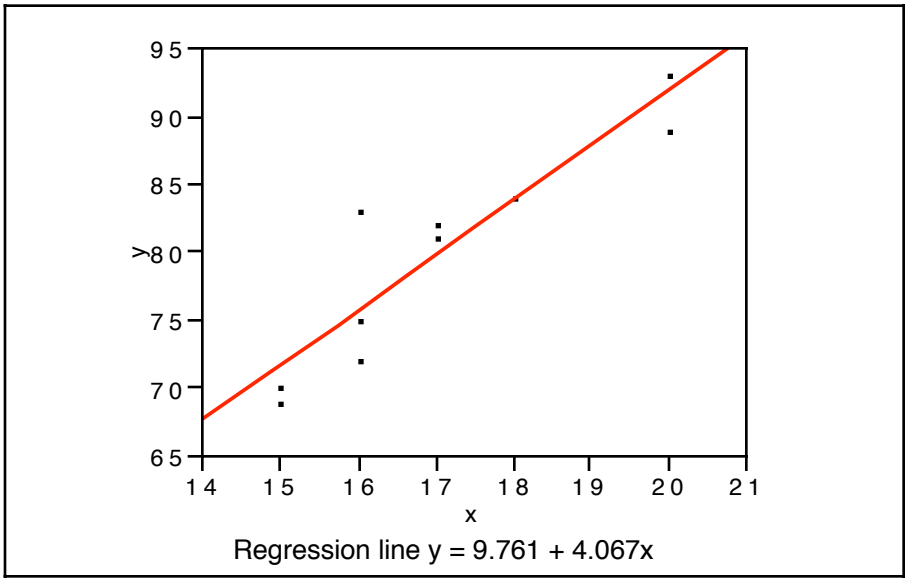
Hence

$$\widehat{\beta} = \frac{s_{xy}}{s_{xx}} = 4.067 \quad \text{and} \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 9.761.$$

Thus the normal regression line is

$$y = 9.761 + 4.067x.$$

This regression line is shown below.




Now we test the hypothesis $H_o : \beta = 4$ against $H_a : \beta \neq 4$ at 0.1 level of significance. From the data, the maximum likelihood estimate of $\sigma$ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right]}$$

$$= \sqrt{\frac{1}{n}\left[S_{yy} - \widehat{\beta}\, S_{xy}\right]}$$

$$= \sqrt{\frac{1}{10}\left[589 - (4.067)(122)\right]}$$

$$= 3.047$$

and

$$|t| = \left|\frac{4.067 - 4}{3.047}\sqrt{\frac{(8)\,(30)}{10}}\right| = 0.528.$$

Hence

$$0.528 = |t| < t_{0.05}(8) = 1.860.$$

Thus we do not reject the null hypothesis that $H_o : \beta = 4$ at a significance level 0.1.

Let $\mu_x = \alpha + \beta\, x$ and write $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\, x$ for an arbitrary but fixed $x$. Then $\widehat{Y}_x$ is an estimator of $\mu_x$. The following theorem gives various properties of this estimator.

**Theorem 19.6.** Let $x$ be an arbitrary but fixed real number. Then
(i) $\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$,
(ii) $\widehat{Y}_x$ is an unbiased estimator of $\mu_x$, and
(iii) $Var\left(\widehat{Y}_x\right) = \left\{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right\}\sigma^2$.

**Proof:** First we show that $\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$. Since

$$\begin{aligned}
\widehat{Y}_x &= \widehat{\alpha} + \widehat{\beta}\, x \\
&= \overline{Y} - \widehat{\beta}\overline{x} + \widehat{\beta}\, x \\
&= \overline{Y} + \widehat{\beta}\,(x - \overline{x}) \\
&= \overline{Y} + \sum_{k=1}^{n} \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\, Y_k \\
&= \sum_{k=1}^{n} \frac{Y_k}{n} + \sum_{k=1}^{n} \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\, Y_k \\
&= \sum_{k=1}^{n} \left(\frac{1}{n} + \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\right) Y_k
\end{aligned}$$

$\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$.

Next, we show that $\widehat{Y}_x$ is an unbiased estimator of $\mu_x$. Since

$$\begin{aligned}
E\left(\widehat{Y}_x\right) &= E\left(\widehat{\alpha} + \widehat{\beta}\, x\right) \\
&= E\left(\widehat{\alpha}\right) + E\left(\widehat{\beta}\, x\right) \\
&= \alpha + \beta\, x \\
&= \mu_x
\end{aligned}$$

$\widehat{Y}_x$ is an unbiased estimator of $\mu_x$.

Finally, we calculate the variance of $\widehat{Y}_x$ using Theorem 19.3. The variance

of $\widehat{Y}_x$ is given by

$$\begin{aligned} Var\left(\widehat{Y}_x\right) &= Var\left(\widehat{\alpha}+\widehat{\beta}\,x\right) \\ &= Var\left(\widehat{\alpha}\right)+Var\left(\widehat{\beta}\,x\right)+2Cov\left(\widehat{\alpha},\,\widehat{\beta}\,x\right) \\ &= \left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)+x^2\,\frac{\sigma^2}{S_{xx}}+2\,x\,Cov\left(\widehat{\alpha},\,\widehat{\beta}\right) \\ &= \left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)-2\,x\,\frac{\overline{x}\,\sigma^2}{S_{xx}} \\ &= \left(\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2. \end{aligned}$$

In this computation we have used the fact that

$$Cov\left(\widehat{\alpha},\,\widehat{\beta}\right)=-\frac{\overline{x}\,\sigma^2}{S_{xx}}$$

whose proof is left to the reader as an exercise. The proof of the theorem is now complete.

By Theorem 19.3, we see that

$$\widehat{\beta}\sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)\qquad\text{and}\qquad\widehat{\alpha}\sim N\left(\alpha,\ \frac{\sigma^2}{n}+\frac{\overline{x}^2\sigma^2}{S_{xx}}\right).$$

Since $\widehat{Y}_x=\widehat{\alpha}+\widehat{\beta}\,x$, the random variable $\widehat{Y}_x$ is also a normal random variable with mean $\mu_x$ and variance

$$Var\left(\widehat{Y}_x\right)=\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2.$$

Hence standardizing $\widehat{Y}_x$, we have

$$\frac{\widehat{Y}_x-\mu_x}{\sqrt{Var\left(\widehat{Y}_x\right)}}\sim N(0,1).$$

If $\sigma^2$ is known, then one can take the statistic $Q=\dfrac{\widehat{Y}_x-\mu_x}{\sqrt{Var\left(\widehat{Y}_x\right)}}$ as a pivotal quantity to construct a confidence interval for $\mu_x$. The $(1-\gamma)100\%$ confidence interval for $\mu_x$ when $\sigma^2$ is known is given by

$$\left[\widehat{Y}_x-z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)},\ \ \widehat{Y}_x+z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)}\ \right].$$

**Example 19.9.** Let the following data on the number chirps per second, $x$ by the striped ground cricket and the temperature, $y$ in Fahrenheit is shown below:

| $x$ | 20 | 16 | 20 | 18 | 17 | 16 | 15 | 17 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 89 | 72 | 93 | 84 | 81 | 75 | 70 | 82 | 69 | 83 |

What is the 95% confidence interval for $\beta$? What is the 95% confidence interval for $\mu_x$ when $x = 14$ and $\sigma = 3.047$?

**Answer:** From Example 19.8, we have

$$n = 10, \quad \widehat{\beta} = 4.067, \quad \widehat{\sigma} = 3.047 \quad \text{and} \quad S_{xx} = 376.$$

The $(1-\gamma)\%$ confidence interval for $\beta$ is given by

$$\left[ \widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\,, \quad \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}} \right].$$

Therefore the 90% confidence interval for $\beta$ is

$$\left[ 4.067 - t_{0.025}(8)\,(3.047)\sqrt{\frac{10}{(8)\,(376)}}\,,\ 4.067 + t_{0.025}(8)\,(3.047)\sqrt{\frac{10}{(8)\,(376)}} \right]$$

which is

$$\left[\, 4.067 - t_{0.025}(8)\,(0.1755)\,, \quad 4.067 + t_{0.025}(8)\,(0.1755)\right].$$

Since from the $t$-table, we have $t_{0.025}(8) = 2.306$, the 90% confidence interval for $\beta$ becomes

$$\left[\, 4.067 - (2.306)\,(0.1755)\,, \quad 4.067 + (2.306)\,(0.1755)\right]$$

which is [3.6623, 4.4717].

If variance $\sigma^2$ is not known, then we can use the fact that the statistic $U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}$ is chi-squares with $n-2$ degrees of freedom to obtain a pivotal quantity for $\mu_x$. This can be done as follows:

$$\begin{aligned} Q &= \frac{\widehat{Y}_x - \mu_x}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{S_{xx} + n\,(x-\overline{x})^2}} \\ &= \frac{\frac{\widehat{Y}_x - \mu_x}{\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \sim t(n-2). \end{aligned}$$

Using this pivotal quantity one can construct a $(1-\gamma)100\%$ confidence interval for mean $\mu$ as

$$\left[\widehat{Y}_x - t_{\frac{\gamma}{2}}(n-2)\sqrt{\frac{S_{xx}+n(x-\overline{x})^2}{(n-2)\,S_{xx}}},\ \ \widehat{Y}_x + t_{\frac{\gamma}{2}}(n-2)\sqrt{\frac{S_{xx}+n(x-\overline{x})^2}{(n-2)\,S_{xx}}}\ \right].$$

Next we determine the 90% confidence interval for $\mu_x$ when $x = 14$ and $\sigma = 3.047$. The $(1-\gamma)100\%$ confidence interval for $\mu_x$ when $\sigma^2$ is known is given by

$$\left[\widehat{Y}_x - z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)},\ \ \widehat{Y}_x + z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)}\ \right].$$

From the data, we have

$$\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\,x = 9.761 + (4.067)\,(14) = 66.699$$

and

$$Var\left(\widehat{Y}_x\right) = \left(\frac{1}{10} + \frac{(14-17)^2}{376}\right)\sigma^2 = (0.124)\,(3.047)^2 = 1.1512.$$

The 90% confidence interval for $\mu_x$ is given by

$$\left[66.699 - z_{0.025}\,\sqrt{1.1512},\ 66.699 + z_{0.025}\,\sqrt{1.1512}\ \right]$$

and since $z_{0.025} = 1.96$ (from the normal table), we have

$$[66.699 - (1.96)\,(1.073),\ 66.699 + (1.96)\,(1.073)]$$

which is $[64.596,\ 68.802]$.

We now consider the predictions made by the normal regression equation $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}x$. The quantity $\widehat{Y}_x$ gives an estimate of $\mu_x = \alpha + \beta x$. Each time we compute a regression line from a random sample we are observing one possible linear equation in a population consisting all possible linear equations. Further, the actual value of $Y_x$ that will be observed for given value of $x$ is normal with mean $\alpha + \beta x$ and variance $\sigma^2$. So the actual observed value will be different from $\mu_x$. Thus, the predicted value for $\widehat{Y}_x$ will be in error from two different sources, namely (1) $\widehat{\alpha}$ and $\widehat{\beta}$ are randomly distributed about $\alpha$ and $\beta$, and (2) $Y_x$ is randomly distributed about $\mu_x$.

Let $y_x$ denote the actual value of $Y_x$ that will be observed for the value $x$ and consider the random variable

$$\mathcal{D} = Y_x - \widehat{\alpha} - \widehat{\beta}\, x.$$

Since $\mathcal{D}$ is a linear combination of normal random variables, $\mathcal{D}$ is also a normal random variable.

The mean of $\mathcal{D}$ is given by

$$\begin{aligned} E(\mathcal{D}) &= E(Y_x) - E(\widehat{\alpha}) - x\, E(\widehat{\beta}) \\ &= \alpha + \beta\, x - \alpha - x\, \beta \\ &= 0. \end{aligned}$$

The variance of $\mathcal{D}$ is given by

$$\begin{aligned} Var(\mathcal{D}) &= Var(Y_x - \widehat{\alpha} - \widehat{\beta}\, x) \\ &= Var(Y_x) + Var(\widehat{\alpha}) + x^2\, Var(\widehat{\beta}) + 2\, x\, Cov(\widehat{\alpha}, \widehat{\beta}) \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{\overline{x}^2\, \sigma^2}{S_{xx}} + x^2\, \frac{\sigma^2}{S_{xx}} - 2\, x\, \frac{\overline{x}}{S_{xx}} \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{(x - \overline{x})^2\, \sigma^2}{S_{xx}} \\ &= \frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2. \end{aligned}$$

Therefore

$$\mathcal{D} \sim N\left(0,\ \frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2\right).$$

We standardize $\mathcal{D}$ to get

$$Z = \frac{\mathcal{D} - 0}{\sqrt{\frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2}} \sim N(0, 1).$$

Since in practice the variance of $Y_x$ which is $\sigma^2$ is unknown, we can not use $Z$ to construct a confidence interval for a predicted value $y_x$.

We know that $U = \frac{n\, \widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$. By Theorem 14.6, the statistic

$\frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2)$. Hence

$$\begin{aligned} Q &= \frac{y_x - \widehat{\alpha} - \widehat{\beta}\, x}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{(n+1)\, S_{xx} + n}} \\ &= \frac{\frac{y_x - \widehat{\alpha} - \widehat{\beta}\, x}{\sqrt{\frac{(n+1)\, S_{xx}+n}{n\, S_{xx}}\, \sigma^2}}}{\sqrt{\frac{n\, \widehat{\sigma}^2}{(n-2)\, \sigma^2}}} \\ &= \frac{\frac{\mathcal{D}-0}{\sqrt{Var(\mathcal{D})}}}{\sqrt{\frac{n\, \widehat{\sigma}^2}{(n-2)\, \sigma^2}}} \\ &= \frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2). \end{aligned}$$

The statistic $Q$ is a pivotal quantity for the predicted value $y_x$ and one can use it to construct a $(1-\gamma)100\%$ confidence interval for $y_x$. The $(1-\gamma)100\%$ confidence interval, $[a, b]$, for $y_x$ is given by

$$\begin{aligned} 1 - \gamma &= P\left(-t_{\frac{\gamma}{2}}(n-2) \leq Q \leq t_{\frac{\gamma}{2}}(n-2)\right) \\ &= P(a \leq y_x \leq b), \end{aligned}$$

where

$$a = \widehat{\alpha} + \widehat{\beta}\, x - t_{\frac{\gamma}{2}}(n-2)\, \widehat{\sigma} \sqrt{\frac{(n+1)\, S_{xx} + n}{(n-2)\, S_{xx}}}$$

and

$$b = \widehat{\alpha} + \widehat{\beta}\, x + t_{\frac{\gamma}{2}}(n-2)\, \widehat{\sigma} \sqrt{\frac{(n+1)\, S_{xx} + n}{(n-2)\, S_{xx}}}.$$

This confidence interval for $y_x$ is usually known as the *prediction interval* for predicted value $y_x$ based on the given $x$. The prediction interval represents an interval that has a probability equal to $1-\gamma$ of containing not a parameter but a future value $y_x$ of the random variable $Y_x$. In many instances the prediction interval is more relevant to a scientist or engineer than the confidence interval on the mean $\mu_x$.

**Example 19.10.** Let the following data on the number chirps per second, $x$ by the striped ground cricket and the temperature, $y$ in Fahrenheit is shown below:

| $x$ | 20 | 16 | 20 | 18 | 17 | 16 | 15 | 17 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 89 | 72 | 93 | 84 | 81 | 75 | 70 | 82 | 69 | 83 |

What is the 95% prediction interval for $y_x$ when $x = 14$?

**Answer:** From Example 19.8, we have

$$n = 10, \quad \widehat{\beta} = 4.067, \quad \widehat{\alpha} = 9.761, \quad \widehat{\sigma} = 3.047 \quad \text{and} \quad S_{xx} = 376.$$

Thus the normal regression line is

$$y_x = 9.761 + 4.067x.$$

Since $x = 14$, the corresponding predicted value $y_x$ is given by

$$y_x = 9.761 + (4.067)\,(14) = 66.699.$$

Therefore

$$\begin{aligned} a &= \widehat{\alpha} + \widehat{\beta}\,x - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\,\sqrt{\frac{(n+1)\,S_{xx} + n}{(n-2)\,S_{xx}}} \\ &= 66.699 - t_{0.025}(8)\,(3.047)\,\sqrt{\frac{(11)\,(376) + 10}{(8)\,(376)}} \\ &= 66.699 - (2.306)\,(3.047)\,(1.1740) \\ &= 58.4501. \end{aligned}$$

Similarly

$$\begin{aligned} b &= \widehat{\alpha} + \widehat{\beta}\,x + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\,\sqrt{\frac{(n+1)\,S_{xx} + n}{(n-2)\,S_{xx}}} \\ &= 66.699 + t_{0.025}(8)\,(3.047)\,\sqrt{\frac{(11)\,(376) + 10}{(8)\,(376)}} \\ &= 66.699 + (2.306)\,(3.047)\,(1.1740) \\ &= 74.9479. \end{aligned}$$

Hence the 95% prediction interval for $y_x$ when $x = 14$ is [58.4501, 74.9479].

### 19.3. The Correlation Analysis

In the first two sections of this chapter, we examine the regression problem and have done an in-depth study of the least squares and the normal regression analysis. In the regression analysis, we assumed that the values of $X$ are not random variables, but are fixed. However, the values of $Y_x$ for

a given value of $x$ are randomly distributed about $E(Y_x) = \mu_x = \alpha + \beta x$. Further, letting $\varepsilon$ to be a random variable with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, one can model the so called regression problem by

$$Y_x = \alpha + \beta\, x + \varepsilon.$$

In this section, we examine the correlation problem. Unlike the regression problem, here both $X$ and $Y$ are random variables and the correlation problem can be modeled by

$$E(Y) = \alpha + \beta\, E(X).$$

From an experimental point of view this means that we are observing random vector $(X, Y)$ drawn from some bivariate population.

Recall that if $(X, Y)$ is a bivariate random variable then the correlation coefficient $\rho$ is defined as

$$\rho = \frac{E\left((X - \mu_X)\,(Y - \mu_Y)\right)}{\sqrt{E\left((X - \mu_X)^2\right)\; E\left((Y - \mu_Y)^2\right)}}$$

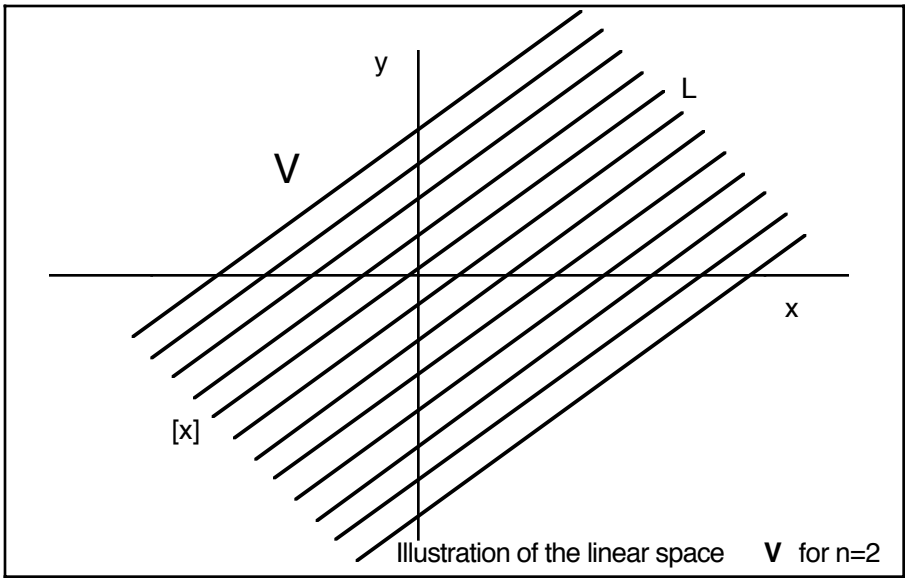
where $\mu_X$ and $\mu_Y$ are the mean of the random variables $X$ and $Y$, respectively.

