MODULE “STATISTICAL DATA ANALYSIS”

By Dr. Dang Quang A and Dr. Bui The Hong
Institute of Information Technology
Hoang Quoc Viet road, Cau giay, HANOI

Preface

Statistics is the science of collecting, organizing and interpreting numerical and non-numerical facts, which we call data.

The collection and study of data is important in the work of many professions, so that training in the science of statistics is valuable preparation for variety of careers, for example, economists and financial advisors, businessmen, engineers and farmers.

Knowledge of probability and statistical methods also are useful for informatics specialists in various fields such as data mining, knowledge discovery, neural networks, and fuzzy systems and so on.

Whatever else it may be, statistics is first and foremost a collection of tools used for converting raw data into information to help decision makers in their work.

The science of data - statistics - is the subject of this course.

Chapter 1 is an introduction into statistical analysis of data. Chapters 2 and 3 deal with statistical methods for presenting and describing data. Chapters 4 and 5 introduce the basic concepts of probability and probability distributions, which are the foundation for our study of statistical inference in later chapters. Sampling and sampling distributions is the subject of Chapter 6. The remaining seven chapters discuss statistical inference - methods for drawing conclusions from properly produced data. Chapter 7 deals with estimating characteristics of a population by observing the characteristic of a sample. Chapters 8 to 13 describe some of the most common methods of inference: for drawing conclusions about means, proportions and variances from one and two samples, about relations in categorical data, regression and correlation and analysis of variance. In every chapter we include examples to illustrate the concepts and methods presented. The use of computer packages such as SPSS and STATGRAPHICS will be evolved.

Audience

This tutorial as an introductory course to statistics is intended mainly for users such as engineers, economists and managers who need to use statistical methods in their work and for students. However, many aspects will be useful for computer trainers.

Objectives

Understanding statistical reasoning
Mastering basic statistical methods for analyzing data such as descriptive and inferential methods
Ability to use methods of statistics in practice with the help of computer software

Entry requirements

High school algebra course (+elements of calculus)
Elementary computer skills

http://www.netnam.vn/unescocourse/statistics/stat_frm.htm

CONTENTS

THE STATISTICAL ANALYSIS OF DATA

Chapter 1 Introduction

CONTENTS

1.1 What is Statistics

The word *statistics* in our everyday life means different things to different people. To a football fan, *statistics* are the information about rushing yardage, passing yardage, and first downs, given a halftime. To a manager of a power generating station, *statistics* may be information about the quantity of pollutants being released into the atmosphere. To a school principal, *statistics* are information on the absenteeism, test scores and teacher salaries. To a medical researcher investigating the effects of a new drug, *statistics* are evidence of the success of research efforts. And to a college student, *statistics* are the grades made on all the quizzes in a course this semester.

Each of these people is using the word *statistics* correctly, yet each uses it in a slightly different way and for a somewhat different purpose*. Statistics* is a word that can refer to quantitative data or to a field of study.

As a field of study, *statistics* is the science of collecting, organizing and interpreting numerical facts, which we call data. We are bombarded by data in our everyday life. The collection and study of data are important in the work of many professions, so that training in the science of *statistics* is valuable preparation for variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisors as well as policy makers in government and business study these data in order to make informed decisions. Farmers study data from field trials of new crop varieties. Engineers gather data on the quality and reliability of manufactured of products. Most areas of academic study make use of numbers, and therefore also make use of methods of statistics.

Whatever else it may be, *statistics* is, first and foremost, a collection of tools used for converting raw data into information to help decision makers in their works.

The science of data - *statistics* - is the subject of this course.

1.2 Populations and samples

In statistics, the data set that is the target of your interest is called a population. Notice that, a statistical population does not refer to people as in our everyday usage of the term; it refers to a collection of data.

Definition 1.1

A population is a collection (or set) of data that describes some phenomenon of interest to you.

Definition 1.2

A sample is a subset of data selected from a population

Example 1.1 The population may be all women in a country, for example, in Vietnam. If from each city or province we select 50 women, then the set of selected women is a sample.

Example 1.2 The set of all whisky bottles produced by a company is a population. For the quality control 150 whisky bottles are selected at random. This portion is a sample.

1.3 Descriptive and inferential statistics

If you have every measurement (or observation) of the population in hand, then statistical methodology can help you to describe this typically large set of data. We will find graphical and numerical ways to make sense out of a large mass of data. The branch of statistics devoted to this application is called descriptive statistics.

Definition 1.3
The branch of statistics devoted to the summarization and description of data (population or sample) is called descriptive statistics.

If it may be too expensive to obtain or it may be impossible to acquire every measurement in the population, then we will want to select a sample of data from the population and use the sample to infer the nature of the population.

Definition 1.4
The branch of statistics concerned with using sample data to make an inference about a population of data is called inferential statistics.

1.4 Brief history of statistics

The word *statistik* comes from the Italian word *statista* (meaning "statesman"). It was first used by Gottfried Achenwall (1719-1772), a professor at Marlborough and Gottingen. Dr. E.A.W. Zimmermam introduced the word *statistics* to England. Its use was popularized by Sir John Sinclair in his work "*Statistical Account of Scotland 1791-1799*". Long before the eighteenth century, however, people had been recording and using data.

Official government statistics are as old as recorded history. The emperor Yao had taken a census of the population in China in the year 2238 B.C. The Old Testament contains several accounts of census taking. Governments of ancient Babylonia, Egypt and Rome gathered detail records of population and resources. In the Middle Age, governments began to register the ownership of land. In A.D. 762 Charlemagne asked for detailed descriptions of church-owned properties. Early, in the ninth century, he completed a statistical enumeration of the serfs attached to the land. About 1086, William and Conqueror ordered the writing of the *Domesday Book*, a record of the ownership, extent, and value of the lands of England. This work was England's first statistical abstract.

Because of Henry VII's fear of the plague, England began to register its dead in 1532. About this same time, French law required the clergy to register baptisms, deaths and marriages. During an outbreak of the plague in the late 1500s, the English government started publishing weekly death statistics. This practice continued, and by 1632 these *Bills of Mortality* listed births and deaths by sex. In 1662, Captain John Graunt used thirty years of these Bills to make predictions about the number of persons who would die from various diseases and the proportion of male and female birth that could be expected. Summarized in his work, *Natural and Political Observations ...Made upon the Bills of Mortality*, Graunt's study was a pioneer effort in statistical analysis. For his achievement in using past records to predict future events, Graund was made a member of the original Royal Society.

The history of the development of statistical theory and practice is a lengthy one. We have only begun to list the people who have made significant contributions to this field. Later we will encounter others whose names are now attached to specific laws and methods. Many people have brought to the study of statistics refinements or innovations that, taken together, form the theoretical basis of what we will study in this course.

1.5 Computer softwares for statistical analysis

Many real problems have so much data that doing the calculations by hand is not feasible. For this reason, most real-world statistical analysis is done on computers. You must prepare the input data and interpret the results of the analysis and take appropriate action, but the machine does all the "number crunching". There many widely-used software packages for statistical analysis. Below we list some of them.

- Minitab (registered trademark of Minitab, Inc., University Park, Pa)
- SAS (registered trademark of SAS Institute, Inc., Cary, N.C.)
- SPSS (registered trademark of SPSS, Inc.,Chicago)
- SYSTAT (registered trademark of SYSTAT, Inc., Evanston,Il)
- STATGRAPHICS (registered trademark of Statistical Graphics Corp., Maryland).

Except for the above listed softwares it is possible to make simple statistical analysis of data by using the part "Data analysis" in Microsoft EXCEL.

Chapter 2 Data presentation

CONTENTS

2.1 Introduction

The objective of data description is to summarize the characteristics of a data set. Ultimately, we want to make the data set more comprehensible and meaningful. In this chapter we will show how to construct charts and graphs that convey the nature of a data set. The procedure that we will use to accomplish this objective in a particular situation depends on the type of data that we want to describe.

2.2 Types of data

Data can be one of two types, qualitative and quantitative**.**

Definition 2.1

Quantitative data are observations measured on a numerical scale.

In other words, quantitative data are those that represent the quantity or amount of something.

Example 2.1 Height (in centimeters), weight (in kilograms) of each student in a group are both quantitative data.

Definition 2.2
Non-numerical data that can only be classified into one of a group of categories are said to be qualitative data.

In other words, qualitative data are those that have no quantitative interpretation, i.e., they can only be classified into categories.

Example 2.2 Education level, nationality, sex of each person in a group of people are qualitative data.

2.3 Qualitative data presentation

When describing qualitative observations, we define the categories in such a way that each observations can fall in one and only one category. The data set is then described by giving the number of observations, or the proportion of the total number of observations that fall in each of the categories.

Definition 2.3
The category frequency for a given category is the number of observations that fall in that category.

Definition 2.4
The category relative frequency for a given category is the proportion of the total number of observations that fall in that category.

$$\text{Relative frequency for a category} = \frac{\text{Number of observations falling in that category}}{\text{Total number of observations}}$$

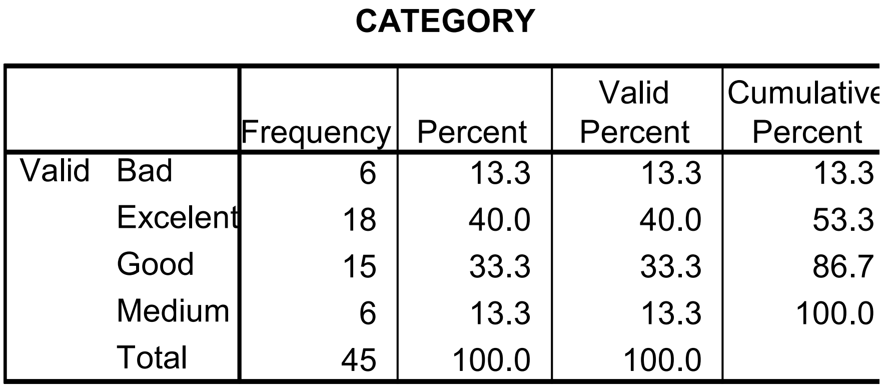
Instead of the relative frequency for a category one usually uses percentage for a category, which is computed as follows

Percentage for a category = Relative frequency for the category x 100%

Example 2.3 The classification of students of a group by the score on the subject “Statistical analysis” is presented in Table 2.0a. The table of frequencies for the data set generated by computer using the software SPSS is shown in Figure 2.1.

Table 2.0a *The classification of students*

No of Stud.	CATEGORY	No of Stud.	CATEGORY	No of Stud.	CATEGORY	No of stud	CATEGORY
1	Bad	13	Good	24	Good	35	Good
2	Medium	14	Excellent	25	Medium	36	Medium
3	Medium	15	Excellent	26	Bad	37	Good
4	Medium	16	Excellent	27	Good	38	Excellent
5	Good	17	Excellent	28	Bad	39	Good
6	Good	18	Good	29	Bad	40	Good
7	Excellent	19	Excellent	30	Good	41	Medium
8	Excellent	20	Excellent	31	Excellent	42	Bad
9	Excellent	21	Good	32	Excellent	43	Excellent
10	Excellent	22	Excellent	33	Excellent	44	Excellent
11	Bad	23	Excellent	34	Good	45	Good
12	Good						



CATEGORY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Bad	6	13.3	13.3	13.3
	Excelent	18	40.0	40.0	53.3
	Good	15	33.3	33.3	86.7
	Medium	6	13.3	13.3	100.0
	Total	45	100.0	100.0	

Figure 2.1 *Output from SPSS showing the frequency table for the variable CATEGORY.*

2.4 Graphical description of qualitative data

Bar graphs and pie charts are two of the most widely used graphical methods for describing qualitative data sets.

Bar graphs give the frequency (or relative frequency) of each category with the height or length of the bar proportional to the category frequency (or relative frequency).

Example 2.4a (Bar Graph) The bar graph generated by computer using SPSS for the variable CATEGORY is depicted in Figure 2.2.

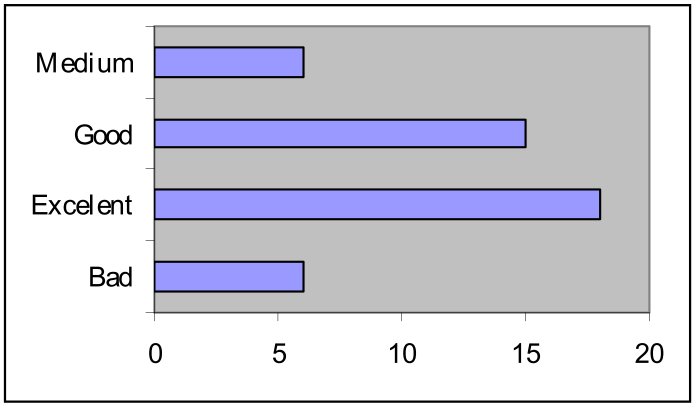


Figure 2.2 *Bar graph showing the number of students of each category*

Pie charts divide a complete circle (a pie) into slices, each corresponding to a category, with the central angle and hence the area of the slice proportional to the category relative frequency.

Example 2.4b (Pie Chart) The pie chart generated by computer using EXCEL CHARTS for the variable CATEGORY is depicted in Figure 2.3.

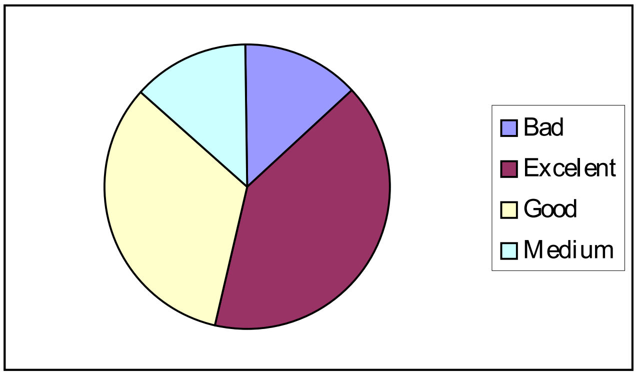


Figure 2.3 *Pie chart showing the number of students of each category*

2.5 Graphical description of quantitative data: Stem and Leaf displays

One of graphical methods for describing quantitative data is the stem and leaf display, which is widely used in exploratory data analysis when the data set is small.

In order to explain what is a stem and what is a leaf we consider the data from the table 2.0b. For this data for a two-digit number, for example, 79, we designate the first digit (7) as its stem; we call the last digit (9) its leaf; and for three-digit number, for example, 112, we designate the first two digit (12) as its stem; we also call the last digit (2) its leaf.

Steps to follow in constructing a Stem and Leaf Display

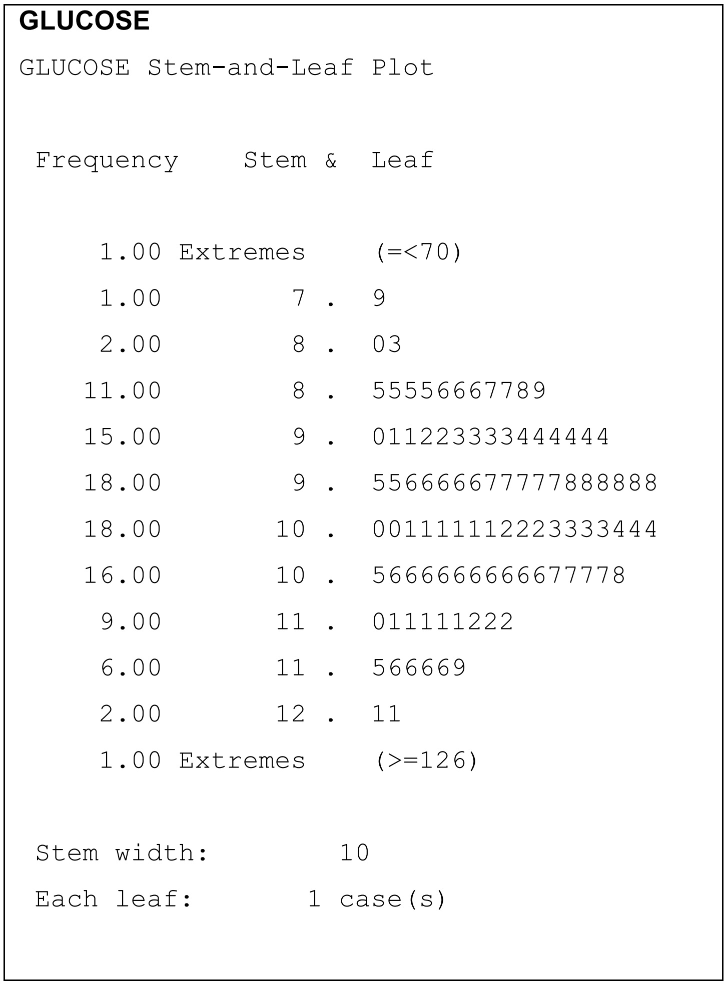
1. Divide each observation in the data set into two parts, the Stem and the Leaf.
2. List the stems in order in a column, starting with the smallest stem and ending with the largest.
3. Proceed through the data set, placing the leaf for each observation in the appropriate stem row.

Depending on the data, a display can use one, two or five lines per stem. Among the different stems, two-line stems are widely used.

Example 2.5 The quantity of glucose in blood of 100 persons is measured and recorded in Table 2.0b (unit is mg %). Using SPSS we obtain the following Stem-and-Leaf display for this data set.

Table 2.0b *Quantity of glucose in blood of 100 students (unit: mg %)*

70	79	80	83	85	85	85	85	86	86
86	87	87	88	89	90	91	91	92	92
93	93	93	93	94	94	94	94	94	94
95	95	96	96	96	96	96	97	97	97
97	97	98	98	98	98	98	98	100	100
101	101	101	101	101	101	102	102	102	103
103	103	103	104	104	104	105	106	106	106
106	106	106	106	106	106	106	107	107	107
107	108	110	111	111	111	111	111	112	112
112	115	116	116	116	116	119	121	121	126



```
GLUCOSE
GLUCOSE Stem-and-Leaf Plot


 Frequency    Stem &  Leaf


     1.00 Extremes    (=<70)
     1.00        7 .  9
     2.00        8 .  03
    11.00        8 .  55556667789
    15.00        9 .  011223333444444
    18.00        9 .  556666677777888888
    18.00       10 .  001111112223333444
    16.00       10 .  5666666666677778
     9.00       11 .  011111222
     6.00       11 .  566669
     2.00       12 .  11
     1.00 Extremes    (>=126)


 Stem width:        10
 Each leaf:       1 case(s)
```

Figure 2.4. *Output from SPSS showing the Stem-and-Leaf display for the data set of glucose*

The stem and leaf display of Figure 2.4 partitions the data set into 12 classes corresponding to 12 stems. Thus, here two-line stems are used. The number of leaves in each class gives the class frequency.