**Definition 19.1.** If $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is a random sample from a bivariate population, then the sample correlation coefficient is defined as

$$R = \frac{\displaystyle\sum_{i=1}^{n} (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (X_i - \overline{X})^2}\;\sqrt{\displaystyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}.$$

The corresponding quantity computed from data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ will be denoted by $r$ and it is an estimate of the correlation coefficient $\rho$.

Now we give a geometrical interpretation of the sample correlation coefficient based on a paired data set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. We can associate this data set with two vectors $\vec{x} = (x_1, x_2, ..., x_n)$ and $\vec{y} = (y_1, y_2, ..., y_n)$ in $\mathbb{R}^n$. Let $\mathcal{L}$ be the subset $\{\lambda\, \vec{e}\,|\, \lambda \in \mathbb{R}\}$ of $\mathbb{R}^n$, where $\vec{e} = (1, 1, ..., 1) \in \mathbb{R}^n$. Consider the linear space $V$ given by $\mathbb{R}^n$ modulo $\mathcal{L}$, that is $V = \mathbb{R}^n/\mathcal{L}$. The linear space $V$ is illustrated in a figure on next page when $n = 2$.




We denote the equivalence class associated with the vector $\vec{x}$ by $[\vec{x}]$. In the linear space $V$ it can be shown that the points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ are collinear if and only if the the vectors $[\vec{x}]$ and $[\vec{y}]$ in $V$ are proportional.

We define an inner product on this linear space $V$ by

$$\langle [\vec{x}], [\vec{y}] \rangle = \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}).$$

Then the angle $\theta$ between the vectors $[\vec{x}]$ and $[\vec{y}]$ is given by

$$\cos(\theta) = \frac{\langle [\vec{x}], [\vec{y}] \rangle}{\sqrt{\langle [\vec{x}], [\vec{x}] \rangle}\,\sqrt{\langle [\vec{y}], [\vec{y}] \rangle}}$$

which is

$$\cos(\theta) = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}\,\sqrt{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})^2}} = r.$$

Thus the sample correlation coefficient $r$ can be interpreted geometrically as the cosine of the angle between the vectors $[\vec{x}]$ and $[\vec{y}]$. From this view point the following theorem is obvious.

**Theorem 19.7.** The sample correlation coefficient $r$ satisfies the inequality

$$-1 \leq r \leq 1.$$

The sample correlation coefficient $r = \pm 1$ if and only if the set of points $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ for $n \geq 3$ are collinear.

To do some statistical analysis, we assume that the paired data is a random sample of size $n$ from a bivariate normal population $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then the conditional distribution of the random variable $Y$ given $X = x$ is normal, that is

$$Y|_x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\, \sigma_2^2(1 - \rho^2)\right).$$

This can be viewed as a normal regression model $E(Y|_x) = \alpha + \beta\, x$ where $\alpha = \mu - \rho\frac{\sigma_2}{\sigma_1}\mu_1$, $\beta = \rho\frac{\sigma_2}{\sigma_1}$, and $Var(Y|_x) = \sigma_2^2(1 - \rho^2)$.

Since $\beta = \rho\frac{\sigma_2}{\sigma_1}$, if $\rho = 0$, then $\beta = 0$. Hence the null hypothesis $H_o : \rho = 0$ is equivalent to $H_o : \beta = 0$. In the previous section, we devised a hypothesis test for testing $H_o : \beta = \beta_o$ against $H_a : \beta \neq \beta_o$. This hypothesis test, at significance level $\gamma$, is "Reject $H_o : \beta = \beta_o$ if $|t| \geq t_{\frac{\gamma}{2}}(n - 2)$", where

$$t = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)\, S_{xx}}{n}}.$$

If $\beta = 0$, then we have

$$t = \frac{\widehat{\beta}}{\widehat{\sigma}}\sqrt{\frac{(n-2)\, S_{xx}}{n}}. \tag{10}$$

Now we express $t$ in term of the sample correlation coefficient $r$. Recall that

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}, \tag{11}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right], \tag{12}$$

and

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}. \tag{13}$$

Now using (11), (12), and (13), we compute

$$\begin{aligned} t &= \frac{\widehat{\beta}}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n}} \\ &= \frac{S_{xy}}{S_{xx}} \frac{\sqrt{n}}{\sqrt{\left[S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy}\right]}} \sqrt{\frac{(n-2)\, S_{xx}}{n}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} \frac{1}{\sqrt{\left[1 - \frac{S_{xy}}{S_{xx}} \frac{S_{xy}}{S_{yy}}\right]}} \sqrt{n-2} \\ &= \sqrt{n-2}\, \frac{r}{\sqrt{1-r^2}}. \end{aligned}$$

Hence to test the null hypothesis $H_o : \rho = 0$ against $H_a : \rho \neq 0$, at significance level $\gamma$, is "Reject $H_o : \rho = 0$ if $|t| \geq t_{\frac{\gamma}{2}}(n-2)$", where $t = \sqrt{n-2}\, \frac{r}{1-r^2}$.

This above test does not extend to test other values of $\rho$ except $\rho = 0$. However, tests for the nonzero values of $\rho$ can be achieved by the following result.

**Theorem 19.8.** Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be a random sample from a bivariate normal population $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. If

$$V = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right) \quad \text{and} \quad m = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right),$$

then

$$Z = \sqrt{n-3}\,(V - m) \to N(0,1) \quad \text{as } n \to \infty.$$

This theorem says that the statistic $V$ is approximately normal with mean $m$ and variance $\frac{1}{n-3}$ when $n$ is large. This statistic can be used to devise a hypothesis test for the nonzero values of $\rho$. Hence to test the null hypothesis $H_o : \rho = \rho_o$ against $H_a : \rho \neq \rho_o$, at significance level $\gamma$, is "Reject $H_o : \rho = \rho_o$ if $|z| \geq z_{\frac{\gamma}{2}}$", where $z = \sqrt{n-3}\,(V - m_o)$ and $m_o = \frac{1}{2} \ln\left(\frac{1+\rho_o}{1-\rho_o}\right)$.

**Example 19.11.** The following data were obtained in a study of the relationship between the weight and chest size of infants at birth:

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$, weight in kg | 2.76 | 2.17 | 5.53 | 4.31 | 2.30 | 3.70 |
| $y$, chest size in cm | 29.5 | 26.3 | 36.6 | 27.8 | 28.3 | 28.6 |

Determine the sample correlation coefficient $r$ and then test the null hypothesis $H_o : \rho = 0$ against the alternative hypothesis $H_a : \rho \neq 0$ at a significance level 0.01.

**Answer:** From the above data we find that

$$\overline{x} = 3.46 \qquad \text{and} \qquad \overline{y} = 29.51.$$

Next, we compute $S_{xx}$, $S_{yy}$ and $S_{xy}$ using a tabular representation.

| $x - \overline{x}$ | $y - \overline{y}$ | $(x - \overline{x})(y - \overline{y})$ | $(x - \overline{x})^2$ | $(y - \overline{y})^2$ |
|---|---|---|---|---|
| $-0.70$ | $-0.01$ | 0.007 | 0.490 | 0.000 |
| $-1.29$ | $-3.21$ | 4.141 | 1.664 | 10.304 |
| 2.07 | 7.09 | 14.676 | 4.285 | 50.268 |
| 0.85 | $-1.71$ | $-1.453$ | 0.722 | 2.924 |
| $-1.16$ | $-1.21$ | 1.404 | 1.346 | 1.464 |
| 0.24 | $-0.91$ | $-0.218$ | 0.058 | 0.828 |
| | | $S_{xy} = 18.557$ | $S_{xx} = 8.565$ | $S_{yy} = 65.788$ |

Hence, the correlation coefficient $r$ is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = \frac{18.557}{\sqrt{(8.565)\,(65.788)}} = 0.782.$$

The computed $t$ value is give by

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = \sqrt{(6-2)}\,\frac{0.782}{\sqrt{1-(0.782)^2}} = 2.509.$$

From the $t$-table we have $t_{0.005}(4) = 4.604$. Since

$$2.509 = |t| \ngeq t_{0.005}(4) = 4.604$$

we do not reject the null hypothesis $H_o : \rho = 0$.

### 19.4. Review Exercises

**1.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both $\beta$ and $\sigma^2$ are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimators of $\widehat{\beta}$ and $\widehat{\sigma}^2$ of $\beta$ and $\sigma^2$.

**2.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both $\beta$ and $\sigma^2$ are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then show that the maximum likelihood estimator of $\widehat{\beta}$ is normally distributed. What are the mean and variance of $\widehat{\beta}$?

**3.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both $\beta$ and $\sigma^2$ are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find an unbiased estimator $\widehat{\sigma}^2$ of $\sigma^2$ and then find a constant $k$ such that $k\,\widehat{\sigma}^2 \sim \chi^2(2n)$.

**4.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both $\beta$ and $\sigma^2$ are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find a pivotal quantity for $\beta$ and using this pivotal quantity construct a $(1-\gamma)100\%$ confidence interval for $\beta$.

**5.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both $\beta$ and $\sigma^2$ are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find a pivotal quantity for $\sigma^2$ and using this pivotal quantity construct a $(1-\gamma)100\%$ confidence interval for $\sigma^2$.

**6.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim EXP(\beta x_i)$, where $\beta$ is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimator of $\widehat{\beta}$ of $\beta$.

**7.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim EXP(\beta x_i)$, where $\beta$ is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the least squares estimator of $\widehat{\beta}$ of $\beta$.

**8.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim POI(\beta x_i)$, where $\beta$ is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the ob-

served values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimator of $\widehat{\beta}$ of $\beta$.

**9.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim POI(\beta x_i)$, where $\beta$ is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the least squares estimator of $\widehat{\beta}$ of $\beta$.

**10.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim POI(\beta x_i)$, where $\beta$ is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, show that the least squares estimator and the maximum likelihood estimator of $\beta$ are both unbiased estimator of $\beta$.

**11.** Let $Y_1, Y_2, ..., Y_n$ be $n$ independent random variables such that each $Y_i \sim POI(\beta x_i)$, where $\beta$ is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, the find the variances of both the least squares estimator and the maximum likelihood estimator of $\beta$.

**12.** Given the five pairs of points $(x, y)$ shown below:

| $x$ | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| $y$ | 50.071 | 0.078 | 0.112 | 0.120 | 0.131 |

What is the curve of the form $y = a + bx + cx^2$ best fits the data by method of least squares?

**13.** Given the five pairs of points $(x, y)$ shown below:

| $x$ | 4 | 7 | 9 | 10 | 11 |
|---|---|---|---|---|---|
| $y$ | 10 | 16 | 22 | 20 | 25 |

What is the curve of the form $y = a + b\,x$ best fits the data by method of least squares?

**14.** The following data were obtained from the grades of six students selected at random:

| Mathematics Grade, $x$ | 72 | 94 | 82 | 74 | 65 | 85 |
|---|---|---|---|---|---|---|
| English Grade, $y$ | 76 | 86 | 65 | 89 | 80 | 92 |

Find the sample correlation coefficient $r$ and then test the null hypothesis $H_o : \rho = 0$ against the alternative hypothesis $H_a : \rho \neq 0$ at a significance level 0.01.

**15.** Given a set of data $\{(x_1, y_2), (x_2, y_2), ..., (x_n, y_n)\}$ what is the least square estimate of $\alpha$ if $y = \alpha$ is fitted to this data set.

**16.** Given a set of data points $\{(2, 3), (4, 6), (5, 7)\}$ what is the curve of the form $y = \alpha + \beta\, x^2$ best fits the data by method of least squares?

**17.** Given a data set $\{(1, 1), (2, 1), (2, 3), (3, 2), (4, 3)\}$ and $Y_x \sim N(\alpha + \beta\, x,\ \sigma^2)$, find the point estimate of $\sigma^2$ and then construct a 90% confidence interval for $\sigma$.

**18.** For the data set $\{(1, 1), (2, 1), (2, 3), (3, 2), (4, 3)\}$ determine the correlation coefficient $r$. Test the null hypothesis $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ at a significance level 0.01.

# Chapter 20
# ANALYSIS OF VARIANCE

In Chapter 19, we examine how a quantitative independent variable $x$ can be used for predicting the value of a quantitative dependent variable $y$. In this chapter we would like to examine whether one or more independent (or predictor) variable affects a dependent (or response) variable $y$. This chapter differs from the last chapter because the independent variable may now be either quantitative or qualitative. It also differs from the last chapter in assuming that the response measurements were obtained for specific settings of the independent variables. Selecting the settings of the independent variables is another aspect of experimental design. It enables us to tell whether changes in the independent variables cause changes in the mean response and it permits us to analyze the data using a method known as analysis of variance (or ANOVA). Sir Ronald Aylmer Fisher (1890-1962) developed the analysis of variance in 1920's and used it to analyze data from agricultural experiments.

The ANOVA investigates independent measurements from several treatments or levels of one or more than one factors (that is, the predictor variables). The technique of ANOVA consists of partitioning the total sum of squares into component sum of squares due to different factors and the error. For instance, suppose there are $Q$ factors. Then the total sum of squares ($\mathrm{SS_T}$) is partitioned as

$$\mathrm{SS_T} = \mathrm{SS_A} + \mathrm{SS_B} + \cdots + \mathrm{SS_Q} + \mathrm{SS_{Error}},$$

where $\mathrm{SS_A}$, $\mathrm{SS_B}$, ..., and $\mathrm{SS_Q}$ represent the sum of squares associated with the factors A, B, ..., and Q, respectively. If the ANOVA involves only one factor, then it is called one-way analysis of variance. Similarly if it involves two factors, then it is called the two-way analysis of variance. If it involves

more then two factors, then the corresponding ANOVA is called the higher order analysis of variance. In this chapter we only treat the one-way analysis of variance.

The analysis of variance is a special case of the linear models that represent the relationship between a continuous response variable $y$ and one or more predictor variables (either continuous or categorical) in the form

$$y = X\,\beta + \epsilon \tag{1}$$

where $y$ is an $m \times 1$ vector of observations of response variable, $X$ is the $m \times n$ design matrix determined by the predictor variables, $\beta$ is $n \times 1$ vector of parameters, and $\epsilon$ is an $m \times 1$ vector of random error (or disturbances) independent of each other and having distribution.

## 20.1. One-Way Analysis of Variance with Equal Sample Sizes

The standard model of one-way ANOVA is given by

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{for } i = 1, 2, ..., m, \quad j = 1, 2, ..., n, \tag{2}$$

where $m \geq 2$ and $n \geq 2$. In this model, we assume that each random variable

$$Y_{ij} \sim N(\mu_i, \sigma^2) \qquad \text{for } i = 1, 2, ..., m, \quad j = 1, 2, ..., n. \tag{3}$$

Note that because of (3), each $\epsilon_{ij}$ in model (2) is normally distributed with mean zero and variance $\sigma^2$.

Given $m$ independent samples, each of size $n$, where the members of the $i^{\text{th}}$ sample, $Y_{i1}, Y_{i2}, ..., Y_{in}$, are normal random variables with mean $\mu_i$ and unknown variance $\sigma^2$. That is,

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right), \qquad i = 1, 2, ..., m, \qquad j = 1, 2, ..., n.$$

We will be interested in testing the null hypothesis

$$\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

against the alternative hypothesis

$$\mathrm{H_a} : \text{not all the means are equal.}$$

In the following theorem we present the maximum likelihood estimators of the parameters $\mu_1, \mu_2, ..., \mu_m$ and $\sigma^2$.

**Theorem 20.1.** Suppose the one-way ANOVA model is given by the equation (2) where the $\epsilon_{ij}$'s are independent and normally distributed random variables with mean zero and variance $\sigma^2$ for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Then the MLE's of the parameters $\mu_i$ $(i = 1, 2, ..., m)$ and $\sigma^2$ of the model are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{nm}\mathrm{SS_W},$$

where $\overline{Y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$ and $\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$ is the within samples sum of squares.

**Proof:** The likelihood function is given by

$$L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_{ij}-\mu_i)^2}{2\sigma^2}}\right\}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{nm} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu_i)^2}.$$

Taking the natural logarithm of the likelihood function $L$, we obtain

$$\ln L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = -\frac{nm}{2}\ln(2\,\pi\,\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu_i)^2. \tag{4}$$

Now taking the partial derivative of (4) with respect to $\mu_1$, $\mu_2, ..., \mu_m$ and $\sigma^2$, we get

$$\frac{\partial ln L}{\partial \mu_i} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(Y_{ij}-\mu_i) \tag{5}$$

and

$$\frac{\partial ln L}{\partial \sigma^2} = -\frac{nm}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu_i)^2. \tag{6}$$

Equating these partial derivatives to zero and solving for $\mu_i$ and $\sigma^2$, respectively, we have

$$\mu_i = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\sigma^2 = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2,$$

where

$$\overline{Y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}.$$

It can be checked that these solutions yield the maximum of the likelihood function and we leave this verification to the reader. Thus the maximum likelihood estimators of the model parameters are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{nm}\mathrm{SS_W},$$

where $\mathrm{SS_W} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. The proof of the theorem is now complete.

Define

$$\overline{Y}_{\bullet\bullet} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}. \tag{7}$$

Further, define

$$\mathrm{SS_T} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{8}$$

$$\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \tag{9}$$

and

$$\mathrm{SS_B} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{10}$$

Here $\mathrm{SS_T}$ is the total sum of square, $\mathrm{SS_W}$ is the within sum of square, and $\mathrm{SS_B}$ is the between sum of square.

Next we consider the partitioning of the total sum of squares. The following lemma gives us such a partition.

**Lemma 20.1.** The total sum of squares is equal to the sum of within and between sum of squares, that is

$$\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}. \tag{11}$$

**Proof:** Rewriting (8) we have

$$\begin{aligned}
\mathrm{SS_T} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij}-\overline{Y}_{\bullet\bullet}\right)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[(Y_{ij}-\overline{Y}_{i\bullet})+(Y_{i\bullet}-\overline{Y}_{\bullet\bullet})\right]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})^2+\sum_{i=1}^{m}\sum_{j=1}^{n}(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})^2 \\
&\qquad\qquad +2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}) \\
&= \mathrm{SS_W}+\mathrm{SS_B}+2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet}).
\end{aligned}$$

The cross-product term vanishes, that is

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})=\sum_{i=1}^{m}(Y_{i\bullet}-Y_{\bullet\bullet})\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})=0.$$

Hence we obtain the asserted result $\mathrm{SS_T}=\mathrm{SS_W}+\mathrm{SS_B}$ and the proof of the lemma is complete.

The following theorem is a technical result and is needed for testing the null hypothesis against the alternative hypothesis.

**Theorem 20.2.** Consider the ANOVA model

$$Y_{ij}=\mu_i+\epsilon_{ij}\qquad i=1,2,...,m,\quad j=1,2,...,n,$$

where $Y_{ij}\sim N\left(\mu_i,\sigma^2\right)$. Then

(a) the random variable $\frac{\mathrm{SS_W}}{\sigma^2}\sim\chi^2\left(m(n-1)\right)$, and

(b) the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Further, if the null hypothesis $\mathrm{H_o}:\mu_1=\mu_2=\cdots=\mu_m=\mu$ is true, then

(c) the random variable $\frac{\mathrm{SS_B}}{\sigma^2}\sim\chi^2(m-1)$,

(d) the statistics $\frac{\mathrm{SS_B}\,m(n-1)}{\mathrm{SS_W}(m-1)}\sim F(m-1,m(n-1))$, and

(e) the random variable $\frac{\mathrm{SS_T}}{\sigma^2}\sim\chi^2(nm-1)$.