Advantages of a stem and leaf display over a frequency distribution *(considered in the next section):*

1. The original data are preserved.
2. A stem and leaf display arranges the data in an orderly fashion and makes it easy to determine certain numerical characteristics to be discussed in the following chapter.
3. The classes and numbers falling in them are quickly determined once we have selected the digits that we want to use for the stems and leaves.

Disadvantage of a stem and leaf display:

Sometimes not much flexibility in choosing the stems.

2.6 Tabulating quantitative data: Relative frequency distributions

Frequency distribution or relative frequency distribution is most often used in scientific publications to describe quantitative data sets. They are better suited to the description of large data sets and they permit a greater flexibility in the choice of class widths.

A frequency distribution is a table that organizes data into classes. It shows the number of observations from the data set that fall into each of classes. It should be emphasized that we always have in mind *non-overlapping classes*, i.e. classes without common items.

Steps for constructing a frequency distribution and relative frequency distribution:

1. Decide the type and number of classes for dividing the data set, lower limit and upper limit of the classes:

   Lower limit < Minimum of values

   Upper limit > Maximum of values

2. Determine the width of class intervals:

$$\text{Width of class intervals} = \frac{\text{Upper limit - Lower limit}}{\text{Total number of classes}}$$

3. For each class, count the number of observations that fall in that class. This number is called the class frequency.
4. Calculate each class relative frequency

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

Except for frequency distribution and relative frequency distribution one usually uses *relative class percentage*, which is calculated by the formula:

Relative class percentage = Class relative frequency x 100%

Example 2.6 Construct frequency table for the data set of quantity of glucose in blood of 100 persons recorded in Table 2.0b (unit is mg %).

Using the software STATGRAPHICS, taking Lower limit = 62, Upper limit = 150 and Total number of classes = 22 we obtained the following table.

Table 2.1 Frequency *distribution for glucose in blood of 100 persons*

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
0	62	66	64	0	0	0	0
1	66	70	68	1	0.01	1	0.01
2	70	74	72	0	0	1	0.01
3	74	78	76	0	0	1	0.01
4	78	82	80	2	0.02	3	0.03
5	82	86	84	8	0.08	11	0.11
6	86	90	88	5	0.05	16	0.16
7	90	94	92	14	0.14	30	0.3
8	94	98	96	18	0.18	48	0.48
9	98	102	100	11	0.11	59	0.59
10	102	106	104	18	0.18	77	0.77
11	106	110	108	6	0.06	83	0.83
12	110	114	112	8	0.08	91	0.91
13	114	118	116	5	0.05	96	0.96
14	118	122	120	3	0.03	99	0.99
15	122	126	124	1	0.01	100	1
16	126	130	128	0	0	100	1
17	130	134	132	0	0	100	1
18	134	138	136	0	0	100	1
19	138	142	140	0	0	100	1
20	142	146	144	0	0	100	1
21	146	150		0	0	100	1

Remarks:

1. All classes of frequency table must be mutually exclusive.
2. Classes may be open-ended when either the lower or the upper end of a quantitative classification scheme is limitless. For example

Class: age
birth to 7
8 to 15
....
64 to 71
72 and older

3. Classification schemes can be either discrete or continuous. Discrete classes are separate entities that do not progress from one class to the next without a break. Such class as the number of children in each family, the number of trucks owned by moving companies. Discrete data are data that can take only a limit number of values. Continuous data do progress from one class to the next without a break. They involve numerical measurement such as the weights of cans of tomatoes, the kilograms of pressure on concrete. Usually, continuous classes are half-open intervals. For example, the classes in Table 2.1 are half-open intervals [62, 66), [66, 70) ...

2.7 Graphical description of quantitative data: histogram and polygon

There is an old saying that "one picture is worth a thousand words". Indeed, statisticians have employed graphical techniques to describe sets of data more vividly. Bar charts and pie charts were presented in Figure 2.2 and Figure 2.3 to describe qualitative data. With quantitative data summarized into frequency, relative frequency tables, however, histograms and polygons are used to describe the data.

2.7.1 Histogram

When plotting histograms, the phenomenon of interest is plotted along the horizontal axis, while the vertical axis represents the number, proportion or percentage of observations per class interval – depending on whether or not the particular histogram is respectively, a frequency histogram, a relative frequency histogram or a percentage histogram.

Histograms are essentially vertical bar charts in which the rectangular bars are constructed at midpoints of classes.

Example 2.7 Below we present the frequency histogram for the data set of quantities of glucose, for which the frequency table is constructed in Table 2.1.

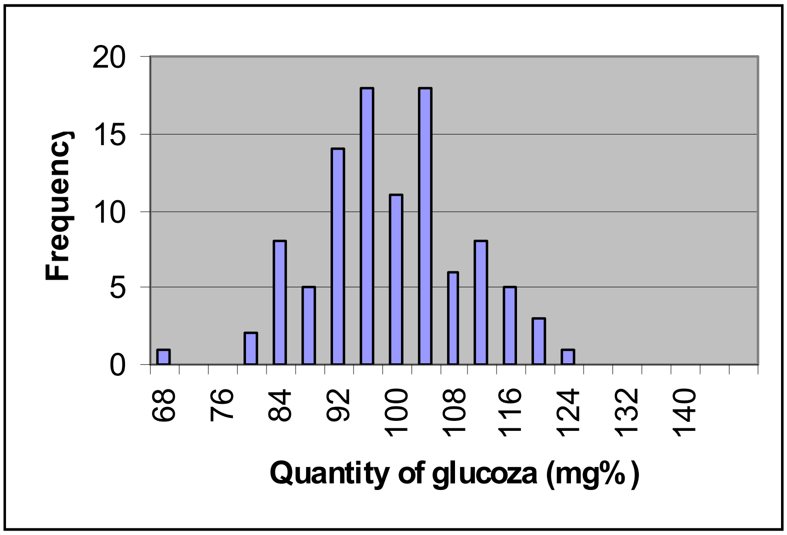


Figure 2.5 *Frequency histogram for quantities of glucose, tabulated in Table 2.1*

Remark: When comparing two or more sets of data, the various histograms can not be constructed on the same graph because superimposing the vertical bars of one on another would cause difficulty in interpretation. For such cases it is necessary to construct relative frequency or percentage polygons.

2.7.2 Polygons

As with histograms, when plotting polygons the phenomenon of interest is plotted along the horizontal axis while the vertical axis represents the number, proportion or percentage of observations per class interval – depending on whether or not the particular polygon is respectively, a frequency polygon, a relative frequency polygon or a percentage polygon. For example, the frequency polygon is a line graph connecting the midpoints of each class interval in a data set, plotted at a height corresponding to the frequency of the class.

Example 2.8 Figure 2.6 is a frequency polygon constructed from data in Table 2.1.

Figure 2.6 *Frequency polygon for data of glucose in Table 2.1*

Advantages of polygons:

- The frequency polygon is simpler than its histogram counterpart.
- It sketches an outline of the data pattern more clearly.
- The polygon becomes increasingly smooth and curve like as we increase the number of classes and the number of observations.

2.8 Cumulative distributions and cumulative polygons

Other useful methods of presentation which facilitate data analysis and interpretation are the construction of cumulative distribution tables and the plotting of cumulative polygons. Both may be developed from the frequency distribution table, the relative frequency distribution table or the percentage distribution table.

A cumulative frequency distribution enables us to see how many observations lie above or below certain values, rather than merely recording the number of items within intervals.

A "less-than" cumulative frequency distribution may be developed from the frequency table as follows:

Suppose a data set is divided into n classes by boundary points x_1, x_2, ..., x_n, x_{n+1}. Denote the classes by $C_1, C_2, ..., C_n$. Thus, the class $C_k = [x_k, x_{k+1})$. See Figure 2.7.

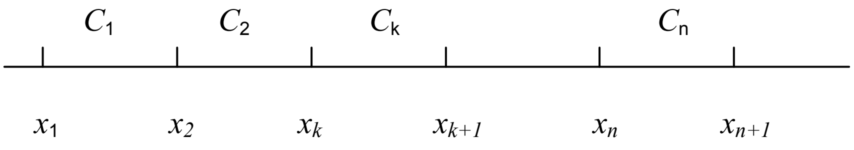


Figure 2.7 *Class intervals*

Suppose the frequency and relative frequency of class C_k is f_k and r_k $(k=1, 2, ..., n)$, respectively. Then the cumulative frequency that observations fall into classes $C_1, C_2, ..., C_k$ or lie below the value $xk+1$ is the sum $f_1+f_2+...+f_k$. The corresponding cumulative relative frequency *is* $r_1 +r_2+...+r_k$.

Example 2.9 Table 2.1 gives frequency, relative frequency, cumulative frequency and cumulative relative frequency distribution for quantity of glucose in blood of 100 students. According to this table the number of students having quantity of glucose less than 90 is 16.

A graph of cumulative frequency distribution is called an "less-than" **ogive** or simply ogive. Figure 2. shows the cumulative frequency distribution for quantity of glucose in blood of 100 students (data from Table 2.1)

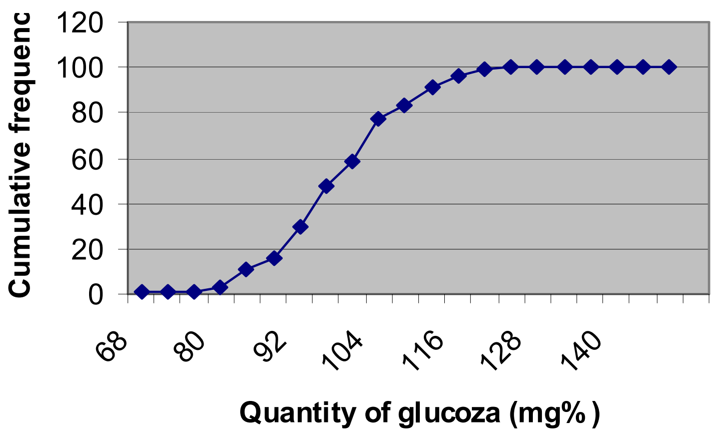


Figure 2.8 *Cumulative frequency distribution for quantity of glucose (for data in Table 2.1)*

2.9 Summary

This chapter discussed methods for presenting data set of qualitative and quantitative variables.

For a qualitative data set we first define categories and the category frequency which is the number of observations falling in each category. Further, the category relative frequency

and the percentage for a category are introduced. Bar graphs and pie charts as the graphical pictures of the data set are constructed.

If the data are quantitative and the number of the observations is small the categorization and the determination of class frequencies can be done by constructing a stem and leaf display. Large sets of data are best described using relative frequency distribution. The latter presents a table that organizes data into classes with their relative frequencies. For describing the quantitative data graphically histogram and polygon are used.

2.10 Exercises

1) A national cancer institure survey of 1,580 adult women recently responded to the question "In your opinion, what is the most serious health problem facing women?" The responses are summarized in the following table:

The most serious health problem for women	**Relative frequency**
Breast cancer	0.44
Other cancers	0.31
Emotional stress	0.07
High blood pressure	0.06
Heart trouble	0.03
Other problems	0.09

a) Use one of graphical methods to describe the data.

b) What proportion of the respondents believe that high blood pressure or heart trouble is the most serious health problem for women?

c) Estimate the percentage of all women who believe that some type of cancer is the most serious health problem for women?

2) The administrator of a hospital has ordered a study of the amount of time a patient must wait before being treated by emergency room personnel. The following data were collected during a typical day:

WAITING TIME (MINUTES)									
12	16	21	20	24	3	11	17	29	18
26	4	7	14	25	2	26	15	16	6

a) Arrange the data in an array from lowest to heighest. What comment can you make about patient waiting time from your data array?
b) Construct a frequency distribution using 6 classes. What additional interpretation can you give to the data from the frequency distribution?
c) Construct the cumulative relative frequency polygon and from this ogive state how long 75% of the patients should expect to wait.

3) Bacteria are the most important component of microbial eco systems in sewage treatment plants. Water management engineers must know the percentage of active bacteria at each stage of the sewage treatment. The accompanying data represent the percentages of respiring bacteria in 25 raw sewage samples collected from a sewage plant.

42.3	50.6	41.7	36.5	28.6
40.7	48.1	48.0	45.7	39.9
32.3	31.7	39.6	37.5	40.8
50.1	39.2	38.5	35.6	45.6
34.9	46.1	38.3	44.5	37.2

a. Construct a relative frequency distribution for the data.

b. Construct a stem and leaf display for the data.

c. Compare the two graphs of parts a and b.

4) At a newspaper office, the time required to set the entire front page in type was recorded for 50 days. The data, to the nearest tenth of a minute, are given below.

20.8	22.8	21.9	22.0	20.7	20.9	25.0	22.2	22.8	20.1
25.3	20.7	22.5	21.2	23.8	23.3	20.9	22.9	23.5	19.5
23.7	20.3	23.6	19.0	25.1	25.0	19.5	24.1	24.2	21.8
21.3	21.5	23.1	19.9	24.2	24.1	19.8	23.9	22.8	23.9
19.7	24.2	23.8	20.7	23.8	24.3	21.1	20.9	21.6	22.7

a) Arrange the data in an array from lowest to heighest.
b) Construct a frequency distribution and a “less-than” cumulative frequency distribution from the data, using intervals of 0.8 minutes.
c) Construct a frequency polygon from the data.
d) Construct a “less-than” ogive from the data.
e) From your ogive, estimate what percentage of the time the front page can be set in less than 24 minutes.

Chapter 3 Data characteristics: descriptive summary statistics

CONTENTS

3.1 Introduction

In the previous chapter data were collected and appropriately summarized into tables and charts. In this chapter a variety of descriptive summary measures will be developed. These descriptive measures are useful for analyzing and interpreting quantitative data, whether collected in raw form (ungrouped data) or summarized into frequency distributions (grouped data)

3.2 Types of numerical descriptive measures

Four types of characteristics which describe a data set pertaining to some numerical variable or phenomenon of interest are:

- Location
- Dispersion
- Relative standing
- Shape

In any analysis and/or interpretation of numerical data, a variety of descriptive measures representing the properties of location, variation, relative standing and shape may be used to extract and summarize the salient features of the data set.

If these descriptive measures are computed from a sample of data they are called **statistics** . In contrast, if these descriptive measures are computed from an entire population of data, they are called **parameters**.

Since statisticians usually take samples rather than use entire populations, our primary emphasis deals with statistics rather than parameters.

3.3 Measures of location (or measures of central tendency)

3.3.1. Mean

Definition 3.1

The arithmetic mean of a sample (or simply the sample mean) of n observations $x_1, x_2, \mathrm{K}\ , x_n$, denoted by $\bar{x}$ is computed as

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Definition 3.1a

The population mean is defined by the formula

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\text{Sum of the values of all observations in population}}{\text{Total number of observations in population}}$$

Note that the definitions of the population mean and the sample mean are the same. It is also valid for the definition of other measures of central tendency. But in the next section we will give different formulas for variances of population and sample.

Example 3.1 Consider 7 observations: 4.2, 4.3, 4.7, 4.8, 5.0, 5.1, 9.0.

By definition

$$\bar{x} = (4.2+ 4.3+ 4.7+ 4.8+ 5.0+ 5.1+ 9.0)/7 = 5.3$$

Advantages of the mean:

- It is a measure that can be calculated and is unique.
- It is useful for performing statistical procedures such as comparing the means from several data sets.

Disadvantages of the mean:

It is affected by extreme values that are not representative of the rest of the data.

Indeed, if in the above example we compute the mean of the first 6 numbers and exclude the 9.0 value, then the mean is 4.7. The one extreme value 9.0 distorts the value we get for the mean. It would be more representative to calculate the mean without including such an extreme value.

3.3.2. Median

Definition 3.2
The median m of a sample of n observations $x_1, x_2, \mathrm{K}\ , x_n$ arranged in ascending or descending order is the middle number that divides the data set into two equal halves: one half of the items lie above this point, and the other half lie below it.

Formula for calculating median of an arranged in ascending order data set

$$m = Median = \begin{cases} x_k \text{ if } n = 2k-1 \ (n \text{ is odd}) \\ \frac{1}{2}(x_k + x_{k+1}) \text{ if } n = 2k \ (n \text{ is even}) \end{cases}$$

Example 3.2 Find the median of the data set consisting of the observations 7, 4, 3, 5, 6, 8, 10.

Solution First, we arrange the data set in ascending order

3 4 5 6 7 8 10.

Since the number of observations is odd, $n = 2$ x 4 - 1, then median $m = x_4 = 6$. We see that a half of the observations, namely, 3, 4, 5 lie below the value 6 and another half of the observations, namely, 7, 8 and 10 lie above the value 6.

Example 3.3 Suppose we have an even number of the observations 7, 4, 3, 5, 6, 8, 10, 1. Find the median of this data set.

Solution First, we arrange the data set in ascending order

1 3 4 5 6 7 8 10.

Since the number of the observations n = 2 x 4, then by Definition

Median = $(x_4+x_5)/2$ = (5+6)/2 = 5.5

Advantage of the median over the mean: Extreme values in data set do not affect the median as strongly as they do the mean.

Indeed, if in Example 3.1 we have

mean = 5.3, median = 4.8.

The extreme value of 9.0 does not affect the median.

3.3.3 Mode

Definition 3.3
The mode of a data set $x_1, x_2, \mathrm{K}, x_n$ is the value of x that occurs with the greatest frequency , i.e., is repeated most often in the data set.

Example 3.4 Find the mode of the data set in Table 3.1.

Table 3.1 Quantity of glucose (mg%) in blood of 25 students

70	88	95	101	106
79	93	96	101	107
83	93	97	103	108
86	93	97	103	112
87	95	98	106	115

Solution First we arrange this data set in the ascending order

70	88	95	101	106
79	93	96	101	107
83	93	97	103	108
86	93	97	103	112
87	95	98	106	115

This data set contains 25 numbers. We see that, the value of 93 is repeated most often. Therefore, the mode of the data set is 93.