**Proof:** In Chapter 13, we have seen in Theorem 13.7 that if $X_1, X_2, ..., X_n$ are independent random variables each one having the distribution $N(\mu, \sigma^2)$, then their mean $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ have the following properties:

(i) $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ are independent, and

(ii) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$.

Now using (i) and (ii), we establish this theorem.

(a) Using (ii), we see that

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(n-1)$$

for each $i = 1, 2, ..., m$. Since

$$\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \sum_{j=1}^{n}\left(Y_{i'j} - \overline{Y}_{i'\bullet}\right)^2$$

are independent for $i' \neq i$, we obtain

$$\sum_{i=1}^{m}\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(m(n-1)).$$

Hence

$$\begin{aligned}\frac{\mathrm{SS_W}}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \\ &= \sum_{i=1}^{m}\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(m(n-1)).\end{aligned}$$

(b) Since for each $i = 1, 2, ..., m$, the random variables $Y_{i1}, Y_{i2}, ..., Y_{in}$ are independent and

$$Y_{i1}, Y_{i2}, ..., Y_{in} \sim N\left(\mu_i, \sigma^2\right)$$

we conclude by (i) that

$$\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \overline{Y}_{i\bullet}$$

are independent. Further

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \overline{Y}_{i'\bullet}$$

are independent for $i' \neq i$. Therefore, each of the statistics

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m$$

is independent of the statistics $\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet}, ..., \overline{Y}_{m\bullet}$, and the statistics

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m$$

are independent. Thus it follows that the sets

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m \qquad \text{and} \qquad \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m$$

are independent. Thus

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2$$

are independent. Hence by definition, the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Suppose the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true.

(c) Under $\mathrm{H_o}$, the random variables $\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet}, ..., \overline{Y}_{m\bullet}$ are independent and identically distributed with $N\left(\mu, \frac{\sigma^2}{n}\right)$. Therefore by (ii)

$$\frac{n}{\sigma^2} \sum_{i=1}^{m} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(m-1).$$

Hence

$$\begin{aligned} \frac{\mathrm{SS_B}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \\ &= \frac{n}{\sigma^2} \sum_{i=1}^{m} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(m-1). \end{aligned}$$

(d) Since

$$\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\left(m(n-1)\right)$$

and

$$\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$$

therefore

$$\frac{\frac{\mathrm{SS_B}}{(m-1)\,\sigma^2}}{\frac{\mathrm{SS_W}}{(n(m-1)\,\sigma^2}} \sim F(m-1, m(n-1)).$$

That is

$$\frac{\frac{\mathrm{SS_B}}{(m-1)}}{\frac{\mathrm{SS_W}}{(n(m-1)}} \sim F(m-1, m(n-1)).$$

(e) Under $\mathrm{H_o}$, the random variables $Y_{ij}$, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$ are independent and each has the distribution $N(\mu, \sigma^2)$. By (ii) we see that

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(nm-1).$$

Hence we have

$$\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(nm-1)$$

and the proof of the theorem is now complete.

From Theorem 20.1, we see that the maximum likelihood estimator of each $\mu_i$ $(i = 1, 2, ..., m)$ is given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet},$$

and since $\overline{Y}_{i\bullet} \sim N\left(\mu_i, \frac{\sigma^2}{n}\right)$,

$$E\left(\widehat{\mu_i}\right) = E\left(\overline{Y}_{i\bullet}\right) = \mu_i.$$

Thus the maximum likelihood estimators are unbiased estimator of $\mu_i$ for $i = 1, 2, ..., m$.

Since

$$\widehat{\sigma^2} = \frac{\mathrm{SS_W}}{mn}$$

and by Theorem 20.2, $\frac{1}{\sigma^2}\mathrm{SS_W} \sim \chi^2(m(n-1))$, we have

$$E\left(\widehat{\sigma^2}\right) = E\left(\frac{\mathrm{SS_W}}{mn}\right) = \frac{1}{mn}\,\sigma^2\,E\left(\frac{1}{\sigma^2}\mathrm{SS_W}\right) = \frac{1}{mn}\,\sigma^2\,m(n-1) \neq \sigma^2.$$

Thus the maximum likelihood estimator $\widehat{\sigma^2}$ of $\sigma^2$ is biased. However, the estimator $\frac{\mathrm{SS_W}}{m(n-1)}$ is an unbiased estimator. Similarly, the estimator $\frac{\mathrm{SS_T}}{mn-1}$ is an unbiased estimator where as $\frac{\mathrm{SS_T}}{mn}$ is a biased estimator of $\sigma^2$.

**Theorem 20.3.** Suppose the one-way ANOVA model is given by the equation (2) where the $\epsilon_{ij}$'s are independent and normally distributed random variables with mean zero and variance $\sigma^2$ for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. The null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is rejected whenever the test statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(m(n-1))} > F_\alpha(m-1,\, m(n-1)), \tag{12}$$

where $\alpha$ is the significance level of the hypothesis test and $F_\alpha(m-1,\, m(n-1))$ denotes the $100(1-\alpha)$ percentile of the $F$-distribution with $m-1$ numerator and $nm-m$ denominator degrees of freedom.

**Proof:** Under the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$, the likelihood function takes the form

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(Y_{ij}-\mu)^2}{2\sigma^2}}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{nm} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu)^2}. \end{aligned}$$

Taking the natural logarithm of the likelihood function and then maximizing it, we obtain

$$\widehat{\mu} = \overline{Y}_{\bullet\bullet} \qquad \text{and} \qquad \widehat{\sigma_{\mathrm{H_o}}} = \frac{1}{mn}\mathrm{SS_T}$$

as the maximum likelihood estimators of $\mu$ and $\sigma^2$, respectively. Inserting these estimators into the likelihood function, we have the maximum of the likelihood function, that is

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2_{\mathrm{H_o}}}}}\right)^{nm} e^{-\frac{1}{2\widehat{\sigma^2_{\mathrm{H_o}}}}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2}.$$

Simplifying the above expression, we see that

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2_{\mathrm{H_o}}}}}\right)^{nm} e^{-\frac{mn}{2\,\mathrm{SS_T}}\mathrm{SS_T}}$$

which is

$$\max L(\mu, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\widehat{\sigma^2_{\mathrm{H_o}}}}} \right)^{nm} e^{-\frac{mn}{2}}. \tag{13}$$

When no restrictions imposed, we get the maximum of the likelihood function from Theorem 20.1 as

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\widehat{\sigma^2}}} \right)^{nm} e^{-\frac{1}{2\widehat{\sigma^2}} \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i\bullet})^2}.$$

Simplifying the above expression, we see that

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\widehat{\sigma^2}}} \right)^{nm} e^{-\frac{mn}{2\,\mathrm{SS_W}}\,\mathrm{SS_W}}$$

which is

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\widehat{\sigma^2}}} \right)^{nm} e^{-\frac{mn}{2}}. \tag{14}$$

Next we find the likelihood ratio statistic $W$ for testing the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$. Recall that the likelihood ratio statistic $W$ can be found by evaluating

$$W = \frac{\max L(\mu, \sigma^2)}{\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2)}.$$

Using (13) and (14), we see that

$$W = \left( \frac{\widehat{\sigma^2}}{\widehat{\sigma^2_{\mathrm{H_o}}}} \right)^{\frac{mn}{2}}. \tag{15}$$

Hence the likelihood ratio test to reject the null hypothesis $\mathrm{H_o}$ is given by the inequality

$$W < k_0$$

where $k_0$ is a constant. Using (15) and simplifying, we get

$$\frac{\widehat{\sigma^2_{\mathrm{H_o}}}}{\widehat{\sigma^2}} > k_1$$

where $k_1 = \left(\frac{1}{k_0}\right)^{\frac{2}{mn}}$. Hence

$$\frac{\mathrm{SS_T}/mn}{\mathrm{SS_W}/mn} = \frac{\widehat{\sigma^2_{\mathrm{H_o}}}}{\widehat{\sigma^2}} > k_1.$$

Using Lemma 20.1 we have

$$\frac{\mathrm{SS_W} + \mathrm{SS_B}}{\mathrm{SS_W}} > k_1.$$

Therefore

$$\frac{\mathrm{SS_B}}{\mathrm{SS_W}} > k \tag{16}$$

where $k = k_1 - 1$. In order to find the cutoff point $k$ in (16), we use Theorem 20.2 (d). Therefore

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(m(n-1))} > \frac{m(n-1)}{m-1}k$$

Since $\mathcal{F}$ has $F$ distribution, we obtain

$$\frac{m(n-1)}{m-1}k = F_\alpha(m-1,\, m(n-1)).$$

Thus, at a significance level $\alpha$, reject the null hypothesis $\mathrm{H_o}$ if

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(m(n-1))} > F_\alpha(m-1,\, m(n-1))$$

and the proof of the theorem is complete.

The various quantities used in carrying out the test described in Theorem 20.3 are presented in a tabular form known as the ANOVA table.

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | $\mathrm{SS_B}$ | $m-1$ | $\mathrm{MS_B} = \frac{\mathrm{SS_B}}{m-1}$ | $\mathcal{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$ |
| Within | $\mathrm{SS_W}$ | $m(n-1)$ | $\mathrm{MS_W} = \frac{\mathrm{SS_W}}{m(n-1)}$ | |
| Total | $\mathrm{SS_T}$ | $mn-1$ | | |

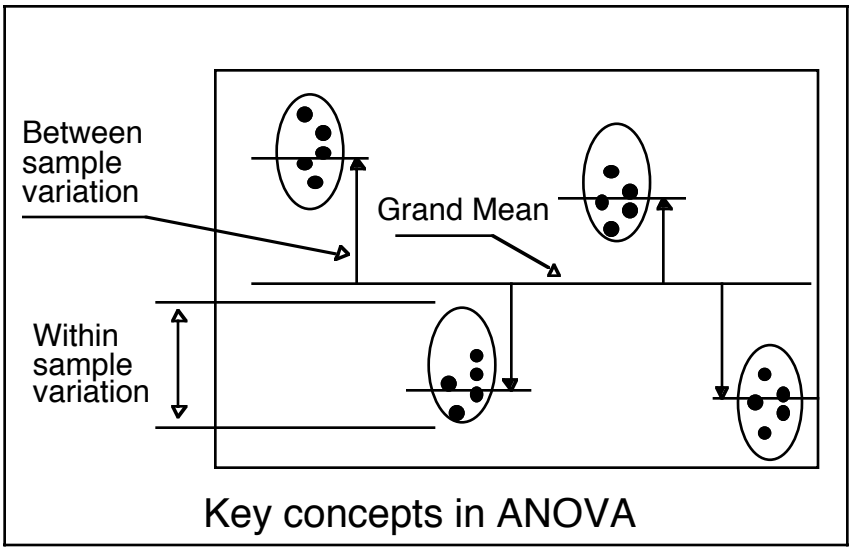
**Table 20.1.** One-Way ANOVA Table

At a significance level $\alpha$, the likelihood ratio test is: "Reject the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ if $\mathcal{F} > F_\alpha(m-1,\, m(n-1))$." One can also use the notion of $p-$value to perform this hypothesis test. If the value of the test statistics is $\mathcal{F} = \gamma$, then the $p$-value is defined as

$$p - \text{value} = P(F(m-1, m(n-1)) \geq \gamma).$$

Alternatively, at a significance level $\alpha$, the likelihood ratio test is: "Reject the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ if $p - \text{value} < \alpha$."

The following figure illustrates the notions of between sample variation and within sample variation.




The ANOVA model described in (2), that is

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

can be rewritten as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where $\mu$ is the mean of the $m$ values of $\mu_i$, and $\sum_{i=1}^{m} \alpha_i = 0$. The quantity $\alpha_i$ is called the effect of the $i^{\text{th}}$ treatment. Thus any observed value is the sum of

an overall mean $\mu$, a treatment or class deviation $\alpha_i$, and a random element from a normally distributed random variable $\epsilon_{ij}$ with mean zero and variance $\sigma^2$. This model is called model I, the fixed effects model. The effects of the treatments or classes, measured by the parameters $\alpha_i$, are regarded as fixed but unknown quantities to be estimated. In this fixed effect model the null hypothesis $\mathrm{H}_0$ is now

$$\mathrm{H_o} : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

and the alternative hypothesis is

$$\mathrm{H_a} : \text{not all the } \alpha_\mathrm{i} \text{ are zero.}$$

The random effects model, also known as model II, is given by

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \qquad \text{for } i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where $\mu$ is the overall mean and

$$A_i \sim N(0, \sigma_A^2) \qquad \text{and} \qquad \epsilon_{ij} \sim N(0, \sigma^2).$$

In this model, the variances $\sigma_A^2$ and $\sigma^2$ are unknown quantities to be estimated. The null hypothesis of the random effect model is $H_o : \sigma_A^2 = 0$ and the alternative hypothesis is $H_a : \sigma_A^2 > 0$. In this chapter we do not consider the random effect model.

Before we present some examples, we point out the assumptions on which the ANOVA is based on. The ANOVA is based on the following three assumptions:

(1) *Independent Samples:* The samples taken from the population under consideration should be independent of one another.

(2) *Normal Population:* For each population, the variable under consideration should be normally distributed.

(3) *Equal Variance:* The variances of the variables under consideration should be the same for all the populations.

**Example 20.1.** The data in the following table gives the number of hours of relief provided by 5 different brands of headache tablets administered to 25 subjects experiencing fevers of $38^o$C or more. Perform the analysis of variance

and test the hypothesis at the 0.05 level of significance that the mean number of hours of relief provided by the tablets is same for all 5 brands.

| Tablets | | | | |
|---|---|---|---|---|
| A | B | C | D | F |
| 5 | 9 | 3 | 2 | 7 |
| 4 | 7 | 5 | 3 | 6 |
| 8 | 8 | 2 | 4 | 9 |
| 6 | 6 | 3 | 1 | 4 |
| 3 | 9 | 7 | 4 | 7 |

**Answer:** Using the formulas (8), (9) and (10), we compute the sum of squares $\mathrm{SS_W}$, $\mathrm{SS_B}$ and $\mathrm{SS_T}$ as

$$\mathrm{SS_W} = 57.60, \qquad \mathrm{SS_B} = 79.94, \qquad \text{and} \qquad \mathrm{SS_T} = 137.04.$$

The ANOVA table for this problem is shown below.

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 79.94 | 4 | 19.86 | 6.90 |
| Within | 57.60 | 20 | 2.88 | |
| Total | 137.04 | 24 | | |

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(4,\, 20) = 2.8661$. Since

$$6.90 = \mathcal{F} > F_{0.05}(4,\, 20) = 2.8661$$

we reject the null hypothesis that the mean number of hours of relief provided by the tablets is same for all 5 brands.

Note that using a statistical package like MINITAB, SAS or SPSS we can compute the $p$-value to be

$$p - \text{value} = P(F(4, 20) \geq 6.90) = 0.001.$$

Hence again we reach the same conclusion since $p$-value is less then the given $\alpha$ for this problem.

**Example 20.2.** Perform the analysis of variance and test the null hypothesis at the 0.05 level of significance for the following two data sets.

| Data Set 1 | | | Data Set 2 | | |
|---|---|---|---|---|---|
| Sample | | | Sample | | |
| A | B | C | A | B | C |
| 8.1 | 8.0 | 14.8 | 9.2 | 9.5 | 9.4 |
| 4.2 | 15.1 | 5.3 | 9.1 | 9.5 | 9.3 |
| 14.7 | 4.7 | 11.1 | 9.2 | 9.5 | 9.3 |
| 9.9 | 10.4 | 7.9 | 9.2 | 9.6 | 9.3 |
| 12.1 | 9.0 | 9.3 | 9.3 | 9.5 | 9.2 |
| 6.2 | 9.8 | 7.4 | 9.2 | 9.4 | 9.3 |

**Answer:** Computing the sum of squares $SS_W$, $SS_B$ and $SS_T$, we have the following two ANOVA tables:

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 0.3 | 2 | 0.1 | 0.01 |
| Within | 187.2 | 15 | 12.5 | |
| Total | 187.5 | 17 | | |

and

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 0.280 | 2 | 0.140 | 35.0 |
| Within | 0.600 | 15 | 0.004 | |
| Total | 0.340 | 17 | | |

At the significance level $\alpha = 0.05$, we find from the F-table that $F_{0.05}(2,\, 15) = 3.68$. For the first data set, since

$$0.01 = \mathcal{F} < F_{0.05}(2,\, 15) = 3.68$$

we do not reject the null hypothesis whereas for the second data set,

$$35.0 = \mathcal{F} > F_{0.05}(2,\, 15) = 3.68$$

we reject the null hypothesis.

**Remark 20.1.** Note that the sample means are same in both the data sets. However, there is a less variation among the sample points in samples of the second data set. The ANOVA finds a more significant differences among the means in the second data set. This example suggests that the larger the variation among sample means compared with the variation of the measurements within samples, the greater is the evidence to indicate a difference among population means.

### 20.2. One-Way Analysis of Variance with Unequal Sample Sizes

In the previous section, we examined the theory of ANOVA when samples are same sizes. When the samples are same sizes we say that the ANOVA is in the balanced case. In this section we examine the theory of ANOVA for unbalanced case, that is when the samples are of different sizes. In experimental work, one often encounters unbalance case due to the death of experimental animals in a study or drop out of the human subjects from a study or due to damage of experimental materials used in a study. Our analysis of the last section for the equal sample size will be valid but have to be modified to accommodate the different sample size.

Consider $m$ independent samples of respective sizes $n_1, n_2, ..., n_m$, where the members of the $i^{\text{th}}$ sample, $Y_{i1}, Y_{i2}, ..., Y_{in_i}$, are normal random variables with mean $\mu_i$ and unknown variance $\sigma^2$. That is,

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right), \qquad i = 1, 2, ..., m, \qquad j = 1, 2, ..., n_i.$$

Let us denote $N = n_1 + n_2 + \cdots + n_m$. Again, we will be interested in testing the null hypothesis

$$\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

against the alternative hypothesis

$$\mathrm{H_a}: \text{not all the means are equal.}$$

Now we defining

$$\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n} Y_{ij}, \tag{17}$$

$$\overline{Y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij}, \tag{18}$$

$$\mathrm{SS_T} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2, \tag{19}$$

$$\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2, \tag{20}$$

and

$$\mathrm{SS_B} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{21}$$

we have the following results analogous to the results in the previous section.

**Theorem 20.4.** Suppose the one-way ANOVA model is given by the equation (2) where the $\epsilon_{ij}$'s are independent and normally distributed random variables with mean zero and variance $\sigma^2$ for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. Then the MLE's of the parameters $\mu_i$ $(i = 1, 2, ..., m)$ and $\sigma^2$ of the model are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{N}\mathrm{SS_W},$$

where $\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and $\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$ is the within samples sum of squares.

**Lemma 20.2.** The total sum of squares is equal to the sum of within and between sum of squares, that is $\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}$.

**Theorem 20.5.** Consider the ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, ..., m, \quad j = 1, 2, ..., n_i,$$

where $Y_{ij} \sim N\left(\mu_i, \sigma^2\right)$. Then

(a) the random variable $\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\,(N-m)$, and

(b) the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Further, if the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true, then

(c) the random variable $\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$,

(d) the statistics $\frac{\mathrm{SS_B}\, m(n-1)}{\mathrm{SS_W}(m-1)} \sim F(m-1, N-m)$, and

(e) the random variable $\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(N-1)$.

**Theorem 20.6.** Suppose the one-way ANOVA model is given by the equation (2) where the $\epsilon_{ij}$'s are independent and normally distributed random variables with mean zero and variance $\sigma^2$ for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. The null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is rejected whenever the test statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(N-m)} > F_\alpha(m-1,\, N-m),$$

where $\alpha$ is the significance level of the hypothesis test and $F_\alpha(m-1,\, N-m)$ denotes the $100(1-\alpha)$ percentile of the $F$-distribution with $m-1$ numerator and $N-m$ denominator degrees of freedom.

The corresponding ANOVA table for this case is

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | $\mathrm{SS_B}$ | $m-1$ | $\mathrm{MS_B} = \frac{\mathrm{SS_B}}{m-1}$ | $\mathcal{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$ |
| Within | $\mathrm{SS_W}$ | $N-m$ | $\mathrm{MS_W} = \frac{\mathrm{SS_W}}{N-m}$ | |
| Total | $\mathrm{SS_T}$ | $N-1$ | | |

**Table 20.2.** One-Way ANOVA Table with unequal sample size

**Example 20.3.** Three sections of elementary statistics were taught by different instructors. A common final examination was given. The test scores are given in the table below. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that there is a difference in the average grades given by the three instructors.

| Elementary Statistics | | |
|---|---|---|
| Instructor A | Instructor B | Instructor C |
| 75 | 90 | 17 |
| 91 | 80 | 81 |
| 83 | 50 | 55 |
| 45 | 93 | 70 |
| 82 | 53 | 61 |
| 75 | 87 | 43 |
| 68 | 76 | 89 |
| 47 | 82 | 73 |
| 38 | 78 | 58 |
| | 80 | 70 |
| | 33 | |
| | 79 | |

**Answer:** Using the formulas (17) - (21), we compute the sum of squares $SS_W$, $SS_B$ and $SS_T$ as

$$SS_W = 10362, \qquad SS_B = 755, \qquad \text{and} \qquad SS_T = 11117.$$

The ANOVA table for this problem is shown below.

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 755 | 2 | 377 | 1.02 |
| Within | 10362 | 28 | 370 | |
| Total | 11117 | 30 | | |

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2, 28) = 3.34$. Since

$$1.02 = \mathcal{F} < F_{0.05}(2, 28) = 3.34$$

we accept the null hypothesis that there is no difference in the average grades given by the three instructors.

Note that using a statistical package like MINITAB, SAS or SPSS we can compute the $p$-value to be

$$p - \text{value} = P(F(2, 28) \geq 1.02) = 0.374.$$

Hence again we reach the same conclusion since $p$-value is less then the given $\alpha$ for this problem.

We conclude this section pointing out the advantages of choosing equal sample sizes (balance case) over the choice of unequal sample sizes (unbalance case). The first advantage is that the $\mathcal{F}$-statistics is insensitive to slight departures from the assumption of equal variances when the sample sizes are equal. The second advantage is that the choice of equal sample size minimizes the probability of committing a type II error.

### 20.3. Pair wise Comparisons

When the null hypothesis is rejected using the $F$-test in ANOVA, one may still wants to know where the difference among the means is. There are several methods to find out where the significant differences in the means lie after the ANOVA procedure is performed. Among the most commonly used tests are Scheffé test and Tuckey test. In this section, we give a brief description of these tests.

In order to perform the Scheffé test, we have to compare the means two at a time using all possible combinations of means. Since we have $m$ means, we need $\binom{m}{2}$ pair wise comparisons. A pair wise comparison can be viewed as a test of the null hypothesis $H_0 : \mu_i = \mu_k$ against the alternative $H_a : \mu_i \neq \mu_k$ for all $i \neq k$.

To conduct this test we compute the statistics

$$F_s = \frac{\left(\overline{Y}_{i\bullet} - \overline{Y}_{k\bullet}\right)^2}{MS_W \left(\frac{1}{n_i} + \frac{1}{n_k}\right)},$$

where $\overline{Y}_{i\bullet}$ and $\overline{Y}_{k\bullet}$ are the means of the samples being compared, $n_i$ and $n_k$ are the respective sample sizes, and $MS_W$ is the mean sum of squared of within group. We reject the null hypothesis at a significance level of $\alpha$ if

$$F_s > (m-1)F_\alpha(m-1, N-m)$$

where $N = n_1 + n_2 + \cdots + n_m$.

**Example 20.4.** Perform the analysis of variance and test the null hypothesis at the 0.05 level of significance for the following data given in the table below. Further perform a Scheffé test to determine where the significant differences in the means lie.

| Sample | | |
|---|---|---|
| 1 | 2 | 3 |
| 9.2 | 9.5 | 9.4 |
| 9.1 | 9.5 | 9.3 |
| 9.2 | 9.5 | 9.3 |
| 9.2 | 9.6 | 9.3 |
| 9.3 | 9.5 | 9.2 |
| 9.2 | 9.4 | 9.3 |

**Answer:** The ANOVA table for this data is given by

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 0.280 | 2 | 0.140 | 35.0 |
| Within | 0.600 | 15 | 0.004 | |
| Total | 0.340 | 17 | | |

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2, 15) = 3.68$. Since

$$35.0 = \mathcal{F} > F_{0.05}(2, 15) = 3.68$$

we reject the null hypothesis. Now we perform the Scheffé test to determine where the significant differences in the means lie. From given data, we obtain $\overline{Y}_{1\bullet} = 9.2$, $\overline{Y}_{2\bullet} = 9.5$ and $\overline{Y}_{3\bullet} = 9.3$. Since $m = 3$, we have to make 3 pair wise comparisons, namely $\mu_1$ with $\mu_2$, $\mu_1$ with $\mu_3$, and $\mu_2$ with $\mu_3$. First we consider the comparison of $\mu_1$ with $\mu_2$. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}\right)^2}{MS_W \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{(9.2 - 9.5)^2}{0.004 \left(\frac{1}{6} + \frac{1}{6}\right)} = 67.5.$$

Since

$$67.5 = F_s > 2\, F_{0.05}(2, 15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_2$ in favor of the alternative $H_a : \mu_1 \neq \mu_2$.

Next we consider the comparison of $\mu_1$ with $\mu_3$. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{1\bullet} - \overline{Y}_{3\bullet}\right)^2}{MS_W\left(\frac{1}{n_1} + \frac{1}{n_3}\right)} = \frac{(9.2 - 9.3)^2}{0.004\left(\frac{1}{6} + \frac{1}{6}\right)} = 7.5.$$

Since

$$7.5 = F_s > 2\, F_{0.05}(2,\, 15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_3$ in favor of the alternative $H_a : \mu_1 \neq \mu_3$.

Finally we consider the comparison of $\mu_2$ with $\mu_3$. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}\right)^2}{MS_W\left(\frac{1}{n_2} + \frac{1}{n_3}\right)} = \frac{(9.5 - 9.3)^2}{0.004\left(\frac{1}{6} + \frac{1}{6}\right)} = 30.0.$$

Since

$$30.0 = F_s > 2\, F_{0.05}(2,\, 15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_3$ in favor of the alternative $H_a : \mu_2 \neq \mu_3$.

Next consider the Tukey test. Tuckey test is applicable when we have a balanced case, that is when the sample sizes are equal. For Tukey test we compute the statistics

$$Q = \frac{\overline{Y}_{i\bullet} - \overline{Y}_{k\bullet}}{\sqrt{\frac{MS_W}{n}}},$$

where $\overline{Y}_{i\bullet}$ and $\overline{Y}_{k\bullet}$ are the means of the samples being compared, $n$ is the size of the samples, and $MS_W$ is the mean sum of squared of within group. At a significance level $\alpha$, we reject the null hypothesis $H_0$ if

$$|Q| > Q_\alpha(m, \nu)$$

where $\nu$ represents the degrees of freedom for the error mean square.

**Example 20.5.** For the data given in Example 20.4 perform a Tukey test to determine where the significant differences in the means lie.

**Answer:** We have seen that $\overline{Y}_{1\bullet} = 9.2$, $\overline{Y}_{2\bullet} = 9.5$ and $\overline{Y}_{3\bullet} = 9.3$.

First we compare $\mu_1$ with $\mu_2$. For this we compute

$$Q = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.2 - 9.3}{\sqrt{\frac{0.004}{6}}} = -11.6189.$$

Since

$$11.6189 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_2$ in favor of the alternative $H_a : \mu_1 \neq \mu_2$.

Next we compare $\mu_1$ with $\mu_3$. For this we compute

$$Q = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{3\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.2 - 9.5}{\sqrt{\frac{0.004}{6}}} = -3.8729.$$

Since

$$3.8729 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_3$ in favor of the alternative $H_a : \mu_1 \neq \mu_3$.

Finally we compare $\mu_2$ with $\mu_3$. For this we compute

$$Q = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.5 - 9.3}{\sqrt{\frac{0.004}{6}}} = 7.7459.$$

Since

$$7.7459 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_3$ in favor of the alternative $H_a : \mu_2 \neq \mu_3$.

Often in scientific and engineering problems, the experiment dictates the need for comparing simultaneously each treatment with a control. Now we describe a test developed by C. W. Dunnett for determining significant differences between each treatment mean and the control. Suppose we wish to test the $m$ hypotheses

$$H_0 : \mu_0 = \mu_i \quad \text{versus} \quad H_a : \mu_0 \neq \mu_i \qquad \text{for } i = 1, 2, ..., m,$$

where $\mu_0$ represents the mean yield for the population of measurements in which the control is used. To test the null hypotheses specified by $H_0$ against two-sided alternatives for an experimental situation in which there are $m$ treatments, excluding the control, and $n$ observation per treatment, we first calculate

$$D_i = \frac{\overline{Y}_{i\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\, MS_W}{n}}}, \quad i = 1, 2, ..., m.$$

At a significance level $\alpha$, we reject the null hypothesis $H_0$ if

$$|D_i| > D_{\frac{\alpha}{2}}(m, \nu)$$

where $\nu$ represents the degrees of freedom for the error mean square. The values of the quantity $D_{\frac{\alpha}{2}}(m, \nu)$ are tabulated for various $\alpha$, $m$ and $\nu$.

**Example 20.6.** For the data given in the table below perform a Dunnett test to determine any significant differences between each treatment mean and the control.

| Control | Sample 1 | Sample 2 |
|---|---|---|
| 9.2 | 9.5 | 9.4 |
| 9.1 | 9.5 | 9.3 |
| 9.2 | 9.5 | 9.3 |
| 9.2 | 9.6 | 9.3 |
| 9.3 | 9.5 | 9.2 |
| 9.2 | 9.4 | 9.3 |

**Answer:** The ANOVA table for this data is given by

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 0.280 | 2 | 0.140 | 35.0 |
| Within | 0.600 | 15 | 0.004 | |
| Total | 0.340 | 17 | | |

At the significance level $\alpha = 0.05$, we find that $D_{0.025}(2, 15) = 2.44$. Since

$$35.0 = D > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis. Now we perform the Dunnett test to determine if there is any significant differences between each treatment mean and the control. From given data, we obtain $\overline{Y}_{0\bullet} = 9.2$, $\overline{Y}_{1\bullet} = 9.5$ and $\overline{Y}_{2\bullet} = 9.3$. Since $m = 2$, we have to make 2 pair wise comparisons, namely $\mu_0$ with $\mu_1$, and $\mu_0$ with $\mu_2$. First we consider the comparison of $\mu_0$ with $\mu_1$. For this case, we find

$$D_1 = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\,MS_W}{n}}} = \frac{9.5 - 9.2}{\sqrt{\frac{2\,(0.004)}{6}}} = 8.2158.$$

Since

$$8.2158 = D_1 > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_0$ in favor of the alternative $H_a : \mu_1 \neq \mu_0$.

Next we find

$$D_2 = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\,MS_W}{n}}} = \frac{9.3 - 9.2}{\sqrt{\frac{2\,(0.004)}{6}}} = 2.7386.$$

Since

$$2.7386 = D_2 > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_0$ in favor of the alternative $H_a : \mu_2 \neq \mu_0$.

### 20.4. Tests for the Homogeneity of Variances

One of the assumptions behind the ANOVA is the equal variance, that is the variances of the variables under consideration should be the same for all population. Earlier we have pointed out that the $\mathcal{F}$-statistics is insensitive to slight departures from the assumption of equal variances when the sample sizes are equal. Nevertheless it is advisable to run a preliminary test for homogeneity of variances. Such a test would certainly be advisable in the case of unequal sample sizes if there is a doubt concerning the homogeneity of population variances.

Suppose we want to test the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$$

versus the alternative hypothesis

$$H_a : \text{not all variances are equal.}$$

A frequently used test for the homogeneity of population variances is the Bartlett test. Bartlett (1937) proposed a test for equal variances that was modification of the normal-theory likelihood ratio test.

We will use this test to test the above null hypothesis $H_0$ against $H_a$. First, we compute the $m$ sample variances $S_1^2, S_2^2, ..., S_m^2$ from the samples of

size $n_1, n_2, ..., n_m$, with $n_1 + n_2 + \cdots + n_m = N$. The test statistics $B_c$ is given by

$$B_c = \frac{(N-m)\,\ln S_p^2 - \sum_{i=1}^{m}(n_i - 1)\,\ln S_i^2}{1 + \frac{1}{3\,(m-1)}\left(\sum_{i=1}^{m}\frac{1}{n_i - 1} - \frac{1}{N-m}\right)}$$

where the pooled variance $S_p^2$ is given by

$$S_p^2 = \frac{\sum_{i=1}^{m}(n_i - 1)\,S_i^2}{N-m} = \mathrm{MS_W}.$$

It is known that the sampling distribution of $B_c$ is approximately chi-square with $m-1$ degrees of freedom, that is

$$B_c \sim \chi^2(m-1)$$

when $(n_i - 1) \geq 3$. Thus the Bartlett test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$ at a significance level $\alpha$ if

$$B_c > \chi^2_{1-\alpha}(m-1),$$

where $\chi^2_{1-\alpha}(m-1)$ denotes the upper $(1-\alpha)100$ percentile of the chi-square distribution with $m-1$ degrees of freedom.

**Example 20.7.** For the following data perform an ANOVA and then apply Bartlett test to examine if the homogeneity of variances condition is met for a significance level 0.05.

| Data | | | |
|---|---|---|---|
| Sample 1 | Sample 2 | Sample 3 | Sample 4 |
| 34 | 29 | 32 | 34 |
| 28 | 32 | 34 | 29 |
| 29 | 31 | 30 | 32 |
| 37 | 43 | 42 | 28 |
| 42 | 31 | 32 | 32 |
| 27 | 29 | 33 | 34 |
| 29 | 28 | 29 | 29 |
| 35 | 30 | 27 | 31 |
| 25 | 37 | 37 | 30 |
| 29 | 44 | 26 | 37 |
| 41 | 29 | 29 | 43 |
| 40 | 31 | 31 | 42 |

**Answer:** The ANOVA table for this data is given by

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 16.2 | 3 | 5.4 | 0.20 |
| Within | 1202.2 | 44 | 27.3 | |
| Total | 1218.5 | 47 | | |

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2,\, 44) = 3.23$. Since

$$0.20 = \mathcal{F} < F_{0.05}(2,\, 44) = 3.23$$

we do not reject the null hypothesis.

Now we compute Bartlett test statistic $B_c$. From the data the variances of each group can be found to be

$$S_1^2 = 35.2836, \qquad S_2^2 = 30.1401, \qquad S_3^2 = 19.4481, \qquad S_4^2 = 24.4036.$$

Further, the pooled variance is

$$S_p^2 = \mathrm{MS}_W = 27.3.$$

The statistics $B_c$ is

$$\begin{aligned} B_c &= \frac{(N-m)\,\ln S_p^2 - \sum_{i=1}^{m}(n_i-1)\,\ln S_i^2}{1+\frac{1}{3\,(m-1)}\left(\sum_{i=1}^{m}\frac{1}{n_i-1}-\frac{1}{N-m}\right)} \\ &= \frac{44\,\ln 27.3 - 11\,[\ln 35.2836 - \ln 30.1401 - \ln 19.4481 - \ln 24.4036\,]}{1+\frac{1}{3\,(4-1)}\left(\frac{4}{12-1}-\frac{1}{48-4}\right)} \\ &= \frac{1.0537}{1.0378} = 1.0153. \end{aligned}$$

From chi-square table we find that $\chi^2_{0.95}(3) = 7.815$. Hence, since

$$1.0153 = B_c < \chi^2_{0.95}(3) = 7.815,$$

we do not reject the null hypothesis that the variances are equal. Hence Bartlett test suggests that the homogeneity of variances condition is met.

The Bartlett test assumes that the $m$ samples should be taken from $m$ normal populations. Thus Bartlett test is sensitive to departures from normality. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality. Levene (1960) proposed a test for the homogeneity of population variances that considers the random variables

$$W_{ij} = \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$$

and apply a one-way analysis of variance to these variables. If the $F$-test is significant, the homogeneity of variances is rejected.

Levene (1960) also proposed using $F$-tests based on the variables

$$W_{ij} = |Y_{ij} - \overline{Y}_{i\bullet}|, \qquad W_{ij} = \ln|Y_{ij} - \overline{Y}_{i\bullet}|, \qquad \text{and} \quad W_{ij} = \sqrt{|Y_{ij} - \overline{Y}_{i\bullet}|}.$$

Brown and Forsythe (1974c) proposed using the transformed variables based on the absolute deviations from the median, that is $W_{ij} = |Y_{ij} - Med(Y_{i\bullet})|$, where $Med(Y_{i\bullet})$ denotes the median of group $i$. Again if the $F$-test is significant, the homogeneity of variances is rejected.

**Example 20.8.** For the data in Example 20.7 do a Levene test to examine if the homogeneity of variances condition is met for a significance level 0.05.

**Answer:** From data we find that $\overline{Y}_{1\bullet} = 33.00$, $\overline{Y}_{2\bullet} = 32.83$, $\overline{Y}_{3\bullet} = 31.83$, and $\overline{Y}_{4\bullet} = 33.42$. Next we compute $W_{ij} = \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. The resulting values are given in the table below.

| Transformed Data | | | |
|---|---|---|---|
| Sample 1 | Sample 2 | Sample 3 | Sample 4 |
| 1 | 14.7 | 0.0 | 0.3 |
| 25 | 0.7 | 4.7 | 19.5 |
| 16 | 3.4 | 3.4 | 2.0 |
| 16 | 103.4 | 103.4 | 29.3 |
| 81 | 3.4 | 0.0 | 2.0 |
| 36 | 14.7 | 1.4 | 0.3 |
| 16 | 23.4 | 8.0 | 19.5 |
| 4 | 8.0 | 23.4 | 5.8 |
| 64 | 17.4 | 26.7 | 11.7 |
| 16 | 124.7 | 34.0 | 12.8 |
| 64 | 14.7 | 0.0 | 91.8 |
| 49 | 3.4 | 0.7 | 73.7 |

Now we perform an ANOVA to the data given in the table above. The ANOVA table for this data is given by

| Source of variation | Sums of squares | Degree of freedom | Mean squares | F-statistics $\mathcal{F}$ |
|---|---|---|---|---|
| Between | 1430 | 3 | 477 | 0.46 |
| Within | 45491 | 44 | 1034 | |
| Total | 46922 | 47 | | |

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(3, 44) = 2.84$. Since

$$0.46 = \mathcal{F} < F_{0.05}(3, 44) = 2.84$$

we do not reject the null hypothesis that the variances are equal. Hence Bartlett test suggests that the homogeneity of variances condition is met.

Although Bartlet test is most widely used test for homogeneity of variances a test due to Cochran provides a computationally simple procedure. Cochran test is one of the best method for detecting cases where the variance of one of the groups is much larger than that of the other groups. The test statistics of Cochran test is give by

$$C = \frac{\max\limits_{1\leq i\leq m} S_i^2}{\sum\limits_{i=1}^{m} S_i^2}.$$

The Cochran test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$ at a significance level $\alpha$ if

$$C > C_\alpha.$$

The critical values of $C_\alpha$ were originally published by Eisenhart *et al* (1947) for some combinations of degrees of freedom $\nu$ and the number of groups $m$. Here the degrees of freedom $\nu$ are

$$\nu = \max_{1\leq i\leq m} (n_i - 1).$$

**Example 20.9.** For the data in Example 20.7 perform a Cochran test to examine if the homogeneity of variances condition is met for a significance level 0.05.