Multimodal distribution: A data set may have several modes. In this case it is called multimodal distribution.

Example 3.5 The data set

0	2	6	9
0	4	6	10
1	4	7	11
1	4	8	11
1	5	9	12

have two modes: 1 and 4. his distribution is called bimodal distribution.

Advantage of the mode: Like the median, the mode is not unduly affected by extreme values. Even if the high values are very high and the low value is very low, we choose the most frequent value of the data set to be the modal value We can use the mode no matter how large, how small, or how spread out the values in the data set happen to be.

Disadvantages of the mode:

- The mode is not used as often to measure central tendency as are the mean and the median. Too often, there is no modal value because the data set contains no values that occur more than once. Other times, every value is the mode because every value occurs for the same number of times. Clearly, the mode is a useless measure in these cases.
- When data sets contain two, three, or many modes, they are difficult to interpret and compare.

Comparing the Mean, Median and Mode

- In general, for data set 3 measures of central tendency: the mean , the median and the mode are different. For example, for the data set in Table 3.1, mean =96.48, median = 97 and mode = 93.
- If all observations in a data set are arranged symmetrically about an observation then this observation is the mean, the median and the mode.
- Which of these three measures of central tendency is better? The best measure of central tendency for a data set depends on the type of descriptive information you want. For most data sets encountered in business, engineering and computer science, this will be the MEAN.

3.3.4 Geometric mean

Definition 3.4
Suppose all the n observations in a data set $x_1, x_2, \ldots, x_n > 0$. Then the geometric mean of the data set is defined by the formula

$$\overline{x}_G = G.M = \sqrt[n]{x_1 x_2 \ldots x_n}$$

The geometric mean is appropriate to use whenever we need to measure the average rate of change (the growth rate) over a period of time.

From the above formula it follows

$$\log \overline{x}_G = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

where log is the logarithmic function of any base.

Thus, the logarithm of the geometric mean of the values of a data set is equal to the arithmetic mean of the logarithms of the values of the data set.

3.4 Measures of data variation

Just as measures of central tendency locate the “center” of a relative frequency distribution, measures of variation measure its “spread”.

The most commonly used measures of data variation are the range, the variance and the standard deviation.

3.4.1 Range

Definition 3.5
The range of a quantitative data set is the difference between the largest and smallest values in the set.

Range = Maximum - Minimum,

where Maximum = Largest value, Minimum = Smallest value.

3.4.2 Variance and standard deviation

Definition 3.6
The population variance of the population of the observations x is defined the formula

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)}{N}$$

where: σ^2 =population variance

x_i = the item or observation

μ = population mean

N = total number of observations in the population.

From the Definition 3.6 we see that the population variance is the average of the squared distances of the observations from the mean.

Definition 3.7

The standard deviation of a population is equal to the square root of the variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}}$$

Note that for the variance, the units are the squares of the units of the data. And for the standard deviation, the units are the same as those used in the data.

Definition 3.6a

The sample variance of the sample of the observations $x_1, x_2, \ldots, x_n$ is defined the formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

where: s^2 =sample variance

$\overline{x}$ = sample mean

n = total number of observations in the sample

The standard deviation of the sample is

$$s = \sqrt{s^2}$$

Remark: In the denominator of the formula for s^2 we use n-1 instead n because statisticians proved that if s^2 is defined as above then s^2 is an unbiased estimate of the variance of the population from which the sample was selected (i.e. the expected value of s^2 is equal to the population variance).

Uses of the standard deviation

The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located in relation to the mean. We can do this according to a theorem devised by the Russian mathematician P.L. Chebyshev (1821-1894).

Chebyshev’s Theorem

For any data set with the mean $\overline{x}$ and the standard deviation *s* at least **75%** of the values will fall within the interval $\overline{x} \pm 2s$ and at least **89%** of the values will fall within the interval $\overline{x} \pm 3s$ **.**

We can measure with even more precision the percentage of items that fall within specific ranges under a symmetrical, bell-shaped curve. In these cases we have:

The Empirical Rule

If a relative frequency distribution of sample data is *bell-shaped* with mean $\overline{x}$ and standard deviation *s*, then the proportions of the total number of observations falling within the intervals $\overline{x} \pm s$, $\overline{x} \pm 2s$, $\overline{x} \pm 3s$ are as follows:

$\overline{x} \pm s$: Close to 68%

$\overline{x} \pm 2s$: Close to 95%

$\overline{x} \pm 3s$: Near 100%

3.4.3 Relative dispersion: The coefficient of variation

The standard deviation is an absolute measure of dispersion that expresses variation in the same units as the original data. For example, the unit of standard deviation of the data set of height of a group of students is centimeter, the unit of standard deviation of the data set of their weight is kilogram. Can we compare the values of these standard deviations? Unfortunately, no, because they are in the different units.

We need a relative measure that will give us a feel for the magnitude of the deviation relative to the magnitude of the mean. The coefficient of variation is one such relative measure of dispersion.

Definition 3.8

The coefficient of variation of a data set is the relation of its standard deviation to its mean

cv = Coefficient of variation = $\dfrac{\text{Standard deviation}}{\text{Mean}} \times 100\,\%$

This definition is applied to both population and sample.

The unit of the coefficient of variation is percent.

Example 3.6 Suppose that each day laboratory technician A completes 40 analyses with a standard deviation of 5. Technician B completes 160 analyses per day with a standard deviation of 15. Which employee shows less variability?

At first glance, it appears that technician B has three times more variation in the output rate than technician A. But B completes analyses at a rate 4 times faster than A. Taking all this information into account, we compute the coefficient of variation for both technicians:

For technician A: cv=5/40 x 100% = 12.5%

For technician B: cv=15/60 x 100% = 9.4%.

So, we find that, technician B who has more absolute variation in output than technician A, has less relative variation.

3.5 Measures of relative standing

In some situations, you may want to describe the relative position of a particular observation in a data set.

Descriptive measures that locate the relative position of an observation in relation to the other observations are called measures of relative standing.

A measure that expresses this position in terms of a percentage is called a percentile for the data set.

Definition 3.9
Suppose a data set is arranged in ascending (or descending) order. The p^{th} percentile is a number such that p% of the observations of the data set fall below and (100-p)% of the observations fall above it.

The median, by definition, is the 50th percentile.
The 25th percentile, the median and 75th percentile are often used to describe a data set because they divide the data set into 4 groups, with each group containing one-fourth (25%) of the observations. They would also divide the relative frequency distribution for a data set into 4 parts, each contains the same are (0.25) , as shown in Figure 3.1. Consequently, the 25th percentile, the median, and the 75th percentile are called the lower quartile, the mid quartile, and the upper quartile, respectively, for a data set.

Definition 3.10
The lower quartile, Q_L, for a data set is the 25th percentile

Definition 3.11

The mid- quartile, M, for a data set is the 50th percentile.

Definition 3.12
The upper quartile, Q_U, for a data set is the 75th percentile.

Definition 3.13

The interquartile range of a data set is Q_U - Q_L .

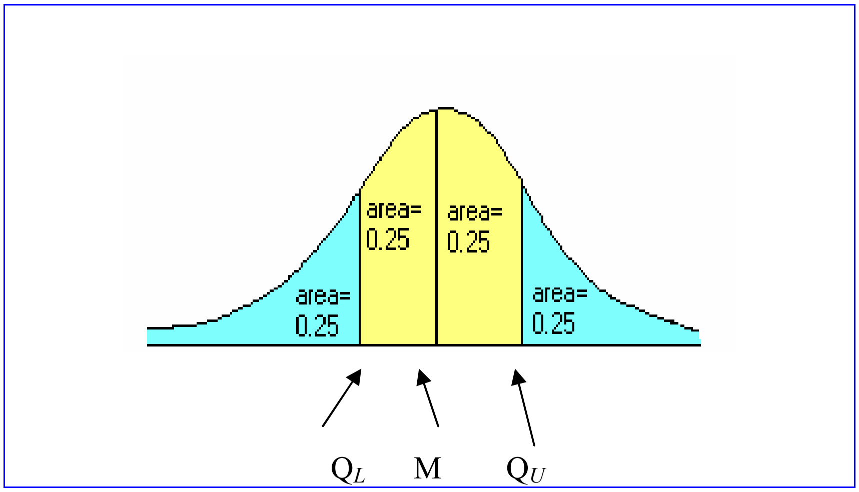


Figure 3.1 *Locating of lower, mid and upper quartiles*

For large data set, quartiles are found by locating the corresponding areas under the relative frequency distribution polygon as in Figure 3. . However, when the sample data set is small, it may be impossible to find an observation in the data set that exceeds, say, exactly 25% of the remaining observations. Consequently, the lower and the upper quartiles for small data set are not well defined. The following box describes a procedure for finding quartiles for small data sets.

Finding quartiles for small data sets:

1. Rank the n observations in the data set in ascending order of magnitude.

2. Calculate the quantity $(n+1)/4$ and round to the nearest integer. The observation with this rank represents the lower quartile. If $(n+1)/4$ falls halfway between two integers, round up.
3. Calculate the quantity $3(n+1)/4$ and round to the nearest integer. The observation with this rank represents the upper quartile. If $3(n+1)/4$ falls halfway between two integers, round down.

Example 3.7 Find the lower quartile, the median, and the upper quartile for the data set in Table 3.1.

Solution For this data set n = 25. Therefore, $(n+1)/4$ = 26/4 = 6.5, $3(n+1)/4$ = 3*26/4 = 19.5. We round 6.5 up to 7 and 19.5 down to 19. Hence, the lower quartile = 7th observation = 93, the upper quartile =19th observation = 103. We also have the median = 13th observation = 97. The location of these quartiles is presented in Figure 3.2.

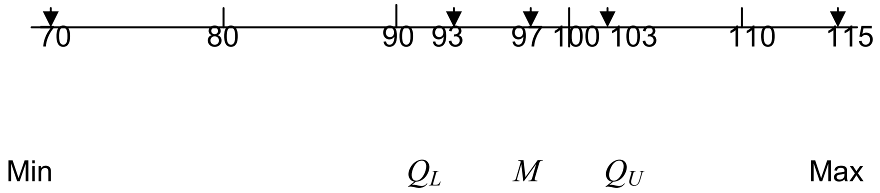


Figure 3.2 *Location of the quartiles for the data set of Table 2.1*

Another measure of real relative standing is the ***z*-score** for an observation (or standard score). It describes how far individual item in a distribution departs from the mean of the distribution. Standard score gives us the number of standard deviations, a particular observation lies below or above the mean.

Definition 3.14

Standard score (or z -score) is defined as follows:

For a population:

$$z\text{-score}= \frac{x-\mu}{\sigma}$$

where x = the observation from the population,

μ = the population mean,

σ = the population standard deviation .

For a sample:

$$z\text{-score}= \frac{x-\bar{x}}{s}$$

where x = the observation from the sample

$\bar{x}$ = the sample mean,

s = the sample standard deviation .

3.6 Shape

The fourth important numerical characteristic of a data set is its shape. In describing a numerical data set its is not only necessary to summarize the data by presenting appropriate measures of central tendency, dispersion and relative standing, it is also necessary to consider the shape of the data – the manner, in which the data are distributed.

There are two measures of the shape of a data set: skewness and kurtosis.

3.6.1 Skewness

If the distribution of the data is not symmetrical, it is called asymmetrical or skewed.

Skewness characterizes the degree of asymmetry of a distribution around its mean. For a sample data, the skewness is defined by the formula:

$$Skewness = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3,$$

where n = the number of observations in the sample,

x_i = ith observation in the sample,

s = standard deviation of the sample.

The direction of the skewness depends upon the location of the extreme values. If the extreme values are the larger observations, the mean will be the measure of location most greatly distorted toward the upward direction. Since the mean exceeds the median and the mode, such distribution is said to be positive or right-skewed. The tail of its distribution is extended to the right. This is depicted in Figure 3.3a.

On the other hand, if the extreme values are the smaller observations, the mean will be the measure of location most greatly reduced. Since the mean is exceeded by the median and the mode, such distribution is said to be negative or left-skewed. The tail of its distribution is extended to the left. This is depicted in Figure 3.3b.

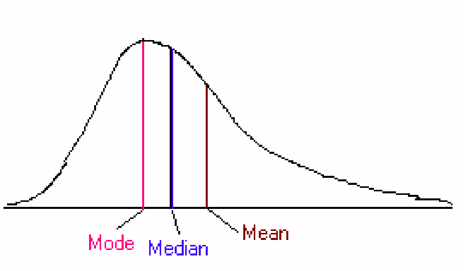


Figure 3.3a *Right-skewed distribution*

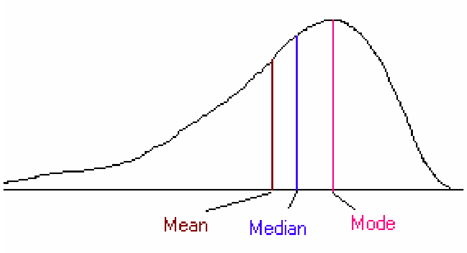


Figure 3.3b *Left-skewed distribution*

3.6.2 Kurtosis

Kurtosis characterizes the relative peakedness or flatness of a distribution compared with the bell-shaped distribution (normal distribution).

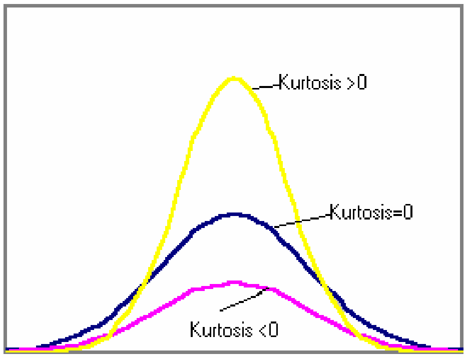
Kurtosis of a sample data set is calculated by the formula:

$$Kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.

The distributions with positive and negative kurtosis are depicted in Figure 3.4 , where the distribution with null kurtosis is normal distribution.

Figure 3.4
The distributions with positive and negative kurtosis



3.7 Methods for detecting outlier

Definition 3.15
An observation (or measurement) that is unusually large or small relative to the other values in a data set is called an **outlier**. Outliers typically are attributable to one of the following causes:

1. The measurement is observed, recorded, or entered into the computer incorrectly.
2. The measurements come from a different population.
3. The measurement is correct, but represents a rare event.

Outliers occur when the relative frequency distribution of the data set is extreme skewed, because such a distribution of the data set has a tendency to include extremely large or small observations.

There are two widely used methods for detecting outliers.

Method of using z-score:

According to Chebyshev theorem almost all the observations in a data set will have z-score less than 3 in absolute value i.e. fall into the interval $(\bar{x}-3s, \bar{x}+3s)$, where $\bar{x}$ the mean and s is is the standard deviation of the sample. Therefore, the observations with z-score greater than 3 will be outliers.

Example 3.8 The doctor of a school has measured the height of pupils in the class 5A. The result (in cm) is follows

Table 3.2 *Heights of the pupils of the class 5A*

130	132	138	136	131	153
131	133	129	133	110	132
129	134	135	132	135	134
133	132	130	131	134	135

For the data set in Table 3.1 $\bar{x}$ = 132.77, s = 6.06, 3s = 18.18, z-score of the observation of 153 is (153-132.77)/6.06=3.34 , z-score of 110 is (110-132.77)/6.06 = -3.76. Since the absolute values of z-score of 153 and 110 are more than 3, the height of 153 cm and the height of 110 cm are outliers in the data set.

Box plot method

Another procedure for detecting outliers is to construct a box plot of the data. Below we present steps to follow in constructing a box plot.

Steps to follow in constructing a box plot

1. Calculate the median M, lower and upper quartiles, Q_L and Q_U, and the interquartile range, $IQR= Q_U - Q_L$, for the data set.
2. Construct a box with Q_L and Q_U located at the lower corners. The base width will then be equal to IQR. Draw a vertical line inside the box to locate the median M.
3. Construct two sets of limits on the box plot: **Inner fences** are located a distance of 1.5 * IQR below Q_L and above $Q_{U;}$ **outer fences** are located a distance of 3 * IQR below Q_L and above Q_U (see Figure 4.5).
4. Observations that fall between the inner and outer fences are called **suspect outliers**. Locate the suspect outliers on the box plot using asterisks (*).Observations that fall outside the outer fences is called **highly suspect outliers**. Use small circles to locate them.

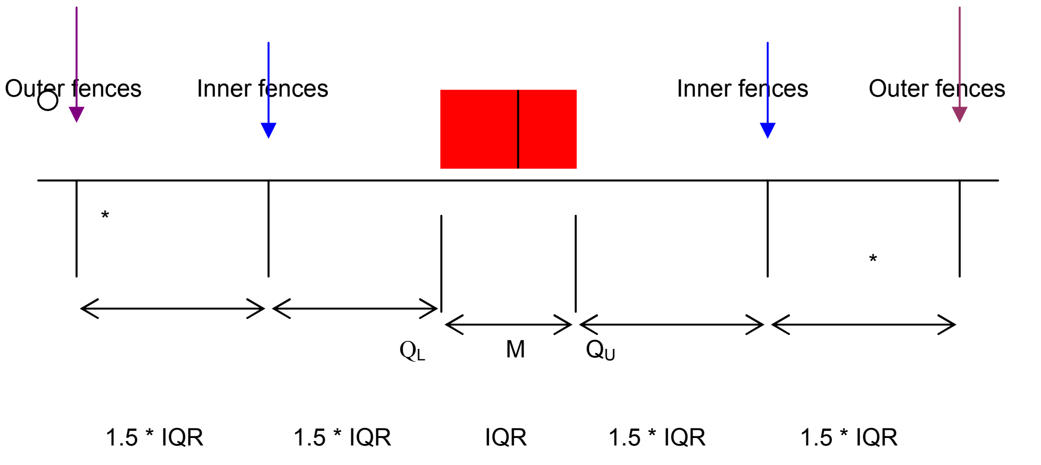


Figure 3.5 *Box plot*

For large data set box plot can be constructed using available statistical computer software.

A computer-generated by SPSS box plot for data set in Table 3.2 is shown in Figure 3.6.

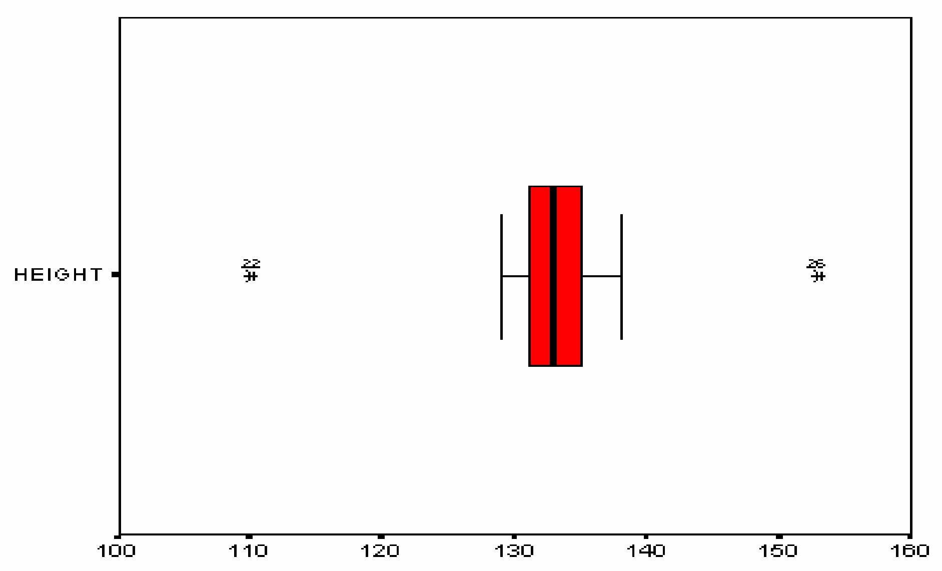


Figure 3.6 *Output from SPSS showing box plot for the data set in Table 3.2*

3.8 Calculating some statistics from grouped data

In Sections 3.3 through 3.6 we gave formulas for computing the mean, median, standard deviation etc. of a data set. However, these formulas apply only to **raw data sets,** i.e., those, in which the value of each of the individual observations in the data set is known. If the data have already been grouped into classes of equal width and arranged in a frequency table, you must use an alternative method to compute the mean, standard deviation etc.

Example 3.9 Suppose we have a frequency table of average monthly checking-account balances of 600 customers at a branch bank.

CLASS (DOLLARS)	FREQUENCY
0 – 49.99	78
50 – 99.99	123
100 – 149.99	187
150 – 199.99	82
150 – 199.99	82
200 – 249.99	51
250 – 299.99	47
300 – 349.99	13
350 – 399.99	9
400 – 449.99	6
450 – 499.99	4

From the information in this table, we can easily compute an estimate of the value of the mean and the standard deviation.

Formulas for calculating the mean and the standard deviation for grouped data:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}, \qquad s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \left(\sum_{i=1}^{k} f_i x_i\right)^2}{n-1},$$

where $\bar{x}$ = mean of the data set, s^2 = standard deviation of the data set

x_i = midpoint of the ith class, f_i = frequency of the ith class,

k = number of classes, n = total number of observations in the data set.

3.9 Computing descriptive summary statistics using computer softwares

All statistical computer softwares have procedure for computing descriptive summary statistics. Below we present outputs from STATGRAPHICS and SPSS for computing descriptive summary statistics for GLUCOSE data in Table 2.0b.

```
Variable:              GLUCOSE.GLUCOSE
---------------------------------------------------------
-------------
Sample size            100.
Average                100.
Median                 100.5
Mode                   106.
Geometric mean          99.482475
Variance               102.767677
Standard deviation      10.137439
Standard error           1.013744
Minimum                 70.
Maximum                126.
Range                   56.
Lower quartile          94.
Upper quartile         106.
Interquartile range     12.
Skewness                -0.051526
Kurtosis                 0.131118
Coeff. of variation     10.137439
```

Figure 4.7 Output from STATGRAPHICS for Glucose data

3.10 Summary

Numerical descriptive measures enable us to construct a mental image of the relative frequency distribution for a data set pertaining to a numerical variable. There are 4 types of these measures: location, dispersion, relative standing and shape.

Three numerical descriptive measures are used to locate a relative frequency distribution are the mean, the median, and the mode. Each conveys a special piece of information. In a sense, the mean is the balancing point for the data. The median, which is insensitive to extreme values, divides the data set into two equal halves: half of the observations will be less than the median and half will be larger. The mode is the observation that occurs with greatest frequency. It is the value of the data set that locates the point where the relative frequency distribution achieves its maximum relative frequency.

The range and the standard deviation measure the spread of a relative frequency distribution. Particularly, we can obtain a very good notion of the way data are distributed around the mean by constructing the intervals and referring to the Chebyshev's theorem and the Empirical rule.

Percentiles, quartiles, and z-scores measure the relative position of an observation in a data set. The lower and upper quartiles and the distance between them called the inter-quartile range can also help us visualize a data set. Box plots constructed from intervals based on the inter-quartile range and z-scores provide an easy way to detect possible outliers in the data.

The two numerical measures of the shape of a data set are skewness and kurtosis. The skewness characterizes the degree of asymmetry of a distribution around its mean. The kurtosis characterizes the relative peakedness or flatness of a distribution compared with the bell-shaped distribution.

3.11 Exercises

1. The ages of a sample of the people attending a training course on networking in IOIT in Hanoi are:

29	20	23	22	30	32	28
23	24	27	28	31	32	33
31	28	26	25	24	23	22
26	28	31	25	28	27	34

a) Construct a frequency distribution with intervals 15-19, 20-24, 25-29, 30-34, 35-39.

b) Compute the mean and the standard deviation of the raw data set.

c) Compute the approximate values for the mean and the standard deviation using the constructed frequency distribution table. Compare these values with ones obtained in b).

2. Industrial engineers periodically conduct “work measurement” analyses to determine the time used to produce a single unit of output. At a large processing plant, the total number of man-hours required per day to perform a certain task was recorded for 50 days. his information will be used in a work measurement analysis. The total man-hours required for each of the 50 days are listed below.

128	119	95	97	124	128	142	98	108	120
113	109	124	97	138	133	136	120	112	146
128	103	135	114	109	100	111	131	113	132
124	131	133	88	118	116	98	112	138	100
112	111	150	117	122	97	116	92	122	125

a) Compute the mean, the median, and the mode of the data set.

b) Find the range, the variance and the standard deviation of the data set.

c) Construct the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. Count the number of observations that fall within each interval and find the corresponding proportions. Compare the results to the Chebyshev theorem. Do you detect any outliers?

e) Find the 75th percentile for the data on total daily man-hours.

3. An engineer tested nine samples of each of three designs of a certain bearing for a new electrical winch. The following data are the number of hours it took for each bearing to fail when

the winch motor was run continuously at maximum output, with a load on the winch equivalent to 1,9 times the intended capacity.

DESIGN		
A	**B**	**C**
16	18	21
16	27	17
53	34	23
21	34	32
17	32	21
25	19	18
30	34	21
21	17	28
45	43	19

a) Calculate the mean and the median for each group.

b) Calculate the standard deviation for each group.

c) Which design is best and why?

4. The projected 30-day storage charges (in US$) for 35 web pages stored on the web server of a university are listed here:

120	125	145	180	175	167	154
143	120	180	175	190	200	145
165	210	120	187	179	167	165
134	167	189	182	145	178	231
185	200	231	240	230	180	154

a) Construct a stem-and-leaf display for the data set.

b) Compute $\overline{x}$, s^2 and s.

c) Calculate the intervals $\overline{x} \pm s$, $\overline{x} \pm 2s$, and $\overline{x} \pm 3s$ and count the number of observations that fall within each interval. Compare your results with the Empirical rule.

Chapter 4 Probability: Basic concepts

CONTENTS

4.1 Experiment, Events and Probability of an Event

Definition 4.1
The process of making an observation or recording a measurement under a given set of conditions is a **trial** or **experiment**

Thus, an experiment is realized whenever the set of conditions is realized.

Definition 4.2
Outcomes of an experiment are called **events.**

We denote events by capital letters $A, B, C,...$

Example 4.1 Consider the following experiment. Toss a coin and observe whether the upside of the coin is Head or Tail. Two events may be occurred:

- H: Head is observed,
- T: Tail is observed.

Example 4.2 Toss a die and observe the number of dots on its upper face. You may observe one, or two, or three, or four, or five or six dots on the upper face of the die. You can not predict this number.

Example 4.3 When you draw one card from a standard 52 card bridge deck, some possible outcomes of this experiment can not be predicted with certainty in advance are:

- A: You draw an ace of hearts
- B: You draw an eight of diamonds

- C: You draw a spade
- D: You do not draw a spade.

The **probability** of an event A, denoted by $P(A)$, in general, is the chance A will happen.

But how to measure the chance of occurrence, i.e., how determine the probability an event?

The answer to this question will be given in the next Sections.

4.2 Approaches to probability

The number of different definitions of probability that have been proposed by various authors is very large. But the majority of definitions can be subdivided into 3 groups:

1. Definitions of probability as a quantitative measure of the "degree of certainty" of the observer of experiment.
2. Definitions that reduce the concept of probability to the more primitive notion of "equal likelihood" (the so-called "classical definition ").
3. Definitions that take as their point of departure the "relative frequency" of occurrence of the event in a large number of trials ("statistical" definition).

According to the first approach to definition of probability, the theory of probability is something not unlike a branch of psychology and all conclusions on probabilistic judgements are deprived of the objective meaning that they have independent of the observer. Those probabilities that depend upon the observer are called subjective probabilities.

In the next sections we shall give the classical and statistical definitions of probability.

4.3 The field of events

Before proceeding to the classical Definition of the concept of probability we shall introduce some **definitions and relations between the events**, which may or may not occur when an experiment is realized.

1. If whenever the event A occurs the event B also occurs, then we say that A **implies** B (or A is contained in B) and write $A \subset B$ or $B \supset A$.
2. If A implies B and at the same time, B implies A, i.e., if for every realization of the experiment either A and B both occur or both do not occur, then we say that the events A and B are **equivalent** and write A=B.
3. The event consisting in the simultaneous occurrence of A and B is called the **product or intersection** of the events A and B, and will be denoted by AB or $A \cap B$.
4. The event consisting in the occurrence of at least one of the events A or B is called the **sum, or union**, of the events A and B, and is denoted by A+B or $A \cup B$.
5. The event consisting in the occurrence of A and the non-occurrence of B is called the **difference** of the events A and B and is denoted by A-B or $A \backslash B$.
6. An event is called **certain (or sure)** if it must inevitably occur whenever the experiment is realized.
7. An event is called **impossible** if it can never occur.

Clearly, all certain events are equivalent to one another. We shall denote these events by the letter *E*. All impossible events are likewise equivalent and denoted by **0**.

8. Two events A and $\overline{A}$ are **complementary** if $A+\overline{A}=E$ and $A\overline{A}$ **= 0** hold simultaneously.

For example, in the experiment of tossing a die the following events are complementary:

- D_{even} = {even number of dots is observed on upper face}
- D_{odd} ={ odd number of dots is observed on upper face}

9. Two events A and B are called **mutually** *exclusive* if when one of the two events occurs in the experiment, the other can not occur, i.e., if their joint occurrence is impossible AB = **0**.

10. If $A=B_1+B_2+...+B_n$ and the events B_i (i =1,2,...,n) are mutually exclusive in pairs (or pair wise mutually exclusive), i.e., B_iB_j = **0** for any $i \neq j$, then we say that the event A is **decomposed into the mutually exclusive events** B_1, B_2, ..., B_n.

For example, in the experiment of tossing a single die, the event consisting of the throw of an even number of dots is decomposed into the mutually exclusive events D_2, D_4 and D_6, where D_k = {observing k dots on the upper face of the die}.

11. An event A is called **simple (or elementary)** if it can not be decomposed into other events. For example, the events D_k that k dots (k=1, 2, 3, 4, 5, 6) are observed in the experiment of tossing a die are simple events.

12. The **sample space** of an experiment is the collection of all its simple events.

13. **Complete list of events**: Suppose that when the experiment is realized there may be a list of events A_1, A_2, ..., A_n with the following properties:

    I. A_1, A_2, ..., A_n are pair wise mutually exclusive events,

    II. $A_1+A_2+...+A_n=E$**.**

Then we say that the list of events A_1, A_2, ..., A_n is complete.

The examples of complete list of events may be:

- List of events Head and Tail in tossing a coin
- List of events D_1, D_2, D_3, D_4, D_5, D_6 in the experiment of tossing a die.
- List of events D_{even} and D_{odd} in the experiment of tossing a die.

All relations between events may be interpreted geometrically by **Venn diagrams**. In theses diagrams the entire sample space is represented by a rectangle and events are represented by parts of the rectangle. If two events are mutually exclusive, their parts of the rectangle will not overlap each other as shown in Figure 4.1a. If two events are not mutually exclusive, their parts of the rectangle will overlap as shown in Figure 4.1b.

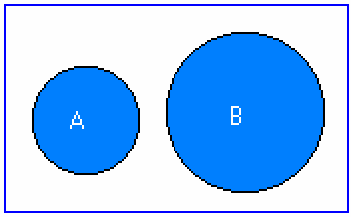


Figure 4.1a *Two mutually exclusive events*

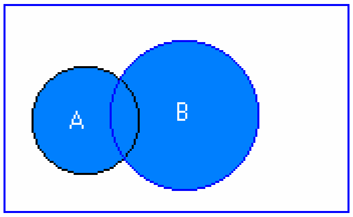
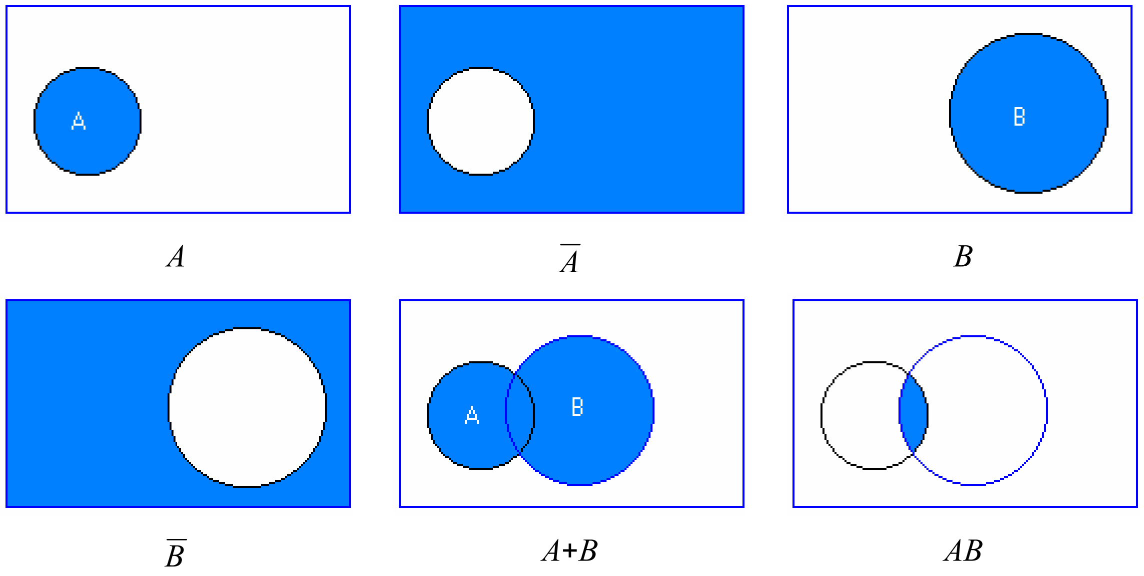


Figure 4.1b *Two non-mutually exclusive events*



A $\overline{A}$ B

$\overline{B}$ $A+B$ AB

Figure 4.2 *Events* $A, \overline{A}, B, \overline{B}$ *and AB*

In every problem in the theory of probability one has to deal with an experiment (under some specific set of conditions) and some specific family of events S.

Definition 4.3

A family $\boldsymbol{S}$ of events is called a **field of events** if it satisfies the following properties:

1. If the event A and B belong to the family $\boldsymbol{S}$, the so do the events AB, $A+B$ and A-B.
2. The family $\boldsymbol{S}$ contains the certain event E and the impossible event **0** .

We see that the sample space of an experiment together with all the events generated from the events of this space by operations "sum", "product" and "complement" constitute a field of events. Thus, for every experiment we have a field of events.

4.4 Definitions of probability

4.4.1 The classical definition of probability

The classical definition of probability reduces the concept of probability to the concept of equiprobability (equal likelihood) of events, which is regarded as a primitive concept and hence not subject to formal definition. For example, in the tossing of a single perfectly cubical die, made of completely homogeneous material, the equally likely events are the appearance of any of the specific number of dots (from 1 to 6) on its upper face.

Thus, for the classical definition of probability we suppose that all possible simple events are equally likely.

Definition 4.4 (The classical definition of probability)

The **probability** $P(A)$ of an event A is equal to the number of possible simple events (outcomes) favorable to A divided by the total number of possible simple events of the experiment, i.e.,

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

where *m*= number of the simple events into which the event A can be decomposed.

Example 4.4 Consider again the experiment of tossing a balanced coin (see Example 4.1). In this experiment the sample space consists of two simple events: H (Head is observed) and T (Tail is observed). These events are equally likely. Therefore, $P(H)$=$P(T)$=1/2.

Example 4.5 Consider again the experiment of tossing a balanced die (see Example 4.2). In this experiment the sample space consists of 6 simple events: D_1, D_2, D_3, D_4, D_5, D_6, where D_k is the event that k dots (k=1, 2, 3, 4, 5, 6) are observed on the upper face of the die. These events are equally likely. Therefore, $P(D_k)$ =1/6 (k=1, 2, 3, 4, 5, 6).

Since $D_{odd} = D_1+D_3+D_5$, $D_{even} = D_2+D_4+D_6$, where D_{odd} is the event that an odd number of dots are observed, D_{even} an even number of dots are observed, we have $P(D_{odd})$=3/6=1/2, $P(D_{even})$ = 3/6 = 1/2. If denote by A the event that a number less than 6 of dots is observed then $P(A)$ = 5/6 because the event $A = D_1+ D_2+D_3+ D_4+ D_5$.