**Answer:** From the data the variances of each group can be found to be

$$S_1^2 = 35.2836, \qquad S_2^2 = 30.1401, \qquad S_3^2 = 19.4481, \qquad S_4^2 = 24.4036.$$

Hence the test statistic for Cochran test is

$$C = \frac{35.2836}{35.2836 + 30.1401 + 19.4481 + 24.4036} = \frac{35.2836}{109.2754} = 0.3328.$$

The critical value $C_{0.5}(3, 11)$ is given by 0.4884. Since

$$0.3328 = C < C_{0.5}(3, 11) = 0.4884.$$

At a significance level $\alpha = 0.05$, we do not reject the null hypothesis that the variances are equal. Hence Cochran test suggests that the homogeneity of variances condition is met.

### 20.5. Exercises

**1.** A consumer organization wants to compare the prices charged for a particular brand of refrigerator in three types of stores in Louisville: discount stores, department stores and appliance stores. Random samples of 6 stores of each type were selected. The results were shown below.

| Discount | Department | Appliance |
|---|---|---|
| 1200 | 1700 | 1600 |
| 1300 | 1500 | 1500 |
| 1100 | 1450 | 1300 |
| 1400 | 1300 | 1500 |
| 1250 | 1300 | 1700 |
| 1150 | 1500 | 1400 |

At the 0.05 level of significance, is there any evidence of a difference in the average price between the types of stores?

**2.** It is conjectured that a certain gene might be linked to ovarian cancer. The ovarian cancer is sub-classified into three categories: stage I, stage II and stage III-IV. There are three random samples available; one from each stage. The samples are labelled with three colors dyes and hybridized on a four channel cDNA microarray (one channel remains unused). The experiment is repeated 5 times and the following data were obtained.

| Microarray Data | | | |
|---|---|---|---|
| Array | mRNA 1 | mRNA 2 | mRNA 3 |
| 1 | 100 | 95 | 70 |
| 2 | 90 | 93 | 72 |
| 3 | 105 | 79 | 81 |
| 4 | 83 | 85 | 74 |
| 5 | 78 | 90 | 75 |

Is there any difference between the averages of the three mRNA samples at 0.05 significance level?

**3.** A stock market analyst thinks 4 stock of mutual funds generate about the same return. He collected the accompaning rate-of-return data on 4 different mutual funds during the last 7 years. The data is given in table below.

| Mutual Funds | | | | |
|---|---|---|---|---|
| Year | A | B | C | D |
| 2000 | 12 | 11 | 13 | 15 |
| 2001 | 12 | 17 | 19 | 11 |
| 2002 | 13 | 18 | 15 | 12 |
| 2004 | 18 | 20 | 25 | 11 |
| 2005 | 17 | 19 | 19 | 10 |
| 2006 | 18 | 12 | 17 | 10 |
| 2007 | 12 | 15 | 20 | 12 |

Do a one-way ANOVA to decide whether the funds give different performance at 0.05 significance level.

**4.** Give a proof of the Theorem 20.4.

**5.** Give a proof of the Lemma 20.2.

**6.** Give a proof of the Theorem 20.5.

**7.** Give a proof of the Theorem 20.6.

**8.** An automobile company produces and sells its cars under 3 different brand names. An autoanalyst wants to see whether different brand of cars have same performance. He tested 20 cars from 3 different brands and recorded the mileage per gallon.

| Brand 1 | Brand 2 | Brand 3 |
|---|---|---|
| 32 | 31 | 34 |
| 29 | 28 | 25 |
| 32 | 30 | 31 |
| 25 | 34 | 37 |
| 35 | 39 | 32 |
| 33 | 36 | |
| 34 | 38 | |
| 31 | | |

Do the data suggest a rejection of the null hypothesis at a significance level 0.05 that the mileage per gallon generated by three different brands are same.

# Chapter 21
# GOODNESS OF FITS TESTS

In point estimation, interval estimation or hypothesis test we always started with a random sample $X_1, X_2, ..., X_n$ of size $n$ from a known distribution. In order to apply the theory to data analysis one has to know the distribution of the sample. Quite often the experimenter (or data analyst) assumes the nature of the sample distribution based on his subjective knowledge.

Goodness of fit tests are performed to validate experimenter opinion about the distribution of the population from where the sample is drawn. The most commonly known and most frequently used goodness of fit tests are the Kolmogorov-Smirnov (KS) test and the Pearson chi-square ($\chi^2$) test. There is a controversy over which test is the most powerful, but the general feeling seems to be that the Kolmogorov-Smirnov test is probably more powerful than the chi-square test in most situations. The KS test measures the distance between distribution functions, while the $\chi^2$ test measures the distance between density functions. Usually, if the population distribution is continuous, then one uses the Kolmogorov-Smirnov where as if the population distribution is discrete, then one performs the Pearson's chi-square goodness of fit test.

### 21.1. Kolmogorov-Smirnov Test

Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$. We hypothesized that the distribution of $X$ is $F(x)$. Further, we wish to test our hypothesis. Thus our null hypothesis is

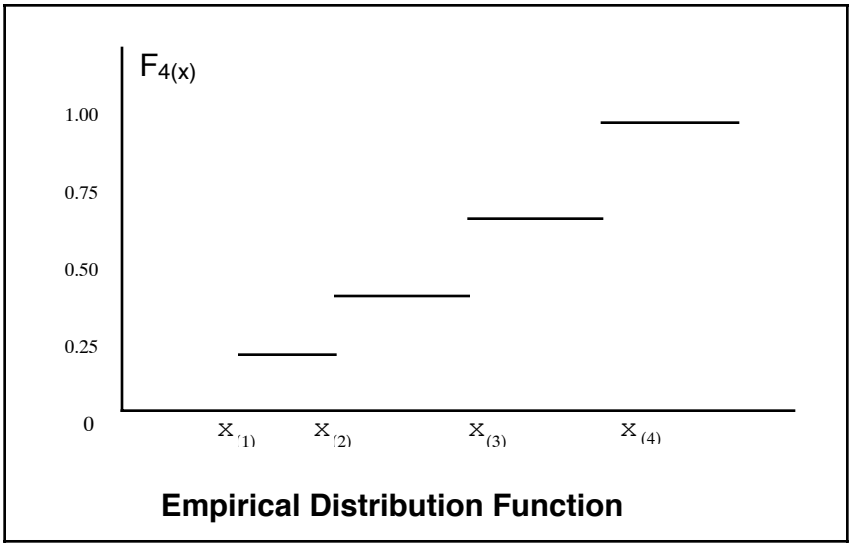
$$H_o : X \sim F(x).$$

We would like to design a test of this null hypothesis against the alternative $H_a : X \not\sim F(x)$.

In order to design a test, first of all we need a statistic which will unbiasedly estimate the unknown distribution $F(x)$ of the population $X$ using the random sample $X_1, X_2, ..., X_n$. Let $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ be the observed values of the ordered statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$. The empirical distribution of the random sample is defined as

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < x_{(1)}, \\ \frac{k}{n} & \text{if} \quad x_{(k)} \le x < x_{(k+1)}, \qquad \text{for } k = 1, 2, ..., n-1, \\ 1 & \text{if} \quad x_{(n)} \le x. \end{cases}$$

The graph of the empirical distribution function $F_4(x)$ is shown below.




For a fixed value of $x$, the empirical distribution function can be considered as a random variable that takes on the values

$$0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, \frac{n}{n}.$$

First we show that $F_n(x)$ is an unbiased estimator of the population distribution $F(x)$. That is,

$$E(F_n(x)) = F(x) \tag{1}$$

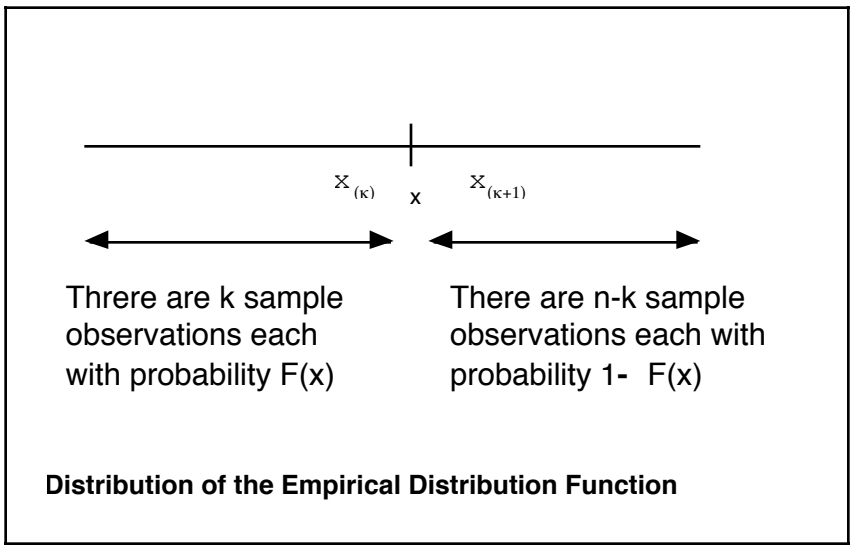
for a fixed value of $x$. To establish (1), we need the probability density function of the random variable $F_n(x)$. From the definition of the empirical distribution we see that if exactly $k$ observations are less than or equal to $x$, then

$$F_n(x) = \frac{k}{n}$$

which is

$$n\,F_n(x) = k.$$

The probability that an observation is less than or equal to $x$ is given by $F(x)$.




Hence (see figure above)

$$\begin{aligned} P(n\,F_n(x) = k) &= P\left(F_n(x) = \frac{k}{n}\right) \\ &= \binom{n}{k} [F(x)]^k \,[1 - F(x)]^{n-k} \end{aligned}$$

for $k = 0, 1, ..., n$. Thus

$$n\,F_n(x) \sim BIN(n, F(x)).$$

Thus the expected value of the random variable $n\,F_n(x)$ is given by

$$\begin{aligned} E(n\,F_n(x)) &= n\,F(x) \\ n\,E(F_n(x)) &= n\,F(x) \\ E(F_n(x)) &= F(x). \end{aligned}$$

This shows that, for a fixed $x$, $F_n(x)$, on an average, equals to the population distribution function $F(x)$. Hence the empirical distribution function $F_n(x)$ is an unbiased estimator of $F(x)$.

Since $n\,F_n(x) \sim BIN(n, F(x))$, the variance of $n\,F_n(x)$ is given by

$$Var(n\,F_n(x)) = n\,F(x)\,[1 - F(x)].$$

Hence the variance of $F_n(x)$ is

$$Var(F_n(x)) = \frac{F(x)\,[1 - F(x)]}{n}.$$

It is easy to see that $Var(F_n(x)) \to 0$ as $n \to \infty$ for all values of $x$. Thus the empirical distribution function $F_n(x)$ and $F(x)$ tend to be closer to each other with large $n$. As a matter of fact, Glivenkno, a Russian mathematician, proved that $F_n(x)$ converges to $F(x)$ uniformly in $x$ as $n \to \infty$ with probability one.

Because of the convergence of the empirical distribution function to the theoretical distribution function, it makes sense to construct a goodness of fit test based on the closeness of $F_n(x)$ and hypothesized distribution $F(x)$.

Let

$$D_n = \max_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

That is $D_n$ is the maximum of all pointwise differences $|F_n(x) - F(x)|$. The distribution of the Kolmogorov-Smirnov statistic, $D_n$ can be derived. However, we shall not do that here as the derivation is quite involved. In stead, we give a closed form formula for $P(D_n \leq d)$. If $X_1, X_2, ..., X_n$ is a sample from a population with continuous distribution function $F(x)$, then

$$P(D_n \leq d) = \begin{cases} 0 & \text{if } d \leq \frac{1}{2\,n} \\ n! \prod_{i=1}^{n} \int_{2\,i-d}^{2\,i-\frac{1}{n}+d} du & \text{if } \frac{1}{2n} < d < 1 \\ 1 & \text{if } d \geq 1 \end{cases}$$

where $du = du_1 du_2 \cdots du_n$ with $0 < u_1 < u_2 < \cdots < u_n < 1$. Further,

$$\lim_{n\to\infty} P(\sqrt{n}\, D_n \le d) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2\,k^2\,d^2}.$$

These formulas show that the distribution of the Kolmogorov-Smirnov statistic $D_n$ is distribution free, that is, it does not depend on the distribution $F$ of the population.

For most situations, it is sufficient to use the following approximations due to Kolmogorov:

$$P(\sqrt{n}\, D_n \le d) \approx 1 - 2e^{-2nd^2} \qquad \text{for } d > \frac{1}{\sqrt{n}}.$$

If the null hypothesis $H_o : X \sim F(x)$ is true, the statistic $D_n$ is small. It is therefore reasonable to reject $H_o$ if and only if the observed value of $D_n$ is larger than some constant $d_n$. If the level of significance is given to be $\alpha$, then the constant $d_n$ can be found from

$$\alpha = P(D_n > d_n \,/\, H_o \text{ is true}) \approx 2e^{-2nd_n^2}.$$

This yields the following hypothesis test: Reject $H_o$ if $D_n \ge d_n$ where

$$d_n = \sqrt{-\frac{1}{2n} \ln\left(\frac{\alpha}{2}\right)}$$

is obtained from the above Kolmogorov's approximation. Note that the approximate value of $d_{12}$ obtained by the above formula is equal to 0.3533 when $\alpha = 0.1$, however more accurate value of $d_{12}$ is 0.34.

Next we address the issue of the computation of the statistics $D_n$. Let us define

$$D_n^+ = \max_{x\in\mathbb{R}} \{F_n(x) - F(x)\}$$

and

$$D_n^- = \max_{x\in\mathbb{R}} \{F(x) - F_n(x)\}.$$

Then it is easy to see that

$$D_n = \max\{D_n^+,\, D_N^-\}.$$

Further, since $F_n(x_{(i)}) = \frac{i}{n}$. it can be shown that

$$D_n^+ = \max\left\{ \max_{1\le i\le n} \left[\frac{i}{n} - F(x_{(i)})\right],\, 0 \right\}$$

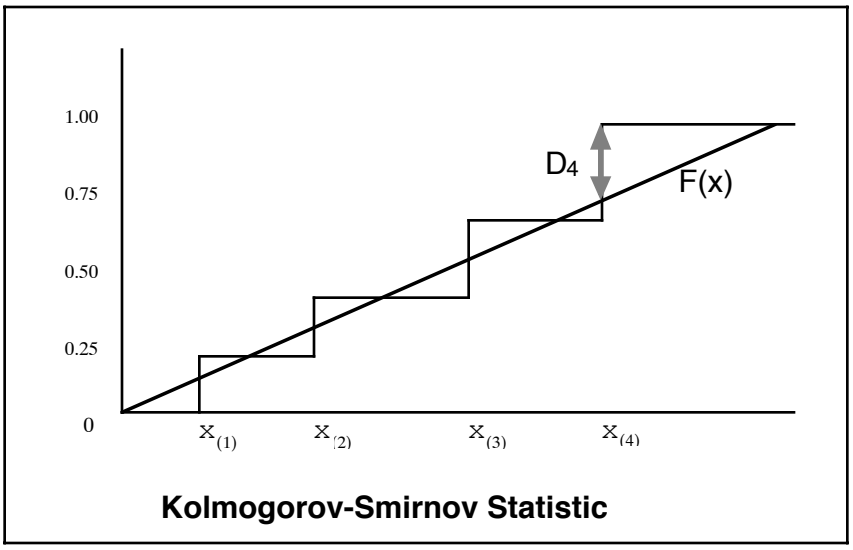
and

$$D_n^- = \max\left\{\max_{1\le i\le n}\left[F(x_{(i)}) - \frac{i-1}{n}\right], 0\right\}.$$

Therefore it can also be shown that

$$D_n = \max_{1\le i\le n}\left\{\max\left[\frac{i}{n} - F(x_{(i)}),\, F(x_{(i)}) - \frac{i-1}{n}\right]\right\}.$$

The following figure illustrates the Kolmogorov-Smirnov statistics $D_n$ when $n = 4$.




**Example 21.1.** The data on the heights of 12 infants are given below: 18.2, 21.4, 22.6, 17.4, 17.6, 16.7, 17.1, 21.4, 20.1, 17.9, 16.8, 23.1. Test the hypothesis that the data came from some normal population at a significance level $\alpha = 0.1$.

**Answer:** Here, the null hypothesis is

$$H_o : X \sim N(\mu, \sigma^2).$$

First we estimate $\mu$ and $\sigma^2$ from the data. Thus, we get

$$\overline{x} = \frac{230.3}{12} = 19.2.$$

and

$$s^2 = \frac{4482.01 - \frac{1}{12}(230.3)^2}{12-1} = \frac{62.17}{11} = 5.65.$$

Hence $s = 2.38$. Then by the null hypothesis

$$F(x_{(i)}) = P\left(Z \leq \frac{x_{(i)} - 19.2}{2.38}\right)$$

where $Z \sim N(0,1)$ and $i = 1, 2, ..., n$. Next we compute the Kolmogorov-Smirnov statistic $D_n$ the given sample of size 12 using the following tabular form.

| $i$ | $x_{(i)}$ | $F(x_{(i)})$ | $\frac{i}{12} - F(x_{(i)})$ | $F(x_{(i)}) - \frac{i-1}{12}$ |
|---|---|---|---|---|
| 1 | 16.7 | 0.1469 | −0.0636 | 0.1469 |
| 2 | 16.8 | 0.1562 | 0.0105 | 0.0729 |
| 3 | 17.1 | 0.1894 | 0.0606 | 0.0227 |
| 4 | 17.4 | 0.2236 | 0.1097 | −0.0264 |
| 5 | 17.6 | 0.2514 | 0.1653 | −0.0819 |
| 6 | 17.9 | 0.2912 | 0.2088 | −0.1255 |
| 7 | 18.2 | 0.3372 | 0.2461 | −0.1628 |
| 8 | 20.1 | 0.6480 | 0.0187 | 0.0647 |
| 9 | 21.4 | 0.8212 | 0.0121 | 0.0712 |
| 10 | 21.4 | | | |
| 11 | 22.6 | 0.9236 | −0.0069 | 0.0903 |
| 12 | 23.1 | 0.9495 | 0.0505 | 0.0328 |

Thus

$$D_{12} = 0.2461.$$

From the tabulated value, we see that $d_{12} = 0.34$ for significance level $\alpha = 0.1$. Since $D_{12}$ is smaller than $d_{12}$ we accept the null hypothesis $H_o : X \sim N(\mu, \sigma^2)$. Hence the data came from a normal population.

**Example 21.2.** Let $X_1, X_2, ..., X_{10}$ be a random sample from a distribution whose probability density function is

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Based on the observed values 0.62, 0.36, 0.23, 0.76, 0.65, 0.09, 0.55, 0.26, 0.38, 0.24, test the hypothesis $H_o : X \sim UNIF(0,1)$ against $H_a : X \not\sim UNIF(0,1)$ at a significance level $\alpha = 0.1$.

**Answer:** The null hypothesis is $H_o : X \sim UNIF(0,1)$. Thus

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

Hence

$$F(x_{(i)}) = x_{(i)} \qquad \text{for } i = 1, 2, ..., n.$$

Next we compute the Kolmogorov-Smirnov statistic $D_n$ the given sample of size 10 using the following tabular form.

| $i$ | $x_{(i)}$ | $F(x_{(i)})$ | $\frac{i}{10} - F(x_{(i)})$ | $F(x_{(i)}) - \frac{i-1}{10}$ |
|---|---|---|---|---|
| 1 | 0.09 | 0.09 | 0.01 | 0.09 |
| 2 | 0.23 | 0.23 | −0.03 | 0.13 |
| 3 | 0.24 | 0.24 | 0.06 | 0.04 |
| 4 | 0.26 | 0.26 | 0.14 | −0.04 |
| 5 | 0.36 | 0.36 | 0.14 | −0.04 |
| 6 | 0.38 | 0.38 | 0.22 | −0.12 |
| 7 | 0.55 | 0.55 | 0.15 | −0.05 |
| 8 | 0.62 | 0.62 | 0.18 | −0.08 |
| 9 | 0.65 | 0.65 | 0.25 | −0.15 |
| 10 | 0.76 | 0.76 | 0.24 | −0.14 |

Thus

$$D_{10} = 0.25.$$

From the tabulated value, we see that $d_{10} = 0.37$ for significance level $\alpha = 0.1$. Since $D_{10}$ is smaller than $d_{10}$ we accept the null hypothesis

$$H_o : X \sim UNIF(0,1).$$

### 21.2 Chi-square Test

The chi-square goodness of fit test was introduced by Karl Pearson in 1900. Recall that the Kolmogorov-Smirnov test is only for testing a specific continuous distribution. Thus if we wish to test the null hypothesis

$$H_o : X \sim BIN(n,p)$$

against the alternative $H_a : X \not\sim BIN(n,p)$, then we can not use the Kolmogorov-Smirnov test. Pearson chi-square goodness of fit test can be used for testing of null hypothesis involving discrete as well as continuous

distribution. Unlike Kolmogorov-Smirnov test, the Pearson chi-square test uses the density function the population $X$.

Let $X_1, X_2, ..., X_n$ be a random sample from a population $X$ with probability density function $f(x)$. We wish to test the null hypothesis

$$H_o : X \sim f(x)$$

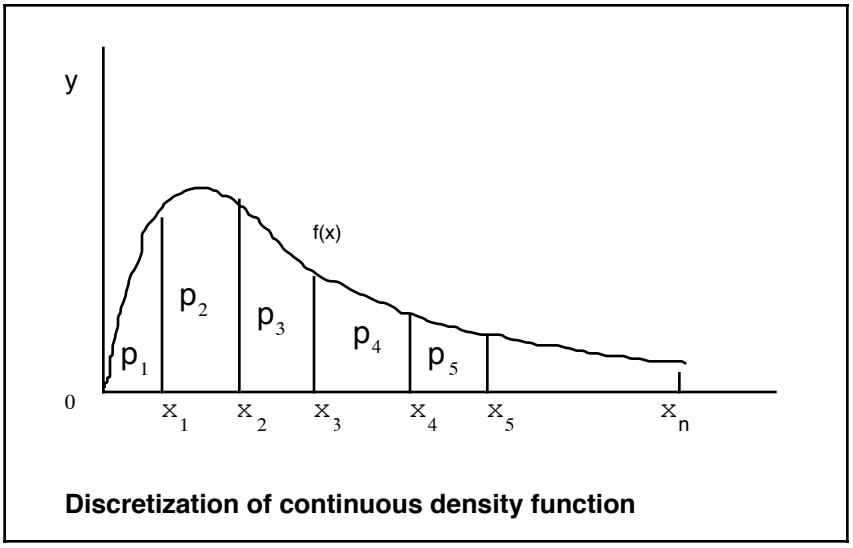
against

$$H_a : X \not\sim f(x).$$

If the probability density function $f(x)$ is continuous, then we divide up the abscissa of the probability density function $f(x)$ and calculate the probability $p_i$ for each of the interval by using

$$p_i = \int_{x_{i-1}}^{x_i} f(x)\, dx,$$

where $\{x_0, x_1, ..., x_n\}$ is a partition of the domain of the $f(x)$.




**Discretization of continuous density function**

Let $Y_1, Y_2, ..., Y_m$ denote the number of observations (from the random sample $X_1, X_2, ..., X_n$) is $1^{\text{st}}, 2^{\text{nd}}, 3^{\text{rd}}, ..., m^{\text{th}}$ interval, respectively.

Since the sample size is $n$, the number of observations expected to fall in the $i^{\text{th}}$ interval is equal to $np_i$. Then

$$Q = \sum_{i=1}^{m} \frac{(Y_i - np_i)^2}{np_i}$$

measures the closeness of observed $Y_i$ to expected number $np_i$. The distribution of $Q$ is chi-square with $m-1$ degrees of freedom. The derivation of this fact is quite involved and beyond the scope of this introductory level book.

Although the distribution of $Q$ for $m > 2$ is hard to derive, yet for $m = 2$ it not very difficult. Thus we give a derivation to convince the reader that $Q$ has $\chi^2$ distribution. Notice that $Y_1 \sim BIN(n, p_1)$. Hence for large $n$ by the central limit theorem, we have

$$\frac{Y_1 - n\,p_1}{\sqrt{n\,p_1\,(1-p_1)}} \sim N(0,1).$$

Thus

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1\,(1-p_1)} \sim \chi^2(1).$$

Since

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1\,(1-p_1)} = \frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(Y_1 - n\,p_1)^2}{n\,(1-p_1)},$$

we have This implies that

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(Y_1 - n\,p_1)^2}{n\,(1-p_1)} \sim \chi^2(1)$$

which is

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(n - Y_2 - n + n\,p_2)^2}{n\,p_2} \sim \chi^2(1)$$

due to the facts that $Y_1 + Y_2 = n$ and $p_1 + p_2 = 1$. Hence

$$\sum_{i=1}^{2} \frac{(Y_i - n\,p_i)^2}{n\,p_i} \sim \chi^2(1),$$

that is, the chi-square statistic $Q$ has approximate chi-square distribution.

Now the simple null hypothesis

$$H_0 : p_1 = p_{10},\ \ p_2 = p_{20},\ \ \cdots\ \ p_m = p_{m0}$$

is to be tested against the composite alternative

$$H_a : \text{at least one } p_i \text{ is not equal to } p_{i0} \text{ for some } i.$$

Here $p_{10}, p_{20}, ..., p_{m0}$ are fixed probability values. If the null hypothesis is true, then the statistic

$$Q = \sum_{i=1}^{m} \frac{(Y_i - n\,p_{i0})^2}{n\,p_{i0}}$$

has an approximate chi-square distribution with $m-1$ degrees of freedom. If the significance level $\alpha$ of the hypothesis test is given, then

$$\alpha = P\left(Q \geq \chi^2_{1-\alpha}(m-1)\right)$$

and the test is "Reject $H_o$ if $Q \geq \chi^2_{1-\alpha}(m-1)$." Here $\chi^2_{1-\alpha}(m-1)$ denotes a real number such that the integral of the chi-square density function with $m-1$ degrees of freedom from zero to this real number $\chi^2_{1-\alpha}(m-1)$ is $1-\alpha$. Now we give several examples to illustrate the chi-square goodness-of-fit test.

**Example 21.3.** A die was rolled 30 times with the results shown below:

| Number of spots | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency ($x_i$) | 1 | 4 | 9 | 9 | 2 | 5 |

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, then what is the value of the chi-square statistic and decision reached?

**Answer:** In this problem, the null hypothesis is

$$H_o : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}.$$

The alternative hypothesis is that not all $p_i$'s are equal to $\frac{1}{6}$. The test will be based on 30 trials, so $n = 30$. The test statistic

$$Q = \sum_{i=1}^{6} \frac{(x_i - n\,p_i)^2}{n\,p_i},$$

where $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Thus

$$n\,p_i = (30)\,\frac{1}{6} = 5$$

and

$$\begin{aligned} Q &= \sum_{i=1}^{6} \frac{(x_i - n\,p_i)^2}{n\,p_i} \\ &= \sum_{i=1}^{6} \frac{(x_i - 5)^2}{5} \\ &= \frac{1}{5}\,[16 + 1 + 16 + 16 + 9] \\ &= \frac{58}{5} = 11.6. \end{aligned}$$

The tabulated $\chi^2$ value for $\chi^2_{0.95}(5)$ is given by

$$\chi^2_{0.95}(5) = 11.07.$$

Since

$$11.6 = Q > \chi^2_{0.95}(5) = 11.07$$

the null hypothesis $H_o : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ should be rejected.

**Example 21.4.** It is hypothesized that an experiment results in outcomes $K$, $L$, $M$ and $N$ with probabilities $\frac{1}{5}$, $\frac{3}{10}$, $\frac{1}{10}$ and $\frac{2}{5}$, respectively. Forty independent repetitions of the experiment have results as follows:

| Outcome | K | L | M | N |
|---|---|---|---|---|
| Frequency | 11 | 14 | 5 | 10 |

If a chi-square goodness of fit test is used to test the above hypothesis at the significance level $\alpha = 0.01$, then what is the value of the chi-square statistic and the decision reached?

**Answer:** Here the null hypothesis to be tested is

$$H_o : p(K) = \frac{1}{5}, p(L) = \frac{3}{10}, p(M) = \frac{1}{10}, p(N) = \frac{2}{5}.$$

The test will be based on $n = 40$ trials. The test statistic

$$\begin{aligned} Q &= \sum_{k=1}^{4} \frac{(x_k - np_k)^2}{n\, p_k} \\ &= \frac{(x_1 - 8)^2}{8} + \frac{(x_2 - 12)^2}{12} + \frac{(x_3 - 4)^2}{4} + \frac{(x_4 - 16)^2}{16} \\ &= \frac{(11 - 8)^2}{8} + \frac{(14 - 12)^2}{12} + \frac{(5 - 4)^2}{4} + \frac{(10 - 16)^2}{16} \\ &= \frac{9}{8} + \frac{4}{12} + \frac{1}{4} + \frac{36}{16} \\ &= \frac{95}{24} = 3.958. \end{aligned}$$

From chi-square table, we have

$$\chi^2_{0.99}(3) = 11.35.$$

Thus

$$3.958 = Q < \chi^2_{0.99}(3) = 11.35.$$

Therefore we accept the null hypothesis.

**Example 21.5.** Test at the 10% significance level the hypothesis that the following data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 06.88 | 06.92 | 04.80 | 09.85 | 07.05 | 19.06 | 06.54 | 03.67 | 02.94 | 04.89 |
| 69.82 | 06.97 | 04.34 | 13.45 | 05.74 | 10.07 | 16.91 | 07.47 | 05.04 | 07.97 |
| 15.74 | 00.32 | 04.14 | 05.19 | 18.69 | 02.45 | 23.69 | 44.10 | 01.70 | 02.14 |
| 05.79 | 03.02 | 09.87 | 02.44 | 18.99 | 18.90 | 05.42 | 01.54 | 01.55 | 20.99 |
| 07.99 | 05.38 | 02.36 | 09.66 | 00.97 | 04.82 | 10.43 | 15.06 | 00.49 | 02.81 |

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

**Answer:** From the data $\overline{x} = 9.74$ and $s = 11.71$. Notice that

$$H_o : X \sim EXP(\theta).$$

Hence we have to partition the domain of the experimental distribution into $m$ parts. There is no rule to determine what should be the value of $m$. We assume $m = 10$ (an arbitrary choice for the sake of convenience). We partition the domain of the given probability density function into 10 mutually disjoint sets of equal probability. This partition can be found as follow.

Note that $\overline{x}$ estimate $\theta$. Thus

$$\widehat{\theta} = \overline{x} = 9.74.$$

Now we compute the points $x_1, x_2, ..., x_{10}$ which will be used to partition the domain of $f(x)$

$$\begin{aligned} \frac{1}{10} &= \int_{x_o}^{x_1} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} \\ &= -\left[e^{-\frac{x}{\theta}}\right]_0^{x_1} \\ &= 1 - e^{-\frac{x_1}{\theta}}. \end{aligned}$$

Hence

$$\begin{aligned} x_1 &= \theta \ln\left(\frac{10}{9}\right) \\ &= 9.74 \ln\left(\frac{10}{9}\right) \\ &= 1.026. \end{aligned}$$

Using the value of $x_1$, we can find the value of $x_2$. That is

$$\frac{1}{10} = \int_{x_1}^{x_2} \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

$$= e^{-\frac{x_1}{\theta}} - e^{-\frac{x_2}{\theta}}.$$

Hence

$$x_2 = -\theta \ln\left(e^{-\frac{x_1}{\theta}} - \frac{1}{10}\right).$$

In general

$$x_k = -\theta \ln\left(e^{-\frac{x_{k-1}}{\theta}} - \frac{1}{10}\right)$$

for $k = 1, 2, ..., 9$, and $x_{10} = \infty$. Using these $x_k$'s we find the intervals $A_k = [x_k, x_{k+1})$ which are tabulates in the table below along with the number of data points in each each interval.

| Interval $A_i$ | Frequency $(o_i)$ | Expected value $(e_i)$ |
|---|---|---|
| [0, 1.026) | 3 | 5 |
| [1.026, 2.173) | 4 | 5 |
| [2.173, 3.474) | 6 | 5 |
| [3.474, 4.975) | 6 | 5 |
| [4.975, 6.751) | 7 | 5 |
| [6.751, 8.925) | 7 | 5 |
| [8.925, 11.727) | 5 | 5 |
| [11.727, 15.676) | 2 | 5 |
| [15.676, 22.437) | 7 | 5 |
| [22.437, $\infty$) | 3 | 5 |
| Total | 50 | 50 |

From this table, we compute the statistics

$$Q = \sum_{i=1}^{10} \frac{(o_i - e_i)^2}{e_i} = 6.4.$$

and from the chi-square table, we obtain

$$\chi^2_{0.9}(9) = 14.68.$$

Since

$$6.4 = Q < \chi^2_{0.9}(9) = 14.68$$

we accept the null hypothesis that the sample was taken from a population with exponential distribution.

### 21.3. Review Exercises

**1.** The data on the heights of 4 infants are: $18.2, 21.4, 16.7$ and $23.1$. For a significance level $\alpha = 0.1$, use Kolmogorov-Smirnov Test to test the hypothesis that the data came from some uniform population on the interval $(15, 25)$. (Use $d_4 = 0.56$ at $\alpha = 0.1$.)

**2.** A four-sided die was rolled 40 times with the following results

| Number of spots | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Frequency | 5 | 9 | 10 | 16 |

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, then what is the value of the chi-square statistic?

**3.** A coin is tossed 500 times and $k$ heads are observed. If the chi-squares distribution is used to test the hypothesis that the coin is unbiased, this hypothesis will be accepted at 5 percents level of significance if and only if $k$ lies between what values? (Use $\chi^2_{0.05}(1) = 3.84$.)

**4.** It is hypothesized that an experiment results in outcomes $A$, $C$, $T$ and $G$ with probabilities $\frac{1}{16}$, $\frac{5}{16}$, $\frac{1}{8}$ and $\frac{3}{8}$, respectively. Eighty independent repetitions of the experiment have results as follows:

| Outcome | A | G | C | T |
|---|---|---|---|---|
| Frequency | 3 | 28 | 15 | 34 |

If a chi-square goodness of fit test is used to test the above hypothesis at the significance level $\alpha = 0.1$, then what is the value of the chi-square statistic and the decision reached?

**5.** A die was rolled 50 times with the results shown below:

| Number of spots | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency ($x_i$) | 8 | 7 | 12 | 13 | 4 | 6 |

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.1$, then what is the value of the chi-square statistic and decision reached?

**6.** Test at the 10% significance level the hypothesis that the following data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 05.88 | 05.92 | 03.80 | 08.85 | 06.05 | 18.06 | 05.54 | 02.67 | 01.94 | 03.89 |
| 70.82 | 07.97 | 05.34 | 14.45 | 06.74 | 11.07 | 17.91 | 08.47 | 06.04 | 08.97 |
| 16.74 | 01.32 | 03.14 | 06.19 | 19.69 | 03.45 | 24.69 | 45.10 | 02.70 | 03.14 |
| 04.79 | 02.02 | 08.87 | 03.44 | 17.99 | 17.90 | 04.42 | 01.54 | 01.55 | 19.99 |
| 06.99 | 05.38 | 03.36 | 08.66 | 01.97 | 03.82 | 11.43 | 14.06 | 01.49 | 01.81 |

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

**7.** Test at the 10% significance level the hypothesis that the following data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.88 | 0.92 | 0.80 | 0.85 | 0.05 | 0.06 | 0.54 | 0.67 | 0.94 | 0.89 |
| 0.82 | 0.97 | 0.34 | 0.45 | 0.74 | 0.07 | 0.91 | 0.47 | 0.04 | 0.97 |
| 0.74 | 0.32 | 0.14 | 0.19 | 0.69 | 0.45 | 0.69 | 0.10 | 0.70 | 0.14 |
| 0.79 | 0.02 | 0.87 | 0.44 | 0.99 | 0.90 | 0.42 | 0.54 | 0.55 | 0.99 |
| 0.94 | 0.38 | 0.36 | 0.66 | 0.97 | 0.82 | 0.43 | 0.06 | 0.49 | 0.81 |

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{if } 0 \le x \le 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

**8.** Test at the 10% significance level the hypothesis that the following data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 06.88 | 06.92 | 04.80 | 09.85 | 07.05 | 19.06 | 06.54 | 03.67 | 02.94 | 04.89 |
| 29.82 | 06.97 | 04.34 | 13.45 | 05.74 | 10.07 | 16.91 | 07.47 | 05.04 | 07.97 |
| 15.74 | 00.32 | 04.14 | 05.19 | 18.69 | 02.45 | 23.69 | 24.10 | 01.70 | 02.14 |
| 05.79 | 03.02 | 09.87 | 02.44 | 18.99 | 18.90 | 05.42 | 01.54 | 01.55 | 20.99 |
| 07.99 | 05.38 | 02.36 | 09.66 | 00.97 | 04.82 | 10.43 | 15.06 | 00.49 | 02.81 |

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta \\ \\ 0 & \text{elsewhere.} \end{cases}$$

**9.** Suppose that in 60 rolls of a die the outcomes 1, 2, 3, 4, 5, and 6 occur with frequencies $n_1$, $n_2$, 14, 8, 10, and 8 respectively. What is the least value of $\sum_{i=1}^{2}(n_i-10)^2$ for which the chi-square test rejects the hypothesis that the die is fair at 1% level of significance level? (Answer: $\sum_{i=1}^{2}(n_i-10)^2 \geq 63.43$.)

**10.** It is hypothesized that of all marathon runners 70% are adult men, 25% are adult women, and 5% are youths. To test this hypothesis, the following data from the a recent marathon are used:

| Adult Men | Adult Women | Youths | Total |
|---|---|---|---|
| 630 | 300 | 70 | 1000 |

A chi-square goodness-of-fit test is used. What is the value of the statistics? (Ans: 25)

# REFERENCES


[1] Aitken, A. C. (1944). *Statistical Mathematics*. 3rd edn. Edinburgh and London: Oliver and Boyd,

[2] Arbous, A. G. and Kerrich, J. E. (1951). Accident statistics and the concept of accident-proneness. *Biometrics*, 7, 340-432.

[3] Arnold, S. (1984). Pivotal quantities and invariant confidence regions. *Statistics and Decisions* 2, 257-280.

[4] Bain, L. J. and Engelhardt. M. (1992). *Introduction to Probability and Mathematical Statistics*. Belmont: Duxbury Press.

[5] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society*, London, Ser. A, 160, 268-282.

[6] Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *J. R. Stat. Soc., Suppli.*, 4, 137-183.

[7] Brown, L. D. (1988). *Lecture Notes*, Department of Mathematics, Cornell University. Ithaca, New York.

[8] Brown, M. B. and Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of American Statistical Association*, 69, 364-367.

[9] Campbell, J. T. (1934). THe Poisson correlation function. *Proc. Edin. Math. Soc.*, Series 2, 4, 18-26.

[10] Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Belmont: Wadsworth.

[11] Castillo, E. (1988). *Extreme Value Theory in Engineering*. San Diego: Academic Press.

[12] Cherian, K. C. (1941). A bivariate correlated gamma-type distribution function. *J. Indian Math. Soc.*, 5, 133-144.