According to the above definition, every event belonging to the field of events $\boldsymbol{S}$ has a well-defined probability. Therefore, the probability $P(A)$ may be regarded as a function of the event A defined over the field of events $\boldsymbol{S}$**.** This function has the following properties, which are easily proved.

The properties of probability:

1. For every event A of the field $\boldsymbol{S}$, $P(A) \geq 0$
2. For the certain event ***E*****,** $P(\boldsymbol{E}) = 1$
3. If the event A is decomposed into the mutually exclusive events B and C belonging to $\boldsymbol{S}$ then $P(A)$=$P(B)$+$P(C)$

This property is called the **theorem on the addition of probabilities**.

4. The probability of the event $\overline{A}$ complementary to the event A is given by the formula $P(\overline{A}) = 1 - P(A)$.
5. The probability of the impossible event is zero, $P(\mathbf{0}) = 0$.
6. If the event A implies the event B then $P(A) \leq P(B)$.
7. The probability of any event A lies between 0 and 1: $0 \leq P(\text{A}) \leq 1$.

Example 4.6 Consider the experiment of tossing two fair coins. Find the probability of the event A = {observe at least one Head} by using the complement relationship.

Solution The experiment of tossing two fair coins has 4 simple events: HH, HT, TH and TT, where H = {Head is observed}, T = {Tail is observed}. We see that the event A consists of the simple events HH, HT, TH. Then the complementary event for A is $\overline{A}$ = { No Heads observed } = TT. We have $P(\overline{A}) = P(TT)$ = 1/4. Therefore, $P(A)$ = 1-$P(\overline{A})$ = 1-1/4 = 3/4.

4.4.2 The statistical definition of probability

The classical definition of probability encounters insurmountable difficulties of a fundamental nature in passing from the simplest examples to a consideration of complex problems. First off all, the question arises in a majority of cases, as to a reasonable way of selecting the "equally likely cases". Thus, for examples, it is difficult to determine the probability that tomorrow the weather will be good, or the probability that a baby to be born is a boy, or to answer to the question "what are the chances that I will blow one of my stereo speakers if I turn my amplifier up to wide open?"

Lengthy observations as to the occurrence or non-occurrence of an event A in large number of repeated trials under the same set of conditions show that for a wide class of phenomena, the number of occurrences or non-occurrences of the event A is subject to a stable law. Namely, *if we denote by m the number of times the event A occurs in N independent trials, then it turns out that for sufficiently large N the ratio m/N in most of such series of observations, assumes an almost constant value. Since this constant is an objective numerical characteristic of the phenomena, it is natural to call it the statistical probability of the random event A under investigation*.

Definition 4.5 (The statistical definition of probability)

The probability of an event A can be approximated by the proportion of times that A occurs when the experiment is repeated a very large number of times.

Fortunately, for the events to which the classical definition of probability is applicable, the statistical probability is equal to the probability in the sense of the classical definition.

4.4.3 Axiomatic construction of the theory of probability (optional)

The classical and statistical definitions of probability reveal some restrictions and shortcomings when deal with complex natural phenomena and especially, they may lead to paradoxical conclusions, for example, the well-known Bertrand's paradox. Therefore, in order to find wide applications of the theory of probability, mathematicians have constructed a rigorous foundation of this theory. The first work they have done is the *axiomatic definition of probability* that *includes as special cases both the classical and statistical definitions of probability and overcomes the shortcomings of each.*

Below we formulate the axioms that define probability.

Axioms for probability

1. With each random event A in a field of events $\boldsymbol{S}$**,** there is associated a non-negative number $P(A)$,called its probability.
2. The probability of the certain event **E** is 1, i.e., $P(\mathbf{E}) = 1$.
3. (Addition axiom) If the event A_1, A_2, ..., A_n are pair wise mutually exclusive events then
$$P(A_1+ A_2+ ...+A_n) = P(A_1)+P(A_2)+ ...+P(A_n)$$
4. (Extended axiom of addition) If the event A is equivalent to the occurrence of at least one of the pair wise mutually exclusive events A_1, A_2, ..., A_n,...then
$$P(A) = P(A_1)+P(A_2)+ ...+P(A_n)+...$$

Obviously, the classical and statistical definitions of probability which deal with finite sum of events, satisfy the formulated above axioms. The necessity for introducing the extended axiom of addition is motivated by the fact that in probability theory we constantly have to consider events that decompose into an infinite number of sub-events.

4.5 Conditional probability and independence

We have said that a certain set of conditions Ç underlies the definition of the probability of an event. If no restrictions other than the conditions Ç are imposed when calculating the probability $P(A)$, then this probability is called unconditional.

However, in many cases, one has to determine the probability of an event under the condition that an other event B whose probability is greater than 0 has already occurred.

Definition 4.6

The probability of an event A, given that an event B has occurred, is called the **conditional probability of A given B** and denoted by the symbol $P(A|B)$.

Example 4.7 Consider the experiment of tossing a fair die. Denote by A and B the following events:

A = {Observing an even number of dots on the upper face of the die},

B = {Observing a number of dots less than or equal to 3 on the upper face of the die}.

Find the probability of the event A, given the event B.

Solution We know that the sample space of the experiment of tossing a fair die consists of 6 simple events: D_1, D_2, D_3, D_4, D_5, D_6, where D_k is the event that k dots (k = 1, 2, 3, 4, 5, 6) are observed on the upper face of the die. These events are equally likely, and $P(D_k) = 1/6$ (k = 1, 2, 3, 4, 5, 6). Since $A = D_2+ D_4+ D_6$, $B = D_1+ D_2+D_3$ we have $P(A) = P(D_2)+ P(D_4)+ P(D_6)$ = 3*1/6 = 1/2, $P(B) = P(D_1)+ P(D_2)+ P(D_3)$ = 3*1/6 = 1/2.

If the event B has occurred then it reduces the sample space of the experiment from 6 simple events to 3 simple events (namely those D_1, D_2, D_3 contained in event B). Since the only even number of three numbers 1, 2, 3 is 2 there is only one simple event D_2 of reduced sample space that is contained in the event A. Therefore, we conclude that the probability that A occurs given that B has occurred is one in three, or 1/3, i.e., $P(A|B) = 1/3$.

For the above example it is easy to verify that $P(A|B) = \dfrac{P(AB)}{P(B)}$. In the general case, we use this formula to define the conditional probability.

Formula for conditional probability

If the probability of an event B is greater 0 then the conditional probability of an event A, given that the event B has occurred, is calculated by the formula

$$P(A|B) = \frac{P(AB)}{P(B)}, \qquad (1)$$

where AB is the event that both A and B occur.

In the same way, if $P(A)>0$, the conditional probability of an event B, given that the event A has occurred, is defined by the formula

$$. \qquad P(B|A) = \frac{P(AB)}{P(A)} \qquad (1')$$

Each of formulas (1) and (1') is equivalent to the so-called Multiplication Theorem.

Multiplication Theorem

The probability of the product of two events is equal to the product of the probability of one of the events by the conditional probability of the other event, given that the first even has occurred, namely

$$P(AB) = P(A)\ P(B|A) = P(B)\ P(A|B).$$

The Multiplication Theorem is also applicable if one of the events A and B is impossible since, in this case, one of the equalities $P(A|B) = 0$ and $P(AB) = 0$ holds along with $P(A) = 0$.

Definition 4.7
We say that an event A is **independent** of an event B if

$$P(A|B) = P(A),$$

i.e., the occurrence of the event B does not affect the probability of the event A.

If the event A is independent of the event B, then it follows from (2) that $P(A)\ P(B|A) = P(B)\ P(A)$. From this we find $P(B|A) = P(B)$ if $P(A)>0$, i.e., the event B is also independent of A. Thus, **independence is a symmetrical relation**.

Example 4.8 Consider the experiment of tossing a fair die and define the following events:

A = {Observe an even number of dots}

B = { Observe a number of dots less or equal to 4}.

Are events A and B independent?

Solution As in Example 4.7 we have $P(A) = 1/2$ and $P(B) = P(D_1)+ P(D_2)+ P(D_3)+P(D_4) = 4*1/6 = 2/3$, where D_k is the event that k dots (k = 1, 2, 3, 4, 5, 6) are observed on the upper face of the die. Since $AB = D_2 + D_4$, we have $P(AB) = P(D_2)+ P(D_4) = 1/6+1/6 = 1/3$.

Now assuming B has occurred, the probability of A given B is

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{1/3}{2/3} = \frac{1}{2} = P(A)\,.$$

Thus, assuming B has occurred does not alter the probability of A. Therefore, the events A and B are independent.

The concept of independence of events plays an important role in the theory of probability and its applications. In particular, the greater part of the results presented in this course is obtained on the assumption that the various events considered are independent.

In practical problems, we rarely resort to verifying that relations $P(A|B) = P(A)$ or $P(B|A) = P(B)$ are satisfied in order to determine whether or not the given events are independent. To determine independence, we usually make use of intuitive arguments based on experience.

The Multiplication Theorem in the case of independent events takes on a simple form.

Multiplication Theorem for independent events
If the events A and B are independent then

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

We next generalize the notion of the independence of two events to that of a collection of events.

Definition 4.8
The events B_1, B_2, ..., B_n are called collectively independent or mutually independent if for any event B_p (p = 1, 2,..., n) and for any group of other events B_q, B_r, ...,B_s of this collection, the event B_p and the event $B_qB_r...B_s$ are independent.

Note that for several events to be mutually independent, it is not sufficient that they be pair wise independent.

4.6 Rules for calculating probability

4.6.1 The addition rule

From the classical definition of probability we deduced the addition theorem, which serves as the addition axiom for the axiomatic definition of probability. Using this axiom we get the following rule:

Addition rule
If the event A_1, A_2, ..., A_n are pair wise mutually exclusive events then

$$P(A_1+ A_2+ ...+A_n) = P(A_1)+P(A_2)+ ...+P(A_n)$$

In the case of two non-mutually exclusive events A and B we have the formula

$$P(A+B) = P(A) + P(B) – P(AB).$$

Example 4.9 In a box there are 10 red balls, 20 blue balls, 10 yellow balls and 10 white balls. At random draw one ball from the box. Find the probability that this ball is color.

Solution Call the event that the ball drawn is red to be R, is blue B, is yellow Y, is white W and is color C. Then $P(R)$ = 10/(10+20+10+10) = 10/50 = 1/5, $P(B)$ = 20/50 = 2/5, $P(Y)$ = 10/50 = 1/5. Since C = R+B+Y and the events R, B and Y are mutually exclusive , we have $P(C)$ = $P(R+B+Y)$ = 1/5+2/5+1/5 = 4/5.

In the preceding section we also got the multiplicative theorem. Below for the purpose of computing probability we recall it.

Multiplicative rule
For any two events A and B from the same field of events there holds the formula

$$P(AB) = P(A)\,P(B|A) = P(B)\,P(A|B).$$

If these events are independent then

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

Now suppose that the event B may occur together with one and only one of n mutually exclusive events $A_1, A_2, ..., A_n$, that is

$$B = A_1B + A_2B + ...+A_nB.$$

By Addition rule we have

$$P(B)= P(A_1B)+P(A_2B)+ ...+P(A_nB).$$

Further, by Multiplicative rule we get a formula, called the formula of total probability.

Formula of total probability
If the event B may occur together with one and only one of n mutually exclusive events $A_1, A_2, ..., A_n$ then

$$P(B)= P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+ ...+P(A_n)P(B|A_n).$$

Example 4.10 There are 5 boxes of lamps:

3 boxes with the content A_1: 9 good lamps and 1 defective lamp,

2 boxes with the content A_2: 4 good lamps and 2 defective lamp.

At random select one box and from this box draw one lamp. Find the probability that the drawn lamp is defective.

Solution Denote by B the event that the drawn lamp is defective and by the same A_1, A_2 the events that the box with content A_1, A_2, respectively, is selected. Since the defective lamp may be drawn from a box of either content A_1 or content A_2 we have $B = A_1B + A_2B$. By the formula of total probability $P(B) = P(A_1)P(B|A_1)+P(A_2)P(B|A_2)$.

Since $P(A_1) = 3/5$, $P(A_2) = 2/5$, $P(B|A_1) = 1/10$, $P(B|A_2) = 2/6 = 1/3$ we have

$$P(B) = 3/5 * 1/10 + 2/5 *1/3 = 29/150 = 0.19.$$

Thus, the probability that the drawn lamp is defective is 0.19.

Now, under the same assumptions and notations as in the formula of total probability, find the probability of the event A_k, given that the event B has occurred.

*A*ccording to the Multiplicative rule,

$$P(A_kB) = P(B)P(A_k|B) = P(A_k)\, P(B|A_k)$$

Hence,

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{P(B)}$$

using the formula of total probability, we then find the following

Bayes's Formula

If the event B may occur together with one and only one of n mutually exclusive events A_1, A_2, ..., A_n then

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)}$$

The formula of Bayes is sometimes called the formula for probabilities of hypotheses.

Example 4.11 As in Example 4.10, there are 5 boxes of lamps:

3 boxes with the content A_1: 9 good lamps and 1 defective lamp,

2 boxes with the content A_2: 4 good lamps and 2 defective lamp.

From one of the boxes, chosen at random, a lamp is withdrawn. It turns out to be a defective (event B). What is the probability, after the experiment has been performed (the aposteriori probability), that the lamp was taken from an box of content A_1?

Solution We have calculated $P(A_1) = 3/5$, $P(A_2) = 2/5$, $P(B|A_1) = 1/10$, $P(B|A_2) = 2/6 = 1/3$, $P(B) = 29/150$. Hence, the formula of Bayes gives

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{3/5*1/10}{29/150} = \frac{9}{29} \approx 0.31\,.$$

Thus, the probability that the lamp was taken from an box of content A_1, given the experiment has been performed, is equal 0.31.

4.7 Summary

In this chapter we introduced the notion of experiment whose outcomes called the events could not be predicted with certainty in advance. The uncertainty associated with these events was measured by their probabilities. But what is the probability? For answer to this question we briefly discussed approaches to probability and gave the classical and statistical definitions of probability. The classical definition of probability reduces the concept of probability to the concept of equiprobability of simple events. According to the classical definition, the probability of an event A is equal to the number of possible simple events favorable to A divided by the total number of possible events of the experiment. In the time, by the statistical definition the probability of an event is approximated by the proportion of times that A occurs when the experiment is repeated very large number of times.

4.8 Exercises

A, B, C are random events.

1) Explain the meaning of the relations:
   a) ABC = A;
   b) $A + B + C = A$.
2) Simplify the expressions
   a) $(A+B)(B+C)$;
   b) $(A+B)(A+\overline{B})$;
   c) $(A+B)(A+\overline{B})(\overline{A}+B)$.
3) A four-volume work is placed on a shelf in random order. What is the probability that the books are in proper order from right to left or left to right?
4) In a lot consisting of *N* items, *M* are defective , n items are selected at random from the lot (n<*N*). Find the probability that $m\ (m \le N)$ of them will be prove to be defective.
5) A quality control inspector examines the articles in a lot consisting of m items of first grade and n items of second grade. A check of the first b articles chosen at random from the lot has shown that all of them are of second grade $(b<m)$. Find the probability that of the next two items selected at random from those remaining at least one proves to be second grade.
6) From a box containing m white balls and n black balls (m>n), one ball after another is drawn at random. What is the probability that at some point the number of white balls and black balls drawn will be the same?
7) Two newly designed data base management systems (DBMS), A and B, are being considered for marketing by a large computer software vendor. To determine whether DBMS users have a preference for one of the two systems, four of the vendor's customers are randomly selected and given the opportunity to evaluate the performances of each of the two systems. After sufficient testing, each user is asked to state which DBMS gave the better performance (measured in terms of CPU utilization, execution time, and disk access).
   a) Count the possible outcomes for this marketing experiment.
   b) If DBMS users actually have no preference for one system over the other (i.e., performances of the two systems are identical), what is the probability that all four sampled users prefer system A?
   c) If all four customers express their preference for system A, can the software vendor infer that DBMS users in general have a preference for one of the two systems?

Chapter 5 Basic Probability distributions

CONTENTS

5.1 Random variables

One of the fundamental concepts of probability theory is that of a random variable.

Definition 5.1
A random variable is a variable that assumes numerical values associated with events of an experiment.

Example 5.1 Observe 100 babies to be born in a clinic. The number of boys, which have been born, is a random variable. It may take values from 0 to 100.

Example 5.2 Number of patients of a clinic daily is a random variable.

Example 5.3 Select one student from an university and measure his/her height and record this height by x. Then x is a random variable, assuming values from, say from 100 cm to 250 cm in dependence upon each specific student.

Example 5.4 The weight of babies at birth also is a random variable. It can assume values in the interval, for example, from 800 grams to 6000 grams.

Classification of random variables: Random variables may be divided into two types: discrete random variables and continuous random variables.

Definition 5.2

A **discrete random variable** is one that can assume only a countable number of values.

A **continuous random variable** can assume any value in one or more intervals on a line.

Among the random variables described above the number of boys in Example 5.1 and the number of patients in Example 5.2 are discrete random variables, the height of students and the weight of babies are continuous random variables.

Example 5.5 Suppose you randomly select a student attending your university. Classify each of the following random variables as discrete or continuous:

a) Number of credit hours taken by the student this semester
b) Current grade point average of the student.

Solution a) The number of credit hours taken by the student this semester is a discrete random variable because it can assume only a countable number of values (for example 10, 11, 12, and so on). It is not continuous since the number of credit hours can not assume values as 11.5678, 15.3456 and 12.9876 hours.

b) The grade point average for the student is a continuous random variable because it could theoretically assume any value (for example, 5.455, 8.986) corresponding to the points on the interval from 0 to 10 of a line.

5.2 The probability distribution for a discrete random variable

Definition 5.3
The **probability distribution** for a discrete random variable x is a table, graph, or formula that gives the probability of observing each value of x. We shall denote the probability of x by the symbol $p(x)$.

Thus, the probability distribution for a discrete random variable x may be given by one of the ways:

1. the table

x	p
x_1	p_1
x_2	p_2
…	…
x_n	p_n

where p_k is the probability that the variable x assume the value x_k (k = 1, 2,..., n).

2. a formula for calculating $p(x_k)$ (k = 1, 2,..., n).
3. a graph presenting the probability of each value x_k .

Example 5.6 A balanced coin is tossed twice and the number x of heads is observed. Find the probability distribution for x.

Solution Let H_k and T_k denote the observation of a head and a tail, respectively, on the k^{th} toss, for k = 1, 2. The four simple events and the associated values of x are shown in Table 5.1.

Table 5.1 *Simple events of the experiment of tossing a coin twice*

SIMPLE EVENT	DESCRIPTION	PROBABILITY	NUMBER OF HEADS
E_1	H_1H_2	0.25	2
E_2	H_1T_2	0.25	1
E_3	T_1H_2	0.25	1
E_4	T_1T_2	0.25	0

The event x = 0 is the collection of all simple events that yield a value of x = 0, namely, the simple event E_4. Therefore, the probability that x assumes the value 0 is

$$P(x = 0) = p(0) = P(E_4) = 0.25.$$

The event x = 1 contains two simple events, E_2 and E_3. Therefore,

$$P(x = 1) = p(1) = P(E_2) + P(E_3) = 0.25 + 0.25 = 0.5.$$

Finally,

$$P(x = 2) = p(2) = P(E_1) = 0.25.$$

The probability distribution $p(x)$ is displayed in tabular form in Table 5.2 and as a probability histogram in Figure 5.1.

Table 5.2 *Probability distribution for x, the number of heads in two tosses of a coin*

x	$p(x)$
0	0.25
1	0.5
2	0.25

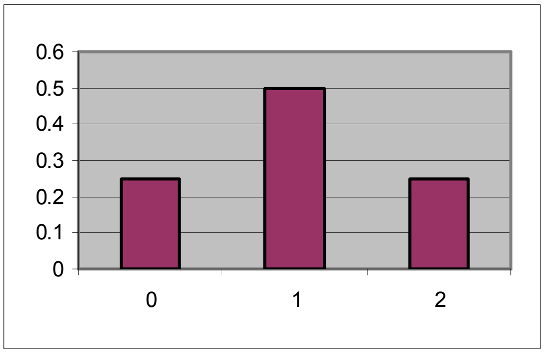


Figure 5.1 *Probability distribution for x, the number of heads in two tosses of a coin*

Properties of the probability distribution for a discrete random variable x

1. $0 \le p(x) \le 1$
2. $\sum_{all\ x} p(x) = 1$

Relationship between the probability distribution for a discrete random variable and the relative frequency distribution of data:

Suppose you were to toss two coins over and over again a very large number of times and record the number x of heads for each toss. A relative frequency distribution for the resulting collection of 0's, 1's and 2's would be very similar to the probability distribution shown in Figure 5.1. In fact, if it were possible to repeat the experiment an infinitely large number of times, the two distributions would be almost identical.

Thus, *the probability distribution of Figure 5.1 provides a model for a conceptual population of values x – the values of x that would be observed if the experiment were to be repeated an infinitely large number of times.*

5.3 Numerical characteristics of a discrete random variable

5.3.1 Mean or expected value

Since a probability distribution for a random variable x is a model for a population relative frequency distribution, we can describe it with numerical descriptive measures, such as its mean and standard deviation, and we can use Chebyshev theorem to identify improbable values of x.

The expected value (or mean) of a random variable x, denoted by the symbol $E(x)$, is defined as follows:

Definition 5.4

Let x be a discrete random variable with probability distribution $p(x)$. Then the **mean or expected value of x** is

$$= E(x) = \sum_{all\ x} xp(x)$$

Example 5.6 Refer to the two-coin tossing experiment of Example 5.5 and the probability distribution for the random variable x, shown in Figure 5.1. Demonstrate that the formula for $E(x)$ gives the mean of the probability distribution for the discrete random variable x.

Solution If we were to repeat the two-coin tossing experiment a large number of times – say 400,000 times, we would expect to observe x = 0 heads approximately 100,000 times, x = 1 head approximately 200,000 times and x = 2 heads approximately 100,000 times. Calculating the mean of these 400,000 values of x, we obtain

$$\mu \approx \frac{\sum x}{n} = \frac{100{,}000(0) + 200{,}000(1) + 100{,}000(2)}{400{,}000} = \frac{1}{4}(0) + \frac{1}{2}(1) + \frac{1}{4}(2) = \sum_{all\ x} p(x)x$$

Thus, the mean of x is $\mu = 1$.

If x is a random variable then any function g(x) of x also is a random variable. The expected value of $g(x)$ is defined as follows:

Definition 5.5

Let x be a discrete random variable with probability distribution $p(x)$ and let $g(x)$ be a function of x . Then the mean or expected value of g(x) is

$$E[g(x)] = \sum_{all\ x} g(x)p(x)$$

5.3.2 Variance and standard deviation

The second important numerical characteristics of random variable are its variance and standard deviation, which are defined as follows:

Definition 5.6

Let x be a discrete random variable with probability distribution $p(x)$. Then the **variance of** x is

$$\sigma^2 = \mathrm{E}[(\mathrm{x} - \mu)^2]$$

The **standard deviation of x** is the positive square root of the variance of x:

$$\sigma = \sqrt{\sigma^2}$$

Example 5.7 Refer to the two-coin tossing experiment and the probability distribution for x, shown in Figure 5.1. Find the variance and standard deviation of x.

Solution In Example 5.6 we found the mean of x is 1. Then

$$\sigma^2 = \mathrm{E}[(\mathrm{x} - \mu)^2] = \sum_{\mathrm{x}=0}^{2} (\mathrm{x} - \mu)^2 p(x) = (0-1)^2\left(\frac{1}{4}\right) + (1-1)^2\left(\frac{1}{2}\right) + (2-1)^2\left(\frac{1}{4}\right) = \frac{1}{2}$$

and

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{2}} \approx 0.707$$

5.4 The binomial probability distribution

Many real-life experiments are analogous to tossing an unbalanced coin a number n of times.

Example 5.8 Suppose that 80% of the jobs submitted to a data-processing center are of a statistical nature. Then selecting a random sample of 10 submitted jobs would be analogous to tossing an unbalanced coin 10 times, with the probability of observing a head (drawing a statistical job) on a single trial equal to 0.80.

Example 5.9 Test for impurities commonly found in drinking water from private wells showed that 30% of all wells in a particular country have impurity A. If 20 wells are selected at random then it would be analogous to tossing an unbalanced coin 20 times, with the probability of observing a head (selecting a well with impurity A) on a single trial equal to 0.30.

Example 5.10 Public opinion or consumer preference polls that elicit one of two responses – Yes or No, Approve or Disapprove,... are also analogous to the unbalanced coin tossing experiment if the size N of the population is large and the size n of the sample is relatively small.

All these experiments are particular examples of a binomial experiment known as a Bernoulli process, after the seventeenth-century Swiss mathematician, Jacob Bernoulli. Such experiments and the resulting binomial random variables have the following characteristics, which form the model of a binomial random variable.

Model (or characteristics) of a binomial random variable

1. The experiment consists of n identical trials
2. There are only 2 possible outcomes on each trial. We will denote one outcome by S (for Success) and the other by F (for Failure).
3. The probability of S remains the same from trial to trial. This probability will be denoted by p, and the probability of F will be denoted by q (q = 1-p).
4. The trials are independent.
5. The binomial random variable x is the number of S' in n trials.

The binomial probability distribution, its mean and its standard deviation are given the following formulas:

The probability distribution, mean and variance for a binomial random variable:

1. **The probability distribution:**

$p(x) = C_n^x p^x q^{n-x}$ (x = 0, 1, 2, ..., n),

where

p = probability of a success on a single trial, q=1-p

n = number of trials, x= number of successes in n trials

$C_n^x = \frac{n!}{x!(n-x)!}$ = combination of x from n.

2. **The mean:** $\mu = np$
3. **The variance:** $\sigma^2 = npq$

Example 5.11 (see also Example 5.9) Test for impurities commonly found in drinking water from private wells showed that 30% of all wells in a particular country have impurity A. If a random sample of 5 wells is selected from the large number of wells in the country, what is the probability that:

a) Exactly 3 will have impurity A?
b) At least 3?
c) Fewer than 3?

Solution First we confirm that this experiment possesses the characteristics of a binomial experiment. This experiment consists of n = 5 trials, one corresponding to each random selected well. Each trial results in an S (the well contains impurity A) or an F (the well does not contain impurity A). Since the total number of wells in the country is large, the probability of drawing a single well and finding that it contains impurity A is equal to 0.30 and this probability will remain the same for each of the 5 selected wells. Further, since the sampling is random, we assume that the outcome on any one well is unaffected by the outcome of any other and that the trials are independent. Finally, we are interested in the number x of wells in the sample of n = 5 that contain impurity A. Therefore, the sampling process represents a binomial experiment with n = 5 and p = 0.30.

a) The probability of drawing exactly x = 3 wells containing impurity A is

$p(x) = C_n^x p^x q^{n-x}$ with n = 5, p = 0.30 and x = 3. We have by this formula

$$p(3) = \frac{5!}{3!2!}(0.30)^3(1-0.30)^{5-3} = 0.1323\,.$$

b) The probability of observing at least 3 wells containing impurity A is

$P(x \geq 3)$ = p(3)+p(4)+p(5). We have calculated p(3) = 0.1323 and we leave to the reader to verify that p(4) = 0.02835, p(5) = 0.00243. In result, P(3) = 0.1323+0.02835+0.00243 = 0.16380.

c) Although $P(x<3)$ = p(0)+p(1)+p(2), we can avoid calculating 3 probabilities by using the complementary relationship $P(x<3)$ = 1-$P(x\geq$ 3) = 1-0.16380 = 0.83692.

5.5 The Poisson distribution

The Poisson probability distribution is named for the French mathematician S.D. Poisson (1871-1840, It is used to describe a number of processes, including the distribution of telephone calls going through a switchboard system, the demand of patients for service at a health institution, the arrivals of trucks and cars at a tollbooth, and the number of accidents at an intersection.

Characteristics defining a Poisson random variable

1. The experiment consists of counting the number x of times a particular event occurs during a given unit of time
2. The probability that an event occurs in a given unit of time is the same for all units.
3. The number of events that occur in one unit of time is independent of the number that occur in other units.
4. The mean number of events in each unit will be denoted by the Greek letter λ

The formulas for the probability distribution, the mean and the variance of a Poisson random variable are shown in the next box.

The probability distribution, mean and variance for a Poisson random variable x:

1. **The probability distribution:**

$$p(x) = \frac{\lambda^{x} e^{-\lambda}}{x!} \quad (x = 0, 1, 2, ...),$$

where

λ = mean number of events during the given time period,

e = 2.71828...(the base of natural logarithm).

2. **The mean:** $\mu = \lambda$

3. **The variance:** $\sigma^2 = \lambda$

Note that instead of time, the Poisson random variable may be considered in the experiment of counting the number x of times a particular event occurs during a given unit of area, volume, etc.

Example 5.12 Suppose that we are investigating the safety of a dangerous intersection. Past police records indicate a mean of 5 accidents per month at this intersection. Suppose the number of accidents is distributed according to a Poisson distribution. Calculate the probability in any month of exactly 0, 1, 2, 3 or 4 accidents.

Solution Since the number of accidents is distributed according to a Poisson distribution and the mean number of accidents per month is 5, we have the probability of happening

accidents in any month $p(x) = \frac{5^x e^{-5}}{x!}$. By this formula we can calculate

$p(0)$ = 0.00674, $p(1)$ = 0.3370, $p(2)$ = 0.08425, $p(3)$ = 0.14042, $p(4)$ = 0.17552.

The probability distribution of the number of accidents per month is presented in Table 5.3 and Figure 5.2.

Table 5.3 *Poisson probability distribution of the number of accidents per month*

X- NUMBER OF ACCIDENTS	$P(\mathrm{X})$ - PROBABILITY
0	0.006738
1	0.03369
2	0.084224
3	0.140374
4	0.175467
5	0.175467
6	0.146223
7	0.104445
8	0.065278
9	0.036266
10	0.018133
11	0.008242
12	0.003434

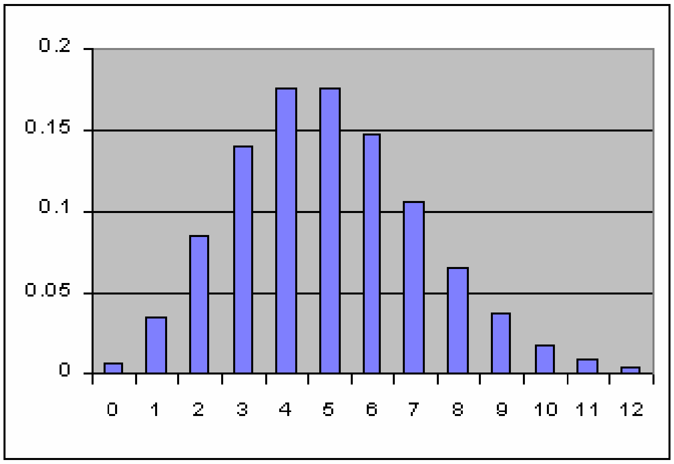


Figure 5.2 *The Poisson probability distribution of the number of accidents*

5.6 Continuous random variables: distribution function and density function

Many random variables observed in real life are not discrete random variables because the number of values they can assume is not countable. In contrast to discrete random variables, these variables can take on any value within an interval. For example, the daily rainfall at some location, the strength of a steel bar and the intensity of sunlight at a particular time of day. In Section 5.1 these random variables were called continuous random variables.

The distinction between discrete random variables and continuous random variables is usually based on the difference in their cumulative distribution functions.

Definition 5.7
Let ξ be a continuous random variable assuming any value in the interval (-∞, +∞). Then **the cumulative distribution function *F(x)*** of the variable ξ is defined as follows

$$F(x) = P(\xi \le x)$$

i.e., $F(x)$ is equal to the probability that the variable ξ assumes values, which are less than or equal to x.

Note that here and from now on we *denote by letter ξ a continuous random variable and denote by x a point on number line.*

From the definition of the cumulative distribution function $F(x)$ it is easy to show the following its properties.

Properties of the cumulative distribution function *F(x)* for a continuous random variable ξ

1. $0 \le F(x) \le 1$,
2. F(x) is a monotonically non-decreasing function, that is, if $a \le b$ then $F(a) \le F(b)$ for any real numbers a and b.
3. $P(a \le \xi \le b) = F(b) - F(a)$
4. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$

In Chapter 2 we described a large data set by means of a relative frequency distribution. If the data represent measurements on a continuous random variable and if the amount of data is very large, we can reduce the width of the class intervals until the distribution appears to be a smooth curve. A probability density is a theoretical model for this distribution.

Definition 5.8

If $F(x)$ is the cumulative distribution function for a continuous random variable ξ then the **density probability function f(x)** for ξ is

$$f(x) = F'(x),$$

i.e., $f(x)$ is the derivative of the distribution function $F(x)$.

The density function for a continuous random variable ξ, the model for some real-life population of data, will usually be a smooth curve as shown in Figure 5.3.

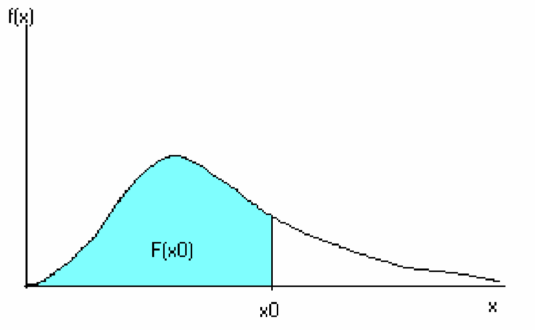


Figure 5.3 *Density function f(x) for a continuous random variable*

It follows from Definition 5.8 that

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Thus, the cumulative area under the curve between $-\infty$ and a point x_0 is equal to $F(x_0)$.

The density function for a continuous random variable must always satisfy the two properties given in the box.

Properties of a density function

1. $f(x) \geq 0$

2. $\int\limits_{-\infty}^{+\infty} f(x)dx = F(\infty) = 1$

5.7 *Numerical characteristics of a continuous random variable*

Definition 5.8

Let ξ be a continuous random variable with density function $f(x)$. Then **the mean or the expected value of** ξ is

$$E(\xi) = \int\limits_{-\infty}^{+\infty} xf(x)dx$$

Definition 5.9

Let ξ be a continuous random variable with density function $f(x)$ and $g(x)$ is a function of x. Then the mean or the expected value of $g(\xi)$ is

$$E[g(\xi)] = \int\limits_{-\infty}^{+\infty} g(x)f(x)dx$$

Definition 5.10

Let ξ be a continuous random variable with the expected value $\mathrm{E}(\xi) = \mu$. Then **the variance** of ξ is

$$\sigma^2 = \mathrm{E}[(\xi - \mu)^2]$$

The standard deviation of ξ is the positive square root of the variance $\sigma = \sqrt{\sigma^2}$

5.8 Normal probability distribution

The normal (or Gaussian) density function was proposed by C.F.Gauss (1777-1855) as a model for the relative frequency distribution of errors, such errors of measurement. Amazingly, this bell-shaped curve provides an adequate model for the relative frequency distributions of data collected from many different scientific areas.

The density function, mean and variance for a normal random variable

The density function:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

The parameters μ and σ^2 are the mean and the variance , respectively, of the normal random variable

There is infinite number of normal density functions – one for each combination of μ and σ. The mean measures the location and the variance measures its spread. Several different normal density functions are shown in Figure 5.4.