[13] Dahiya, R., and Guttman, I. (1982). Shortest confidence and prediction intervals for the log-normal. *The canadian Journal of Statistics* 10, 777-891.

[14] David, F.N. and Fix, E. (1961). Rank correlation and regression in a non-normal surface. *Proc. 4th Berkeley Symp. Math. Statist. & Prob.*, 1, 177-197.

[15] Desu, M. (1971). Optimal confidence intervals of fixed width. *The American Statistician* 25, 27-29.

[16] Dynkin, E. B. (1951). Necessary and sufficient statistics for a family of probability distributions. English translation in *Selected Translations in Mathematical Statistics and Probability*, 1 (1961), 23-41.

[17] Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen, *Ann. Phys.* 17, 549560.

[18] Eisenhart, C., Hastay, M. W. and Wallis, W. A. (1947). *Selected Techniques of Statistical Analysis*, New York: McGraw-Hill.

[19] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume I.* New York: Wiley.

[20] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Volume II.* New York: Wiley.

[21] Ferentinos, K. K. (1988). On shortest confidence intervals and their relation uniformly minimum variance unbiased estimators. *Statistical Papers* 29, 59-75.

[22] Freund, J. E. and Walpole, R. E. (1987). *Mathematical Statistics.* Englewood Cliffs: Prantice-Hall.

[23] Galton, F. (1879). The geometric mean in vital and social statistics. *Proc. Roy. Soc.*, 29, 365-367.

[24] Galton, F. (1886). Family likeness in stature. With an appendix by J.D.H. Dickson. *Proc. Roy. Soc.*, 40, 42-73.

[25] Ghahramani, S. (2000). *Fundamentals of Probability.* Upper Saddle River, New Jersey: Prentice Hall.

[26] Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*, Vol. 1. New YorK: McGraw-Hill.

[27] Guenther, W. (1969). Shortest confidence intervals. *The American Statistician* 23, 51-53.

[28] Guldberg, A. (1934). On discontinuous frequency functions of two variables. *Skand. Aktuar.*, 17, 89-117.

[29] Gumbel, E. J. (1960). Bivariate exponetial distributions. *J. Amer. Statist. Ass.*, 55, 698-707.

[30] Hamedani, G. G. (1992). Bivariate and multivariate normal characterizations: a brief survey. *Comm. Statist. Theory Methods*, 21, 2665-2688.

[31] Hamming, R. W. (1991). *The Art of Probability for Scientists and Engineers* New York: Addison-Wesley.

[32] Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics.* New York: Macmillan.

[33] Hogg, R. V. and Tanis, E. A. (1993). *Probability and Statistical Inference.* New York: Macmillan.

[34] Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, 51, 241-245.

[35] Kapteyn, J. C. (1903). *Skew Frequency Curves in Biology and Statistics.* Astronomical Laboratory, Noordhoff, Groningen.

[36] Kibble, W. F. (1941). A two-variate gamma type distribution. *Sankhya*, 5, 137-150.

[37] Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Erg. Math., Vol 2, Berlin: Springer-Verlag.

[38] Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability.* New York: Chelsea Publishing Company.

[39] Kotlarski, I. I. (1960). On random variables whose quotient follows the Cauchy law. *Colloquium Mathematicum.* 7, 277-284.

[40] Isserlis, L. (1914). The application of solid hypergeometrical series to frequency distributions in space. *Phil. Mag.*, 28, 379-403.

[41] Laha, G. (1959). On a class of distribution functions where the quotient follows the Cauchy law. *Trans. Amer. Math. Soc.* 93, 205-215.

[42] Levene, H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. I. Olkin *et. al.* eds., Stanford University Press, 278-292.

[43] Lundberg, O. (1934). *On Random Processes and their Applications to Sickness and Accident Statistics.* Uppsala: Almqvist and Wiksell.

[44] Mardia, K. V. (1970). *Families of Bivariate Distributions.* London: Charles Griffin & Co Ltd.

[45] Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *J. Amer. Statist. Ass.*, 62. 30-44.

[46] McAlister, D. (1879). The law of the geometric mean. *Proc. Roy. Soc.*, 29, 367-375.

[47] McKay, A. T. (1934). Sampling from batches. *J. Roy. Statist. Soc.*, Suppliment, 1, 207-216.

[48] Meyer, P. L. (1970). *Introductory Probability and Statistical Applications.* Reading: Addison-Wesley.

[49] Mood, A., Graybill, G. and Boes, D. (1974). *Introduction to the Theory of Statistics* (3rd Ed.). New York: McGraw-Hill.

[50] Moran, P. A. P. (1967). Testing for correlation between non-negative variates. *Biometrika*, 54, 385-394.

[51] Morgenstern, D. (1956). Einfache Beispiele zweidimensionaler Verteilungen. *Mitt. Math. Statist.*, 8, 234-235.

[52] Papoulis, A. (1990). *Probability and Statistics.* Englewood Cliffs: Prantice-Hall.

[53] Pearson, K. (1924). On the moments of the hypergeometrical series. *Biometrika*, 16, 157-160.

[54] Pestman, W. R. (1998). *Mathematical Statistics: An Introduction* New York: Walter de Gruyter.

[55] Pitman, J. (1993). *Probability.* New York: Springer-Verlag.

[56] Plackett, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Ass.*, 60, 516-522.

[57] Rice, S. O. (1944). Mathematical analysis of random noise. *Bell. Syst. Tech. J.*, 23, 282-332.

[58] Rice, S. O. (1945). Mathematical analysis of random noise. *Bell. Syst. Tech. J.*, 24, 46-156.

[59] Rinaman, W. C. (1993). *Foundations of Probability and Statistics.* New York: Saunders College Publishing.

[60] Rosenthal, J. S. (2000). *A First Look at Rigorous Probability Theory.* Singapore: World Scientific.

[61] Ross, S. (1988). *A First Course in Probability.* New York: Macmillan.

[62] Ross, S. M. (2000). *Introduction to Probability and Statistics for Engineers and Scientists.* San Diego: Harcourt Academic Press.

[63] Roussas, G. (2003). *An Introduction to Probability and Statistical Inference.* San Diego: Academic Press.

[64] Sahai, H. and Ageel, M. I. (2000). *The Analysis of Variance.* Boston: Birkhauser.

[65] Seshadri, V. and Patil, G. P. (1964). A characterization of a bivariate distribution by the marginal and the conditional distributions of the same component. *Ann. Inst. Statist. Math.*, 15, 215-221.

[66] H. Scheffé (1959). *The Analysis of Variance.* New York: Wiley.

[67] Smoluchowski, M. (1906). Zur kinetischen Theorie der Brownschen Molekularbewe-gung und der Suspensionen, *Ann. Phys.* 21, 756780.

[68] Snedecor, G. W. and Cochran, W. G. (1983). *Statistical Methods.* 6th eds. Iowa State University Press, Ames, Iowa.

[69] Sveshnikov, A. A. (1978). *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions.* New York: Dover.

[70] Tardiff, R. M. (1981). L'Hospital rule and the central limit theorem. *American Statistician*, 35, 43-44.

[71] Taylor, L. D. (1974). *Probability and Mathematical Statistics.* New York: Harper & Row.

[72] Tweedie, M. C. K. (1945). Inverse statistical variates. *Nature*, 155, 453.

[73] Waissi, G. R. (1993). A unifying probability density function. Appl. Math. Lett. 6, 25-26.

[74] Waissi, G. R. (1994). An improved unifying density function. Appl. Math. Lett. 7, 71-73.

[75] Waissi, G. R. (1998). Transformation of the unifying density to the normal distribution. Appl. Math. Lett. 11, 45-28.

[76] Wicksell, S. D. (1933). On correlation functions of Type III. *Biometrika*, 25, 121-133.

[77] Zehna, P. W. (1966). Invariance of maximum likelihood estimators. *Annals of Mathematical Statistics*, 37, 744.

# ANSWERES TO SELECTED REVIEW EXERCISES

## CHAPTER 1

**1.** $\frac{7}{1912}$.
**2.** 244.
**3.** 7488.
**4.** (a) $\frac{4}{24}$, (b) $\frac{6}{24}$ and (c) $\frac{4}{24}$.
**5.** 0.95.
**6.** $\frac{4}{7}$.
**7.** $\frac{2}{3}$.
**8.** 7560.
**10.** $4^3$.
**11.** 2.
**12.** 0.3238.
**13.** $S$ has countable number of elements.
**14.** $S$ has uncountable number of elements.
**15.** $\frac{25}{648}$.
**16.** $(n-1)(n-2)\left(\frac{1}{2}\right)^{n+1}$.
**17.** $(5!)^2$.
**18.** $\frac{7}{10}$.
**19.** $\frac{1}{3}$.
**20.** $\frac{n+1}{3n-1}$.
**21.** $\frac{6}{11}$.
**22.** $\frac{1}{5}$.

## CHAPTER 2

**1.** $\frac{1}{3}$.

**2.** $\frac{(6!)^2}{(21)^6}$.

**3.** 0.941.

**4.** $\frac{4}{5}$.

**5.** $\frac{6}{11}$.

**6.** $\frac{255}{256}$.

**7.** 0.2929.

**8.** $\frac{10}{17}$.

**9.** $\frac{30}{31}$.

**10.** $\frac{7}{24}$.

**11.** $\frac{\left(\frac{4}{10}\right)\left(\frac{3}{6}\right)}{\left(\frac{4}{10}\right)\left(\frac{3}{6}\right)+\left(\frac{6}{10}\right)\left(\frac{2}{5}\right)}$.

**12.** $\frac{(0.01)\,(0.9)}{(0.01)\,(0.9)+(0.99)\,(0.1)}$.

**13.** $\frac{1}{5}$.

**14.** $\frac{2}{9}$.

**15.** (a) $\left(\frac{2}{5}\right)\left(\frac{4}{52}\right)+\left(\frac{3}{5}\right)\left(\frac{4}{16}\right)$ and (b) $\frac{\left(\frac{3}{5}\right)\left(\frac{4}{16}\right)}{\left(\frac{2}{5}\right)\left(\frac{4}{52}\right)+\left(\frac{3}{5}\right)\left(\frac{4}{16}\right)}$.

**16.** $\frac{1}{4}$.

**17.** $\frac{3}{8}$.

**18.** 5.

**19.** $\frac{5}{42}$.

**20.** $\frac{1}{4}$.

## CHAPTER 3

**1.** $\frac{1}{4}$.

**2.** $\frac{k+1}{2k+1}$.

**3.** $\frac{1}{\sqrt[3]{2}}$.

**4.** Mode of $X = 0$ and median of $X = 0$.

**5.** $\theta \ln\left(\frac{10}{9}\right)$.

**6.** $2 \ln 2$.

**7.** 0.25.

**8.** $f(2) = 0.5$, $f(3) = 0.2$, $f(\pi) = 0.3$.

**9.** $f(x) = \frac{1}{6}x^3 e^{-x}$.

**10.** $\frac{3}{4}$.

**11.** $a = 500$, mode $= 0.2$, and $P(X \geq 0.2) = 0.6766$.

**12.** 0.5.

**13.** 0.5.

**14.** $1 - F(-y)$.

**15.** $\frac{1}{4}$.

**16.** $R_X = \{3, 4, 5, 6, 7, 8, 9\}$;
$f(3) = f(4) = \frac{2}{20}$, $f(5) = f(6) = f(7) = \frac{4}{20}$, $f(8) = f(9) = \frac{2}{20}$.

**17.** $R_X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$;
$f(2) = \frac{1}{36}$, $f(3) = \frac{2}{36}$, $f(4) = \frac{3}{36}$, $f(5) = \frac{4}{36}$, $f(6) = \frac{5}{36}$, $f(7) = \frac{6}{36}$, $f(8) = \frac{5}{36}$, $f(9) = \frac{4}{36}$, $f(10) = \frac{3}{36}$, $f(11) = \frac{2}{36}$, $f(12) = \frac{1}{36}$.

**18.** $R_X = \{0, 1, 2, 3, 4, 5\}$;
$f(0) = \frac{59049}{10^5}$, $f(1) = \frac{32805}{10^5}$, $f(2) = \frac{7290}{10^5}$, $f(3) = \frac{810}{10^5}$, $f(4) = \frac{45}{10^5}$, $f(5) = \frac{1}{10^5}$.

**19.** $R_X = \{1, 2, 3, 4, 5, 6, 7\}$;
$f(1) = 0.4$, $f(2) = 0.2666$, $f(3) = 0.1666$, $f(4) = 0.0952$, $f(5) = 0.0476$, $f(6) = 0.0190$, $f(7) = 0.0048$.

**20.** $c = 1$ and $P(X = even) = \frac{1}{4}$.

**21.** $c = \frac{1}{2}$, $P(1 \leq X \leq 2) = \frac{3}{4}$.

**22.** $c = \frac{3}{2}$ and $P\left(X \leq \frac{1}{2}\right) = \frac{3}{16}$.

## CHAPTER 4

**1.** $-0.995$.

**2.** (a) $\frac{1}{33}$, (b) $\frac{12}{33}$, (c) $\frac{65}{33}$.

**3.** (c) 0.25, (d) 0.75, (e) 0.75, (f) 0.

**4.** (a) 3.75, (b) 2.6875, (c) 10.5, (d) 10.75, (e) $-71.5$.

**5.** (a) 0.5, (b) $\pi$, (c) $\frac{3}{10}\,\pi$.

**6.** $\frac{17}{24}\,\frac{1}{\sqrt{\theta}}$.

**7.** $\sqrt[4]{\frac{1}{E(x^2)}}$.

**8.** $\frac{8}{3}$.

**9.** 280.

**10.** $\frac{9}{20}$.

**11.** 5.25.

**12.** $a = \frac{4h^3}{\sqrt{\pi}}$, $E(X) = \frac{2}{h\sqrt{\pi}}$, $Var(X) = \frac{1}{h^2}\,\left[\frac{3}{2} - \frac{4}{\pi}\right]$.

**13.** $E(X) = \frac{7}{4}$, $E(Y) = \frac{7}{8}$.

**14.** $-\frac{2}{38}$.

**15.** $-38$.

**16.** $M(t) = 1 + 2t + 6t^2 + \cdots$.

**17.** $\frac{1}{4}$.

**18.** $\beta^n\,\prod_{i=1}^{n-1}(k+i)$.

**19.** $\frac{1}{4}\,\left[3e^{2t} + e^{3t}\right]$.

**20.** 120.

**21.** $E(X) = 3,\ Var(X) = 2$.

**22.** 11.

**23.** $c = E(X)$.

**24.** $F(c) = 0.5$.

**25.** $E(X) = 0,\ Var(X) = 2$.

**26.** $\frac{1}{625}$.

**27.** 38.

**28.** $a = 5$ and $b = -34$ or $a = -5$ and $b = 36$.

**29.** $-0.25$.

**30.** 10.

**31.** $-\frac{1}{1-p}\,p\ln p$.

## CHAPTER 5

**1.** $\frac{5}{16}$.

**2.** $\frac{5}{16}$.

**3.** $\frac{25}{72}$.

**4.** $\frac{4375}{279936}$.

**5.** $\frac{3}{8}$.

**6.** $\frac{11}{16}$.

**7.** 0.008304.

**8.** $\frac{3}{8}$.

**9.** $\frac{1}{4}$.

**10.** 0.671.

**11.** $\frac{1}{16}$.

**12.** 0.0000399994.

**13.** $\frac{n^2-3n+2}{2^{n+1}}$.

**14.** 0.2668.

**15.** $\frac{\binom{6}{3-k}\binom{4}{k}}{\binom{10}{3}}, \quad 0 \le k \le 3.$

**16.** 0.4019.

**17.** $1 - \frac{1}{e^2}$.

**18.** $\binom{x-1}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{x-3}$.

**19.** $\frac{5}{16}$.

**20.** 0.22345.

**21.** 1.43.

**22.** 24.

**23.** 26.25.

**24.** 2.

**25.** 0.3005.

**26.** $\frac{4}{e^4-1}$.

**27.** 0.9130.

**28.** 0.1239.

## CHAPTER 6

**1.** $f(x) = e^{-x} \qquad 0 < x < \infty.$

**2.** $Y \sim UNIF(0, 1).$

**3.** $f(w) = \frac{1}{w\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{\ln w - \mu}{\sigma}\right)^2}.$

**4.** 0.2313.

**5.** $3 \ln 4.$

**6.** $20.1\sigma.$

**7.** $\frac{3}{4}.$

**8.** 2.0.

**9.** 53.04.

**10.** 44.5314.

**11.** 75.

**12.** 0.4649.

**13.** $\frac{n!}{\theta^n}.$

**14.** 0.8664.

**15.** $e^{\frac{1}{2}k^2}.$

**16.** $\frac{1}{a}.$

**17.** 64.3441.

**18.** $g(y) = \begin{cases} \frac{4}{y^3} & \text{if } 0 < y < \sqrt{2} \\ 0 & \text{otherwise.} \end{cases}$

**19.** 0.5.

**20.** 0.7745.

**21.** 0.4.

**22.** $\frac{2}{3\theta}\, y^{-\frac{1}{3}} e^{-\frac{y^{\frac{2}{3}}}{\theta}}.$

**23.** $\frac{4}{\sqrt{2\pi}}\, y\, e^{-\frac{y^4}{2}}.$

**24.** $\ln(X) \sim \bigwedge\!\!\backslash(\mu, \sigma^2).$

**25.** $e^{\mu - \sigma^2}.$

**26.** $e^{\mu}.$

**27.** 0.3669.

**29.** $Y \sim GBETA(\alpha, \beta, a, b).$

**32.** (i) $\frac{1}{2}\sqrt{\pi}$, (ii) $\frac{1}{2}$, (iii) $\frac{1}{4}\sqrt{\pi}$, (iv) $\frac{1}{2}$.

**33.** (i) $\frac{1}{180}$, (ii) $(100)^{13}\, \frac{5!\, 7!}{13!}$, (iii) $\frac{1}{360}$.

**35.** $\left(1 - \frac{\alpha}{\beta}\right)^2.$

**36.** $E(X^n) = \theta^n\, \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)}.$

## CHAPTER 7

**1.** $f_1(x) = \frac{2x+3}{21}$, and $f_2(y) = \frac{3y+6}{21}$.

**2.** $f(x,y) = \begin{cases} \frac{1}{36} & \text{if } 1 < x < y = 2x < 12 \\ \frac{2}{36} & \text{if } 1 < x < y < 2x < 12 \\ 0 & \text{otherwise.} \end{cases}$

**3.** $\frac{1}{18}$.

**4.** $\frac{1}{2e^4}$.

**5.** $\frac{1}{3}$.

**6.** $f_1(x) = \begin{cases} 2(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$

**7.** $\frac{(e^2-1)(e-1)}{e^5}$.

**8.** 0.2922.

**9.** $\frac{5}{7}$.

**10.** $f_1(x) = \begin{cases} \frac{5}{48}x(8-x^3) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$

**11.** $f_2(y) = \begin{cases} 2y & \text{if } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$

**12.** $f(y/x) = \begin{cases} \frac{1}{1+\sqrt{1-(x-1)^2}} & \text{if } \ (x-1)^2 + (y-1)^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}$