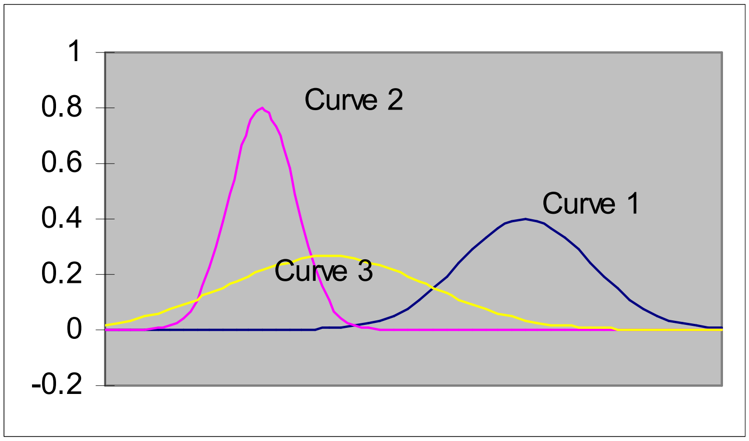


Figure 5.4 *Several normal distributions: Curve 1 with* $\mu=3,\sigma=1$, *Curve 2 with* $\mu=-1,\sigma=0$, *and Curve 3 with* $\mu=0,\sigma=1.5$,

If μ = 0 and σ =1 then $f(x)=\frac{1}{\sqrt{2\pi}}e^{-(x-\mu)^2/2}$. The distribution with this density function is called the standardized normal distribution. The graph of the standardized normal density distribution is shown in Figure 5.5.

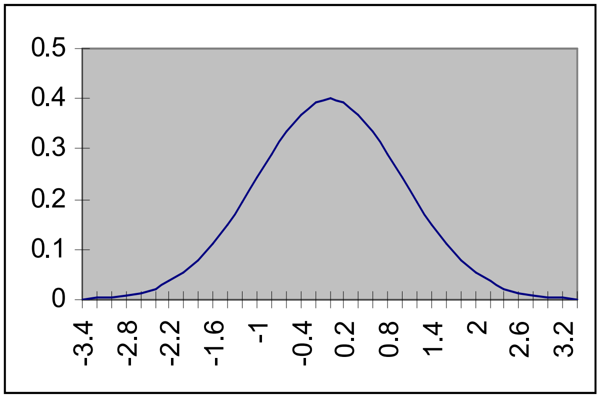


Figure 5.5 *The standardized normal density distribution*

If ξ is a normal random variable with the mean μ and variance σ then

1) the variable

$$z = \frac{\xi - \mu}{\sigma}$$

is the standardized normal random variable.

2) $P(|\xi - \mu| \le n\sigma\ ) = 2\Phi(n)$, where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt$$

This function is called the Laplace function and it is tabulated.

In particular, we have

$$P(|\xi - \mu| \le \sigma\ ) = 0.6826$$

$$P(|\xi - \mu| \le 2\sigma\ ) = 0.9544$$

$$P(|\xi - \mu| \le 3\sigma\ ) = 0.9973$$

These equalities are known as σ, 2σ and σ rules, respectively and are often used in statistics. Namely, if a population of measurements has approximately a normal distribution the probability that a random selected observation falls within the intervals $(\mu - \sigma, \mu + \sigma)$, $(\mu - 2\sigma, \mu + 2\sigma)$, and $(\mu - 3\sigma, \mu + 3\sigma)$, is approximately 0.6826, 0.9544 and 0.9973, respectively.

The normal distribution as an approximation to various discrete probability distributions

Although the normal distribution is continuous, it is interesting to note that it can sometimes be used to approximate discrete distributions. Namely, we can use normal distribution to approximate binomial probability distribution.

Suppose we have a binomial distribution defined by two parameters: the number of trials n and the probability of success p. The normal distribution with the parameters μ and σ will be a good approximation for that binomial distribution if both

$\mu - 2\sigma = np - 2\sqrt{\mathrm{np}(1-\mathrm{p})}$ and $\mu + 2\sigma = np + 2\sqrt{np(1-p)}$ lie between 0 and n.

For example, the binomial distribution with n = 10 and p = 0.5 is well approximated by the normal distribution with μ = np = 10*0.5 = 5.0 and $= \sqrt{np(1-p)}$ = 0.5*$\sqrt{10}$ = 1.58. See Figure 5.6 or Table 5.4.

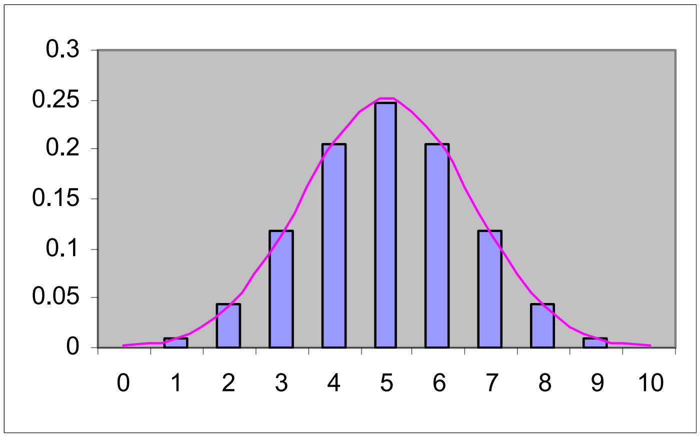
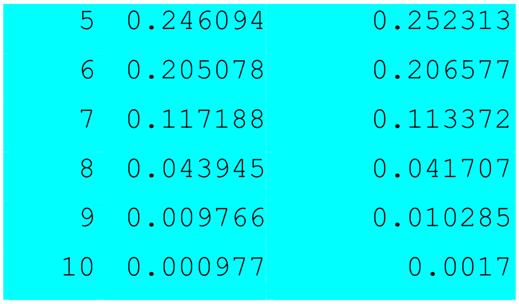


Figure 5.6 *Approximation of binomial distribution (bar graph) with n=10, p=0.5 by a normal distribution (smoothed curve)*

Table 5.4 *The binomial and normal probability distributions for the same values of x*

x	Binomial distribution	Normal distribution
0	0.000977	0.0017
1	0.009766	0.010285
2	0.043945	0.041707
3	0.117188	0.113372
4	0.205078	0.206577



5	0.246094	0.252313
6	0.205078	0.206577
7	0.117188	0.113372
8	0.043945	0.041707
9	0.009766	0.010285
10	0.000977	0.0017

5.9. Summary

This chapter introduces the notion of a random variable – one of the fundamental concepts of the probability theory. It is a rule that assigns one and only one value of a variable x to each simple event in the sample space. A variable is said to be discrete if it can assume only a countable number of values.

The probability distribution of a discrete random variable is a table, graph or formula that gives the probability associated with each value of x. The expected value $E(x) = \mu$ is the mean of this probability distribution and $E[(x-\mu)] = \sigma^2$ is its variance.

Two discrete random variables – the binomial, and the Poisson – were presented, along with their probability distributions.

In contrast to discrete random variables, continuous random variable can assume value corresponding to the infinitely large number can assume value corresponding to the infinitely large number of points contained in one or more intervals on the real line. The relative frequency distribution for a population of data associated with a continuous random variable can be modeled using a probability density function. The expected value (or mean) of a continuous random variable x is defined in the same manner as for discrete random variables, except that integration is substituted for summation. The most important probability distribution – the normal distribution - with its properties is considered.

5.10 Exercises

1) The continuous random variable ξ is called a uniform random variable if its density function is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

Show that for this variable, the mean $\mu = \dfrac{a+b}{2}$ and the variance $\sigma^2 = \dfrac{(b-a)^2}{12}$.

2) The continuous random variable ξ is called a exponential random variable if its density function is

$$f(x) = \frac{e^{-x/\beta}}{\beta} \quad (0 \le x \le \infty)$$

Show that for this random variable $\mu = \beta, \quad \sigma^2 = \beta^2$.

3) Find the area beneath a standardized normal curve between the mean z = 0 and the point z = -1.26.

4) Find the probability that a normally distributed random variable ξ lie more than z = 2 standard deviations above its mean.

5) Suppose *y* is normally distributed random variable with mean 10 and standard deviation 2.1.

a) Find $P(y \ge 11)$.

b) Find $P(7.6 \le y \le 12.2)$

Chapter 6. Sampling Distributions

CONTENTS

6.1 Why the method of sampling is important

Much of our statistical information comes in the form of samples from populations of interests. In order to develop and evaluate methods for using sample information to obtain knowledge of the population, it is necessary to know how closely a descriptive quantity such as the mean or the median of a sample resembles the corresponding population quantity. In this chapter, the ideas of probabilities will be used to study the sample-to-sample variability of these descriptive quantities.

We now return to the objective of statistics - namely, the use of sample information to infer the nature of a population. We will explain why the method of sampling is important through an example.

Example 6.1 **The Vietnam Demographic and Health Survey (VNDHS) was a nationwide representative sample survey conducted in May 1988 to collect data on fertility and a few indicators of child and maternal health. In the survey a total of 4,171 eligible women, ale aged 15 to 49 years old were interviewed. The survey data was given in Appendix A by the format of Excel. The relative frequency distribution for number of children ever born for 4,171 women appears as in the Table 6.1 and in Figure 6.1. In actual practice, the entire population of 4,171 women's number of children ever born may not be easily accessible. Now, we draw two samples of 50 women from the population of 4,171 women. The relative frequency distributions of the two samples are given in Table 6.2a and 6.2b and graphed in Figures 6.2a and 6.2b.**
Click here for Simulation in SPSS.

Compare the distributions of number of children ever born for two samples. Which appears to better characterize number of children ever born for the population?

Solution **It is clear that the two samples lead to quite different conclusions about the same population from which they were both selected. From Figure 6.2a, we see that only 18% of the sampled women bore 3 children, whereas from Figure 6.2b, we see that 26% of the sampled women bore such number of children. This may be compared to the relative frequency distribution for the population (shown in Figure 6.1), in which we observe that 18% of all the women bore 3 children. In addition, note that none of the women in the second sample (Figure 6.2b) had no children, whereas 10% of the women**

in the first sample (Figure 6.2a) had no child. This value from the first sample compare favorably with the 7% of "no children" of the entire population (Figure 6.1).

Table 6.1 *Frequency distribution of number of children ever born for 4,171 women*

Number of Children	Frequency	Relative Frequency
0	312	0.07
1	708	0.17
2	881	0.21
3	737	0.18
4	570	0.14
5	354	0.08
6	243	0.06
7	172	0.04
>7	194	0.05
Total	4171	1.00

Figure 6.1 *Relative frequency distribution of number of children ever born for 4,171 women*

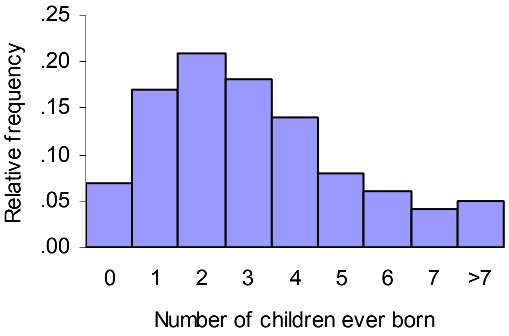


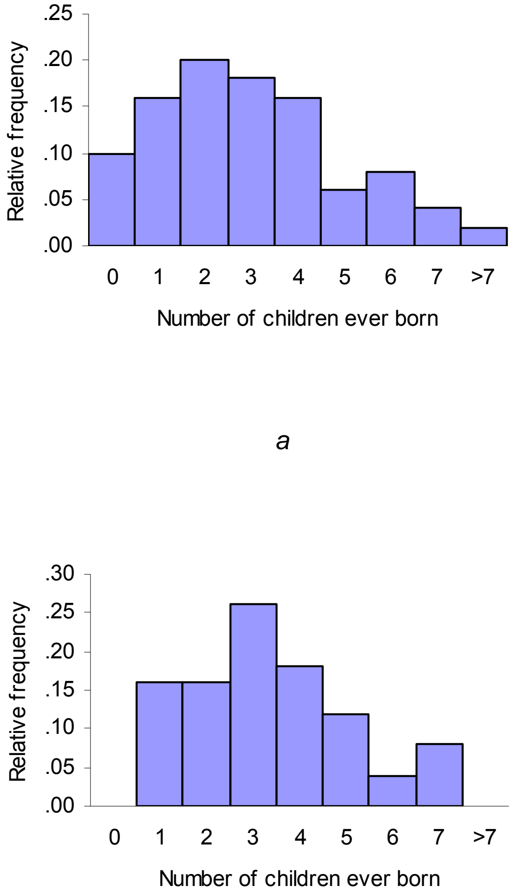
Table 6.2 *Frequency distribution of number of children ever born for each of two samples of 50 women selected from 4,171 women*

Number of Children	Frequency	Relative Frequency
0	5	0.10
1	8	0.16
2	10	0.20
3	9	0.18
4	8	0.16
5	3	0.06
6	4	0.08
7	2	0.04
>7	1	0.02
Total	50	1.00

a

Number of Children	Frequency	Relative Frequency
0	0	0.00
1	8	0.16
2	8	0.16
3	13	0.26

Figure 6.2 *Frequency distribution of number of children ever born for each of two samples of 50 women selected from 4,171 women*



4	9	0.18
5	6	0.12
6	2	0.04
7	4	0.08
Total	50	1.00

b *b*

To rephrase the question posed in the example, we could ask: Which of the two samples is more representative of, or characteristics of, the number of children ever born for all 4,171 of the VNDHS's women? Clearly, the information provided by the first sample (Table and Figure 6.2a) gives a better picture of the actual population of numbers of children ever born. Its relative frequency distribution is closer to that for the entire population (Table and Figure 6.1) than is the one provided by the second sample (Table and Figure 6.2b). Thus, if we were to rely on information from the second sample only, we may have a distorted, or biased, impression of the true situation with respect to numbers of children ever born.

How is it possible that two samples from the same population can provide contradictory information about the population? The key issue is the method by which the samples are obtained. The examples in this section demonstrate that great care must be taken in order to select a sample that will give an unbiased picture of the population about which inferences are to be made. One way to cope with this problem is to use random sampling. Random sampling eliminates the possibility of bias in selecting a sample and, in addition, provides a probabilistic basic for evaluating the reliability of an inference. We will have more to say about random sampling in Section 6.2.

6.2 Obtaining a Random Sample

In the previous section, we demonstrated the importance of obtaining a sample that exhibits characteristics similar to those possessed by the population from which it came, the population about which we wish to make inferences. One way to satisfy this requirement is to select the sample in such a way that every different sample of size *n* has an equal probability of being selected. This procedure is called *random sampling* and the resulting sample is called a *random sample of size n*. In this section we will explain how to draw a random sample, and will then employ random sampling in sections that follow.

Definition 6.1

A random sample of n experimental units is one selected in such a way that every different sample of size n has an equal probability of selection.

Example 6.2 **A city purchasing agent can obtain stationery and office supplies from any of eight companies. If the purchasing agent decides to use three suppliers in a given year and wants to avoid accusations of bias in their selection, the sample of three suppliers should be selected from among the eight.**

a. How many different samples of three suppliers can be chosen from among the eight?
b. List them.
c. State the criterion that must be satisfied in order for the selected sample to be random.

Solution **In this example, the population of interest consists of eight suppliers (call them A, B, C, D, E, F, G, H). from which we want to select a sample of size n = 3. The numbers of different samples of n = 3 elements that can be selected from a population of N = 8 elements is**

$$C_n^N = \frac{N!}{n!(N-n)!} = \frac{8!}{3!5!} = \frac{8*7*6*5*4*3*2*1}{(3*2*1)(5*4*2*1)} = 56$$

a. The following is a list of 56 samples:

A, B, C
A, B, D
A, B, E
A, B, F
A, B, G
A, B, H
A, C, D
A, C, E
A, C, F
A, C, G
A, C, H
A, D, E
A, D, F
A, D, G
A, D, H
A, E, F
A, E, G
A, E, H
A, F, G
A, F, H
A, G, H
B, C, D
B, C, E
B, C, F
B, C, G
B, C, H
B, D, E
B, D, F
B, D, G
B, D, H
B, E, F
B, E, G
B, E, H
B, F, G
B, F, H
B, G, H
C, D, E
C, D, F
C, D, G
C, D, H
C, E, F
C, E, G
C, E, H
C, F, G
C, F, H
C, G, H
D, E, F
D, E, G
D, E, H
D, F, G
D, F, H
D, G, H
E, F, G
E, F, H
E, G, H
F, G, H

b. Each sample must have the same chance of being selected in order to ensure that we have a random sample. Since there are 56 possible samples of size n = 3, each must have a probability equal to 1/56 of being selected by the sampling procedure.

What procedures may one use to generate a random sample? If the population is not too large, each observation may be recorded on a piece of paper and placed in a suitable container. After the collection of papers is thoroughly mixed, the researcher can remove n pieces of paper from container; the elements named on these n pieces of paper would be ones included in the sample.

However, this method has the following drawbacks: It is not feasible when the population consists of a lager number of observations; and since it is very difficult to achieve a thorough mixing, the procedure provides only an approximation to random sample.

A more practical method of generating a random sample, and one that may be used with lager populations, is to use a table of random numbers. At present, in almost statistical program packages this method is used to select random samples. For example, SPSS PC - a comprehensive system for analyzing data, provides a procedure to select a random sample based on an approximate percentage or an exact number of observations. Two samples in Example 6.1 were drawn by the SPSS's "Select cases" procedure from the data on fertilities of 4,171 women recorded in Appendix A.

For the first sample, the mean is

$$\bar{x} = \frac{\sum vf}{n} = \frac{0*5+1*8+2*10+3*9+4*8+5*3+6*4+7*2+8*1}{50} = 2.96$$

For the second sample, the mean is

$$\bar{x}=\frac{\sum vf}{n}=\frac{0*0+1*8+2*8+3*13+4*9+5*6+6*2+7*4+8*0}{50}=3.38$$

where the mean for all 4,171 observations is 3.15. In the next section, we discuss how to judge the performance of a statistic computed from a random sample.

6.3 Sampling Distribution

In the previous section, we learned how to generate a random sample from a population of interest, the ultimate goal being to use information from the sample to make an inference about the nature of the population. In many situations, the objective will be to estimate a numerical characteristic of the population, called a parameter, using information from sample. For example, from the first sample of *50* women in the Example 6.1, we computed $\bar{x}=2.96$, the mean number of children ever born from the sample of n = *50*. In other word, we used the sample information to compute a statistic - namely, the sample mean, $\bar{x}$.

Definition 6.2

A numerical descriptive measure of a population is called a parameter.

Definition 6.3

A quantity computed from the observations in a random sample is called a statistic.