**13.** $\frac{6}{7}$.

**14.** $f(y/x) = \begin{cases} \frac{1}{2x} & \text{if } 0 < y < 2x < 1 \\ 0 & \text{otherwise.} \end{cases}$

**15.** $\frac{4}{9}$.

**16.** $g(w) = 2e^{-w} - 2e^{-2w}$.

**17.** $g(w) = \left(1 - \frac{w^3}{\theta^3}\right) \frac{6w^2}{\theta^3}$.

**18.** $\frac{11}{36}$.

**19.** $\frac{7}{12}$.

**20.** $\frac{5}{6}$.

**21.** No.

**22.** Yes.

**23.** $\frac{7}{32}$.

**24.** $\frac{1}{4}$.

**25.** $\frac{1}{2}$.

**26.** $x\, e^{-x}$.

## CHAPTER 8

**1.** 13.

**2.** $Cov(X,Y) = 0$. Since $0 = f(0,0) \neq f_1(0)f_2(0) = \frac{1}{4}$, $X$ and $Y$ are not independent.

**3.** $\frac{1}{\sqrt{8}}$.

**4.** $\frac{1}{(1-4t)(1-6t)}$.

**5.** $X + Y \sim BIN(n+m, p)$.

**6.** $\frac{1}{2}\left(X^2 - Y^2\right) \sim EXP(1)$.

**7.** $M(s,t) = \frac{e^s-1}{s} + \frac{e^t-1}{t}$.

**9.** $-\frac{15}{16}$.

**10.** $Cov(X,Y) = 0$. No.

**11.** $a = \frac{6}{8}$ and $b = \frac{9}{8}$.

**12.** $Cov = -\frac{45}{112}$.

**13.** $Corr(X,Y) = -\frac{1}{5}$.

**14.** 136.

**15.** $\frac{1}{2}\sqrt{1+\rho}$.

**16.** $(1-p+pe^t)(1-p+pe^{-t})$.

**17.** $\frac{\sigma^2}{n}[1+(n-1)\rho]$.

**18.** 2.

**19.** $\frac{4}{3}$.

**20.** 1.

**21.** $\frac{1}{2}$.

## CHAPTER 9

**1.** 6.

**2.** $\frac{1}{2}(1+x^2)$.

**3.** $\frac{1}{2}\,y^2$.

**4.** $\frac{1}{2}+x$.

**5.** $2x$.

**6.** $\mu_X = -\frac{22}{3}$ and $\mu_Y = \frac{112}{9}$.

**7.** $\frac{1}{3}\,\frac{2+3y-28y^3}{1+2y-8y^2}$.

**8.** $\frac{3}{2}\,x$.

**9.** $\frac{1}{2}\,y$.

**10.** $\frac{4}{3}\,x$.

**11.** 203.

**12.** $15-\frac{1}{\pi}$.

**13.** $\frac{1}{12}\,(1-x)^2$.

**14.** $\frac{1}{12}\,\left(1-x^2\right)^2$.

**15.** $\frac{5}{192}$.

**16.** $\frac{1}{12}$.

**17.** 180.

**19.** $\frac{x}{6}+\frac{5}{12}$.

**20.** $\frac{x}{2}+1$.

## CHAPTER 10

**1.** $g(y) = \begin{cases} \frac{1}{2} + \frac{1}{4\sqrt{y}} & \text{for } 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$

**2.** $g(y) = \begin{cases} \frac{3}{16}\,\frac{\sqrt{y}}{m\sqrt{m}} & \text{for } 0 \le y \le 4m \\ 0 & \text{otherwise.} \end{cases}$

**3.** $g(y) = \begin{cases} 2y & \text{for } 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$

**4.** $g(z) = \begin{cases} \frac{1}{16}\,(z+4) & \text{for } -4 \le z \le 0 \\ \frac{1}{16}\,(4-z) & \text{for } 0 \le z \le 4 \\ 0 & \text{otherwise.} \end{cases}$

**5.** $g(z,x) = \begin{cases} \frac{1}{2}\,e^{-x} & \text{for } 0 < x < z < 2 + x < \infty \\ 0 & \text{otherwise.} \end{cases}$

**6.** $g(y) = \begin{cases} \frac{4}{y^3} & \text{for } 0 < y < \sqrt{2} \\ 0 & \text{otherwise.} \end{cases}$

**7.** $g(z) = \begin{cases} \frac{z^3}{15000} - \frac{z^2}{250} + \frac{z}{25} & \text{for } 0 \le z \le 10 \\ \frac{8}{15} - \frac{2z}{25} - \frac{z^2}{250} - \frac{z^3}{15000} & \text{for } 10 \le z \le 20 \\ 0 & \text{otherwise.} \end{cases}$

**8.** $g(u) = \begin{cases} \frac{4a^2}{u^3}\,\ln\left(\frac{u-a}{a}\right) + \frac{2a\,(u-2a)}{u^2\,(u-a)} & \text{for } 2a \le u < \infty \\ 0 & \text{otherwise.} \end{cases}$

**9.** $h(y) = \frac{3z^2-2z+1}{216}, \;\; z = 1, 2, 3, 4, 5, 6.$

**10.** $g(z) = \begin{cases} \frac{4h^3}{m\sqrt{\pi}}\,\sqrt{\frac{2z}{m}}\,e^{-\frac{2h^2 z}{m}} & \text{for } 0 \le z < \infty \\ 0 & \text{otherwise.} \end{cases}$

**11.** $g(u,v) = \begin{cases} -\frac{3u}{350} + \frac{9v}{350} & \text{for } 10 \le 3u + v \le 20,\; u \ge 0,\; v \ge 0 \\ 0 & \text{otherwise.} \end{cases}$

**12.** $g_1(u) = \begin{cases} \frac{2u}{(1+u)^3} & \text{if } 0 \le u < \infty \\ 0 & \text{otherwise.} \end{cases}$

**13.** $g(u,v) = \begin{cases} \frac{5\left[9v^3-5u^2v+3uv^2+u^3\right]}{32768} & \text{for } 0 < 2v+2u < 3v-u < 16 \\ 0 & \text{otherwise.} \end{cases}$

**14.** $g(u,v) = \begin{cases} \frac{u+v}{32} & \text{for } 0 < u+v < 2\sqrt{5v-3u} < 8 \\ 0 & \text{otherwise.} \end{cases}$

**15.** $g_1(u) = \begin{cases} 2+4u+2u^2 & \text{if } -1 \le u \le 0 \\ 2\sqrt{1-4u} & \text{if } 0 \le u \le \frac{1}{4} \\ 0 & \text{otherwise.} \end{cases}$

**16.** $g_1(u) = \begin{cases} \frac{4}{3}\,u & \text{if } 0 \le u \le 1 \\ \frac{4}{3}\,u^{-5} & \text{if } 1 \le u < \infty \\ 0 & \text{otherwise.} \end{cases}$

**17.** $g_1(u) = \begin{cases} 4u^{\frac{1}{3}} - 4u & \text{if } 0 \le u \le 1 \\ 0 & \text{otherwise.} \end{cases}$

**18.** $g_1(u) = \begin{cases} 2\,u^{-3} & \text{if } 1 \le u < \infty \\ 0 & \text{otherwise.} \end{cases}$

**19.** $f(w) = \begin{cases} \frac{w}{6} & \text{if } 0 \le w \le 2 \\ \frac{2}{6} & \text{if } 2 \le w \le 3 \\ \frac{5-w}{6} & \text{if } 3 \le w \le 5 \\ 0 & \text{otherwise.} \end{cases}$

**20.** $BIN(2n,p)$

**21.** $GAM(\theta,2)$

**22.** $CAU(0)$

**23.** $N(2\mu, 2\sigma^2)$

**24.** $f_1(\alpha) = \begin{cases} \frac{1}{4}\,(2-|\alpha|) & \text{if } |\alpha| \le 2 \\ 0 & \text{otherwise,} \end{cases}$ $f_2(\beta) = \begin{cases} -\frac{1}{2}\,\ln(|\beta|) & \text{if } |\beta| \le 1 \\ 0 & \text{otherwise.} \end{cases}$

## CHAPTER 11

**2.** $\frac{7}{10}$.

**3.** $\frac{960}{7^5}$.

**6.** 0.7627.

# CHAPTER 12

## CHAPTER 13

**3.** 0.115.

**4.** 1.0.

**5.** $\frac{7}{16}$.

**6.** 0.352.

**7.** $\frac{6}{5}$.

**8.** 100.64.

**9.** $\frac{1+\ln(2)}{2}$.

**10.** $[1 - F(x_6)]^5$.

**11.** $\theta + \frac{1}{5}$.

**12.** $2\, e^{-w}\, [1 - e^{-w}]$.

**13.** $6\, \frac{w^2}{\theta^3} \left(1 - \frac{w^3}{\theta^3}\right)$.

**14.** $N(0, 1)$.

**15.** 25.

**16.** $X$ has a degenerate distribution with MGF $M(t) = e^{\frac{1}{2}t}$.

**17.** $POI(1995\lambda)$.

**18.** $\left(\frac{1}{2}\right)^n (n + 1)$.

**19.** $\frac{8^8}{11^9}\, 35$.

**20.** $f(x) = \frac{60}{\theta} \left(1 - e^{-\frac{x}{\theta}}\right)^3 e^{-\frac{3x}{\theta}}$ for $0 < x < \infty$.

**21.** $X_{(n+1)} \sim Beta(n + 1, n + 1)$.

## CHAPTER 14

**1.** $N(0, 32)$.

**2.** $\chi^2(3)$; the MGF of $X_1^2 - X_2^2$ is $M(t) = \frac{1}{\sqrt{1-4t^2}}$.

**3.** $t(3)$.

**4.** $f(x_1, x_2, x_3) = \frac{1}{\theta^3} e^{-\frac{(x_1+x_2+x_3)}{\theta}}$.

**5.** $\sigma^2$

**6.** $t(2)$.

**7.** $M(t) = \frac{1}{\sqrt{(1-2t)(1-4t)(1-6t)(1-8t)}}$.

**8.** 0.625.

**9.** $\frac{\sigma^4}{n^2} 2(n-1)$.

**10.** 0.

**11.** 27.

**12.** $\chi^2(2n)$.

**13.** $t(n+p)$.

**14.** $\chi^2(n)$.

**15.** $(1, 2)$.

**16.** 0.84.

**17.** $\frac{2\sigma^2}{n^2}$.

**18.** 11.07.

**19.** $\chi^2(2n-2)$.

**20.** 2.25.

**21.** 6.37.

## CHAPTER 15

**1.** $\sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}$.

**2.** $\frac{1}{\bar{X}-1}$.

**3.** $\frac{2}{\bar{X}}$.

**4.** $-\frac{n}{\sum_{i=1}^{n} \ln X_i}$.

**5.** $\frac{n}{\sum_{i=1}^{n} \ln X_i} - 1$.

**6.** $\frac{2}{\bar{X}}$.

**7.** 4.2

**8.** $\frac{19}{26}$.

**9.** $\frac{15}{4}$.

**10.** 2.

**11.** $\hat{\alpha} = 3.534$ and $\hat{\beta} = 3.409$.

**12.** 1.

**13.** $\frac{1}{3}\max\{x_1, x_2, ..., x_n\}$.

**14.** $\sqrt{1 - \frac{1}{\max\{x_1, x_2, ..., x_n\}}}$.

**15.** 0.6207.

**18.** 0.75.

**19.** $-1 + \frac{5}{\ln(2)}$.

**20.** $\frac{\bar{X}}{1+\bar{X}}$.

**21.** $\frac{\bar{X}}{4}$.

**22.** 8.

**23.** $\frac{n}{\sum_{i=1}^{n} |X_i - \mu|}$.

**24.** $\frac{1}{N}$.

**25.** $\sqrt{\overline{X}}$.

**26.** $\hat{\lambda} = \frac{n\bar{X}}{(n-1)S^2}$ and $\hat{\alpha} = \frac{n\bar{X}^2}{(n-1)S^2}$.

**27.** $\frac{10\,n}{p\,(1-p)}$.

**28.** $\frac{2n}{\theta^2}$.

**29.** $\begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$.

**30.** $\begin{pmatrix} \frac{n\lambda}{\mu^3} & 0 \\ 0 & \frac{n}{2\lambda^2} \end{pmatrix}$.

**31.** $\widehat{\alpha} = \frac{\overline{X}}{\widehat{\beta}}, \quad \widehat{\beta} = \frac{1}{\overline{X}}\,\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}\,\right]$.

**32.** $\widehat{\theta}$ is obtained by solving numerically the equation $\sum_{i=1}^{n} \frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0$.

**33.** $\widehat{\theta}$ is the median of the sample.

**34.** $\frac{n}{\lambda}$.

**35.** $\frac{n}{(1-p)\,p^2}$.

## CHAPTER 16

**1.** $b = \frac{\sigma_2^2 - cov(T_1,T_2)}{\sigma_1^2+\sigma_2^2-2cov(T_1,T_2}$.

**2.** $\widehat{\theta} = \overline{|X|}$, $E(\overline{|X|}) = \theta$, unbiased.

**4.** $n = 20$.

**5.** $k = \frac{1}{2}$.

**6.** $a = \frac{25}{61}$, $b = \frac{36}{61}$, $\widehat{c} = 12.47$.

**7.** $\sum_{i=1}^{n} X_i^3$.

**8.** $\sum_{i=1}^{n} X_i^2$, no.

**9.** $k = \frac{4}{\pi}$.

**10.** $k = 2$.

**11.** $k = 2$.

**13.** $\ln \prod_{i=1}^{n} (1 + X_i)$.

**14.** $\sum_{i=1}^{n} X_i^2$.

**15.** $X_{(1)}$, and sufficient.

**16.** $X_{(1)}$ is biased and $\overline{X} - 1$ is unbiased. $X_{(1)}$ is efficient then $\overline{X} - 1$.

**17.** $\sum_{i=1}^{n} \ln X_i$.

**18.** $\sum_{i=1}^{n} X_i$.

**19.** $\sum_{i=1}^{n} \ln X_i$.

**22.** Yes.

**23.** Yes.

**24.** Yes.

**25.** Yes.

**26.** $\widehat{\theta} = 3\,\overline{X}$.

**27.** $\widehat{\theta} = \frac{50}{30}\,X$.

## CHAPTER 17

**7.** The pdf of $Q$ is $g(q) = \begin{cases} n\, e^{-n\, q} & \text{if } 0 < q < \infty \\ 0 & \text{otherwise.} \end{cases}$

The confidence interval is $\left[X_{(1)} - \frac{1}{n}\, \ln\left(\frac{2}{\alpha}\right),\ \ X_{(1)} - \frac{1}{n}\, \ln\left(\frac{2}{2-\alpha}\right)\right]$.

**8.** The pdf of $Q$ is $g(q) = \begin{cases} \frac{1}{2}\, e^{-\frac{1}{2}\, q} & \text{if } 0 < q < \infty \\ 0 & \text{otherwise.} \end{cases}$

The confidence interval is $\left[X_{(1)} - \frac{1}{n}\, \ln\left(\frac{2}{\alpha}\right),\ \ X_{(1)} - \frac{1}{n}\, \ln\left(\frac{2}{2-\alpha}\right)\right]$.

**9.** The pdf of $Q$ is $g(q) = \begin{cases} n\, q^{n-1} & \text{if } 0 < q < 1 \\ 0 & \text{otherwise.} \end{cases}$

The confidence interval is $\left[X_{(1)} - \frac{1}{n}\, \ln\left(\frac{2}{\alpha}\right),\ \ X_{(1)} - \frac{1}{n}\, \ln\left(\frac{2}{2-\alpha}\right)\right]$.

**10.** The pdf $g(q)$ of $Q$ is given by $g(q) = \begin{cases} n\, q^{n-1} & \text{if } 0 \le q \le 1 \\ 0 & \text{otherwise.} \end{cases}$

The confidence interval is $\left[\left(\frac{2}{\alpha}\right)^{\frac{1}{n}}\ X_{(n)},\ \ \left(\frac{2}{2-\alpha}\right)^{\frac{1}{n}}\ X_{(n)}\right]$.

**11.** The pdf of $Q$ is given by $g(q) = \begin{cases} n\,(n-1)\, q^{n-2}\,(1-q) & \text{if } 0 \le q \le 1 \\ 0 & \text{otherwise.} \end{cases}$

**12.** $\left[X_{(1)} - z_{\frac{\alpha}{2}}\, \frac{1}{\sqrt{n}},\ X_{(1)} + z_{\frac{\alpha}{2}}\, \frac{1}{\sqrt{n}}\right]$.

**13.** $\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\, \frac{\widehat{\theta}+1}{\sqrt{n}},\ \widehat{\theta} + z_{\frac{\alpha}{2}}\, \frac{\widehat{\theta}+1}{\sqrt{n}}\right]$, where $\widehat{\theta} = -1 + \frac{n}{\sum_{i=1}^{n} \ln x_i}$.

**14.** $\left[\frac{2}{\overline{X}} - z_{\frac{\alpha}{2}} \sqrt{\frac{2}{n\, \overline{X}^2}},\ \frac{2}{\overline{X}} + z_{\frac{\alpha}{2}} \sqrt{\frac{2}{n\, \overline{X}^2}}\right]$.

**15.** $\left[\overline{X} - 4 - z_{\frac{\alpha}{2}}\, \frac{\overline{X}-4}{\sqrt{n}},\ \overline{X} - 4 + z_{\frac{\alpha}{2}}\, \frac{\overline{X}-4}{\sqrt{n}}\right]$.

**16.** $\left[X_{(n)} - z_{\frac{\alpha}{2}}\, \frac{X_{(n)}}{(n+1)\, \sqrt{n+2}},\ X_{(n)} + z_{\frac{\alpha}{2}}\, \frac{X_{(n)}}{(n+1)\, \sqrt{n+2}}\right]$.

**17.** $\left[\frac{1}{4}\, \overline{X} - z_{\frac{\alpha}{2}}\, \frac{\overline{X}}{8\, \sqrt{n}},\ \frac{1}{4}\, \overline{X} + z_{\frac{\alpha}{2}}\, \frac{\overline{X}}{8\, \sqrt{n}}\right]$.

## CHAPTER 18

**1.** $\alpha = 0.03125$ and $\beta = 0.763$.

**2.** Do not reject $H_o$.

**3.** $\alpha = 0.0511$ and $\beta(\lambda) = 1 - \sum_{x=0}^{7} \frac{(8\lambda)^x e^{-8\lambda}}{x!},\ \lambda \neq 0.5$.

**4.** $\alpha = 0.08$ and $\beta = 0.46$.

**5.** $\alpha = 0.19$.

**6.** $\alpha = 0.0109$.

**7.** $\alpha = 0.0668$ and $\beta = 0.0062$.

**8.** $C = \{(x_1, x_2) \,|\, \overline{x}^2 \geq 3.9395\}$.

**9.** $C = \{(x_1, ..., x_{10}) \,|\, \overline{x} \geq 0.3\}$.

**10.** $C = \{x \in [0, 1] \,|\, x \geq 0.829\}$.

**11.** $C = \{(x_1, x_2) \,|\, x_1 + x_2 \geq 5\}$.

**12.** $C = \{(x_1, ..., x_8) \,|\, \overline{x} - \overline{x} \ln \overline{x} \leq a\}$.

**13.** $C = \{(x_1, ..., x_n) \,|\, 35 \ln \overline{x} - \overline{x} \leq a\}$.

**14.** $C = \left\{(x_1, ..., x_5) \,|\, \left(\frac{\overline{x}}{2\overline{x}-2}\right)^{5\overline{x}-5} \overline{x}^5 \leq a\right\}$.

**15.** $C = \{(x_1, x_2, x_3) \mid \ |\overline{x} - 3| \geq 1.96\}$.

**16.** $C = \left\{(x_1, x_2, x_3) \mid \ \overline{x}\, e^{-\frac{1}{3}\overline{x}} \leq a\right\}$.

**17.** $C = \left\{(x_1, x_2, ..., x_n) \mid \ \left(\frac{e}{10\,\overline{x}}\right)^{3\overline{x}} \leq a\right\}$.

**18.** $\frac{1}{3}$.

**19.** $C = \left\{(x_1, x_2, x_3) \mid \ x_{(3)} \leq \sqrt[3]{117}\ \right\}$.

**20.** $C = \{(x_1, x_2, x_3) \mid \ \overline{x} \geq 12.04\}$.

**21.** $\alpha = \frac{1}{16}$ and $\beta = \frac{255}{256}$.

**22.** $\alpha = 0.05$.