You may have observed that the value of a population parameter (for example, the mean μ) is a constant (although it is usually unknown to us); its value does not vary from sample to sample. However, the value of a sample statistic (for example, the sample mean $\bar{x}$) is highly dependent on the particular sample that is selected. As seen in the previous section, the means of two samples with the same size of n = 50 are different.

Since statistics vary from sample to sample, any inferences based on them will necessarily be subject to some uncertainty. How, then, do we judge the reliability of a sample statistic as a tool in making an inference about the corresponding population parameter? Fortunately, the uncertainty of a statistic generally has characteristic properties that are known to us, and that are reflected in its sampling distribution. Knowledge of the sampling distribution of a particular statistic provides us with information about its performance over the long run.

Definition 6.4

A sampling distribution of a sample statistic (based on *n* observations) is the relative frequency distribution of the values of the statistic theoretically generated by taking repeated random samples of size *n* and computing the value of the statistic for each sample. (See Figure 6.3.)

We will illustrate the notion of a sampling distribution with an example, which our interest focuses on the numbers of children ever born of 4,171 women in VNDHS 1988. The data are given in Appendix A. In particular, we wish to estimate the mean number of children ever born to all women. In this case, the 4,171 observations constitute the entire population and we know that the true value of μ, the mean of the population, is 3.15 children.

Example 6.3 **How could we generate the sampling distribution of $\bar{x}$, the mean of a random sample of *n* = 5 observations from population of 4,171 numbers of children ever born in Appendix A?**

Solution **The sampling distribution for the statistic $\bar{x}$, based on a random sample of *n* = 5 measurements, would be generate in this manner: Select a random sample of five measurements from the population of 4,171 observations on number of children ever born in Appendix A; compute and record the value of $\bar{x}$ for this sample. Then return these five measurements to the population and repeat the procedure. (See Figure 6.3). If this sampling procedure could be repeated an infinite number of times, the infinite number of values of $\bar{x}$ obtained could be summarized in a relative frequency distribution, called the sampling distribution of $\bar{x}$.**

The task described in Example 6.3, which may seem impractical if not impossible, is not performed in actual practice. Instead, the sampling distribution of a statistic is obtained by applying mathematical theory or computer simulation, as illustrated in the next example.

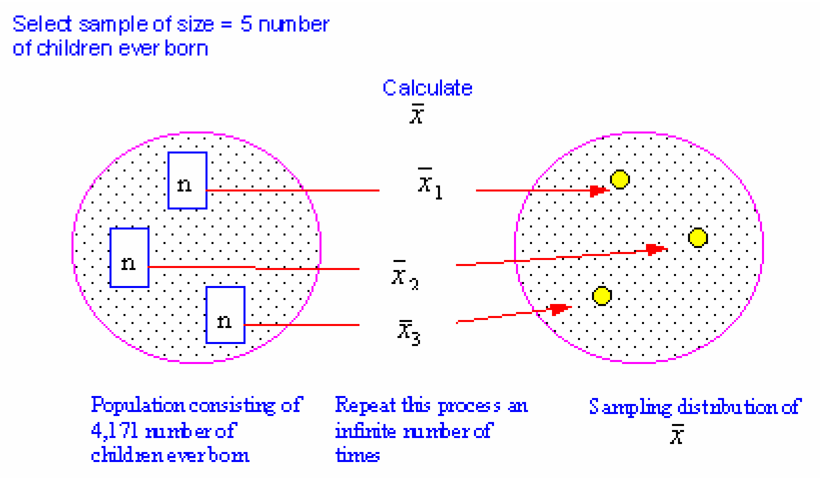

Select sample of size = 5 number of children ever born
Calculate $\bar{x}$
n
n
n
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
Population consisting of 4,171 number of children ever born
Repeat this process an infinite number of times
Sampling distribution of $\bar{x}$

Figure 6.3 *Generating the theoretical sampling distribution of the sample mean $\bar{x}$*

Example 6.4 **Use computer simulation to find the approximate sampling distribution of $\bar{x}$, the mean of a random sample of *n* = *5* observations from the population of 4,171 number of children ever born in Appendix A.**

Solution **We used a statistical program, for example SPSS, to obtain 100 random samples of size *n* = *5* from target population. The first ten of these samples are presented in Table 6.3.**

Table 6.3 *The first ten of samples of n = 5 measurement from population of numbers of children ever born of 4,171 women*

Sample	*Number of children ever born*					*Mean ($\bar{x}$)*
1	1	1	1	2	2	1.4
2	1	2	3	3	3	2.4
3	0	0	4	6	7	3.4

4	0	1	2	2	3	1.6
5	2	2	3	4	7	3.6
6	1	2	3	5	8	3.8
7	1	2	2	5	6	3.2
8	1	2	2	3	6	2.8
9	2	2	3	3	11	4.2
10	0	0	2	3	4	1.8

For each sample of five observations, the sample mean $\bar{x}$ was computed. The relative frequency distribution of the number of children ever born for the entire population of 4,171 women was plotted in Figure 6.4 and the 100 values of $\bar{x}$ are summarized in the relative frequency distribution shown in Figure 6.5.

Click here to see some scripts and print outs from sampling and case summarize procedures in SPSS with sample size of n = 5.

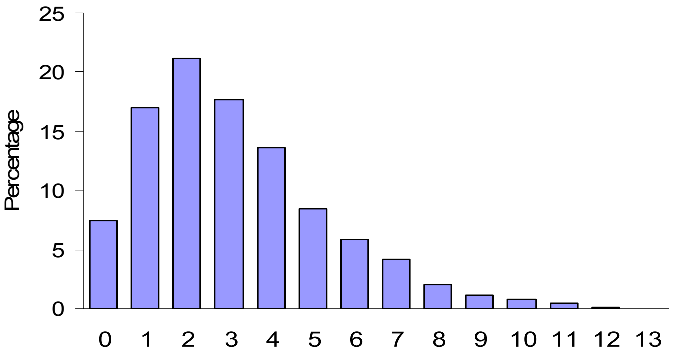


Figure 6.4 *Relative frequency distribution for 4,171 numbers of children ever born*

We can see that the value of $\bar{x}$ in Figure 6.5 tend to cluster around the population mean, μ = 3.15 children. Also, the values of the sample mean are less spread out (that is, they have less variation) than the population values shown in Figure 6.4. These two observations are borne out by comparing the means and standard deviations of the two sets of observations, as shown in Table 6.4.

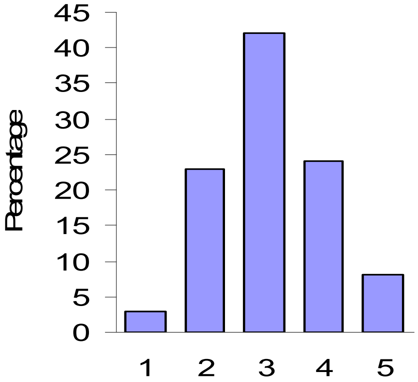


Figure 6.5 *Sampling distribution of $\bar{x}$: Relative frequency distribution of $\bar{x}$ based on 100 samples of size n = 5*

Table 6.4 *Comparison of the population and the approximate sampling distribution of $\bar{x}$ based on 100 samples of size n = 5*

	Mean	**Standard Deviation**
Population of 4,171 numbers of children ever born (Fig. 6.4)	μ = 3.15	σ = 2.229
100 values of $\bar{x}$ based on samples of size n = 5 (Fig. 6.5)	3.11	.920

Example 6.5 **Refer to Example 6.4. Simulate the sampling distribution of $\bar{x}$ for samples size *n* = *25* from population of 4,171 observations of number of children ever born. Compare result with the sampling distribution of $\bar{x}$ based on samples of *n* = 5, obtained in Example 6.4.**

Solution **We obtained 100 computer-generated random samples of size *n* = 25 from target population. A relative frequency distribution for 100 corresponding values of $\bar{x}$ is shown in Figure 6.6.**

It can be seen that, as with the sampling distribution based on samples of size n = 5, the values of $\bar{x}$ tend to center about the population mean. However, a visual inspection shows that the variation of the $\bar{x}$ *-values* about their mean in Figure 6.6 is less than the variation in the values of $\bar{x}$ based on samples of size n = 5 (Figure 6.5). The mean and standard deviation for these 100 values of $\bar{x}$ are shown in Table 6.5 for comparison with previous results.

Table 6.5 *Comparison of the population distribution and the approximate sampling distributions of $\bar{x}$, based on 100 samples of size n = 5 and n = 25*

	Mean	**Standard Deviation**
Population of 4,171 numbers of children ever born (Fig. 6.4)	μ = 3.15	σ = 2.229
100 values of $\bar{x}$ based on samples of size n = 5 (Fig. 6.5)	3.11	.920
100 values of $\bar{x}$ based on samples of size n = 25 (Fig. 6.6)	3.14	.492

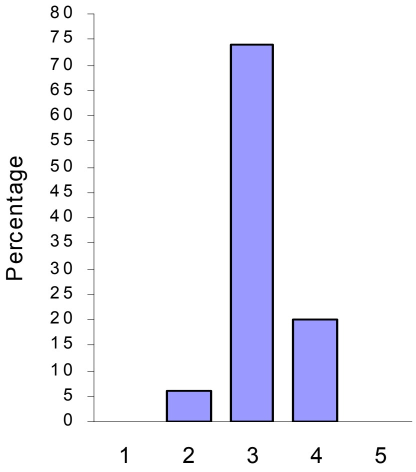


Figure 6.6 *Relative frequency distribution of $\bar{x}$ based on 100 samples of size n = 25*

Click here to see some scripts and print outs from sampling and case summarize procedures in SPSSS with sample size of n = 25.

From Table 6.5 we observe that, as the sample size increases, there is less variation in the sampling distribution of $\bar{x}$; that is, the values of $\bar{x}$ tend to cluster more closely about the population mean as n gets larger. This intuitively appealing result will be stated formally in the next section.

6.4 The sampling distribution of $\bar{x}$: the Central Limit Theorem

Estimating the mean number of children ever born for a population of women, or the mean height for all 3-year old boys in a day-care center are examples of practical problems in which the goal is to make an inference about the mean, μ, of some target population. In previous sections, we have indicated that the mean $\bar{x}$ is often used as a tool for making an inference about the corresponding population parameter μ, and we have shown how to approximate its sampling distribution. The following theorem, of fundamental importance in statistics, provides information about the actual sampling distribution of $\bar{x}$.

The Central Limit Theorem

If the size is sufficiently large, the mean $\bar{x}$ of a random sample from a population has a sampling distribution that is approximately normal, regardless of the shape of the relative frequency distribution of the target population. As the sample size increases, the better will be the normal approximation to the sampling distribution. (*)

The sampling distribution of $\bar{x}$, in addition to being approximately normal, has other known characteristics, which are summarized as follows.

Properties of Sampling Distribution of $\bar{x}$

If $\bar{x}$ is the mean of a random sample of size n from a population with mean μ and standard deviation σ, then:

1. The sampling distribution of $\bar{x}$ has a mean equal to the mean of the population from which the sample was selected. That is, if we let $\mu_{\bar{x}}$ denote the mean of the sampling distribution of $\bar{x}$, then

$$\mu_{\bar{x}} = \mu$$

2. The sampling distribution of $\bar{x}$ has a standard deviation equal to the standard deviation of the population from which the sample was selected, divided by the square root of the sample size. That is, if we let $\sigma_{\bar{x}}$ denote the standard deviation of the sampling distribution of $\bar{x}$, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(*) This is why the normal distribution is so important!

Example 6.6 **Show that the empirical evidence obtained in Examples 6.4 and 6.5 supports the Central Limit Theorem and two properties of the sampling distribution of $\bar{x}$. Recall that in Examples 6.4 and 6.5, we obtained repeated random samples of size *n* = *5* and *n* = *25* from the population of numbers of children ever born in Appendix A. For this target population, we know that the values of the parameters μ and σ:**

Population mean: μ *= 3.15 children*

Population standard deviation: σ *= 2.229 children*

Solution In Figures 6.4 and 6.5, we note that the values of $\bar{x}$ tend to cluster about the population mean, μ = *3.15.* This is guaranteed that by property 1, which implies that, in the long run, the average of all values of $\bar{x}$ that would be generated in infinite repeated sampling would be equal to μ.

We also observed, from Table 6.5, that the standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases from n = *5* to n = *25*. Property 2 quantifies the decrease and relates it to the sample size. As an example, note that, for our approximate sampling distribution based on samples of size n = *5*, we obtained a standard deviation of .920, whereas property 2 tells us that, for the actual sampling distribution of $\bar{x}$, the standard deviation is equal to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.229}{\sqrt{5}} = .997$$

Similarly, for samples of size n = 25, the sampling distribution of $\bar{x}$ actually has a standard deviation of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.229}{\sqrt{25}} = .446$$

The value we obtained by simulation was .492

Finally, the Central Limit Theorem guarantees an approximately normal distribution for $\bar{x}$, regardless of the shapes of the original population. In our examples, the population from which the samples were selected is seen in Figure 6.4 to be moderately skewed to the right. Note from Figure 6.5 and 6.6 that, although the sampling distribution of $\bar{x}$ tends to be bell-shaped in each case, the normal approximation improves when the sample size is increased from n = *5* (Figure 6.5) to n = *25* (Figure 6.6).

Example 6.7 **In research on the health and nutrition of children in a rural area of Vietnam 1988, it was reported that the average height of 823 three-year old children in rural areas in 1988 was 89.67 centimeters with a standard deviation of 4.99 centimeters. These observations are given in Appendix B. In order to check these figures, we will randomly sample 100 three-year old children from the rural area and monitor their heights.**

a. Assuming the report's figures is true, describe the sampling distribution of the mean height for a random sample of 100 three year old children in the rural.
b. Assuming the report's figures are true, what is probability that the sample mean height will be at least 91 centimeters?

Solution

a. Although we have no information about the shape of the relative frequency distribution of the heights of the children, we can apply the Central Limit Theorem to conclude that the sampling distribution of the sample mean height of the 100 three year old children is approximately normally distributed. In addition, the mean $\mu_{\bar{x}}$, and the standard deviation, $\sigma_{\bar{x}}$, of the sampling distribution are given by

$$\mu_{\bar{x}} = \mu = 91\ cm \quad \text{and}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.99}{\sqrt{100}} = .499\,cm$$

assuming that the reported values of μ and σ are correct.

b. If the reported values are correct, then $P(\bar{x} \geq 91)$, the probability of observing a mean height of 91 cm or higher in the sample of 100 observations, is equal to the greened area shown in Figure 6.7.

Since the sampling distribution is approximately normal, with mean and standard deviation as obtained in part a, we can compute the desired area by obtaining the z-score for $\bar{x}$ = *91*

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{91 - 89.67}{.499} = 2.67$$

Thus, $P(\bar{x} \geq 91) = P(z \geq 2.67)$, and this probability (area) may be found in Table 1 of Appendix C.

$$\begin{aligned} P(\bar{x} \geq 91) &= P(z \geq 2.67) \\ &= .5 - A \text{ (see Figure 6.7)} \\ &= .5 - .4962 \\ &= .0038 \end{aligned}$$

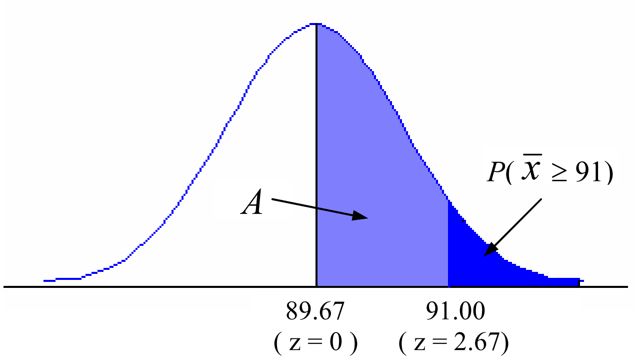


Figure 6.7 *Sampling distribution of $\bar{x}$ in Example 6.7*

The probability that we would obtain a sample mean height of 91 cm or higher is only .0038, if the reported values are true. If the 100 randomly selected three year old children have an average height of 91 cm or higher, we would have strong evidence that the reported values are false, because such a larger sample mean is very unlikely to occur if the research is true.

In practical terms, the Central Limit Theorem and two properties of the sampling distribution of

$\bar{x}$

$\bar{x}$ assure us that the sample mean $\bar{x}$ is a reasonable statistic to use in making inference about the population mean μ, and they allow us to compute a measure of the reliability of references made about μ. As we notice earlier, we will not be required to obtain sampling distributions by simulation or by mathematical arguments. Rather, for all the statistics to be used in this course, the sampling distribution and its properties will be presented as the need arises.

6.5 Summary

The objective of most statistical investigations is to make an inference about a population parameter. Since we often base inferences upon information contained in a sample from the target population, it is essential that the sample be properly selected. A procedure for obtaining a random sample using statistical software (SPSS) was described in this chapter.

After the sample has been selected, we compute a statistic that contains information about the target parameter. The sampling distribution of the statistic, characterizes the relative frequency distribution of values of the statistic over an, infinitely large number of samples.

The Central Limit Theorem provides information about the sampling distribution of the sample mean, $\bar{x}$. In particular, if you have used random sampling, the sampling distribution of $\bar{x}$ will be approximately normal if the sample size is sufficiently large.

6.6 Exercises

6.1 Use command Select cases of SPSS/PC to obtain 30 random samples of size n = 5 from "population" of 4,171 number of children ever born from Appendix A.

a. Calculate $\bar{x}$ for each of the 30 samples. Construct a relative frequency distribution for the 30 sample means. Compare with the population relative frequency distribution shown in Table 6.1.
b. Compute the average of the 30 sample means.
c. Compute the standard deviation of the 3o sample means.

6.2 Repeat parts a, b, and c of Exercise 7.1, using random samples of size n = 10. Compare relative frequency distribution with that of Exercise 7.1. Do the values of $\bar{x}$ generated from samples of size n = 10 tend cluster more closely about μ?

6.3 Suppose a random sample of n measurements is selected from a population with mean μ = 60 and variance σ^2 =100. For each of the following values of n, give the mean and standard deviation of the sampling distribution of the sample means, $\bar{x}$:

a. n = 10 b. n *= 25* c. n = 50

d. n = 75 e. n = 100 f. n = 500

6.4 A random sample of n = 225 observations is selected from a population with μ = 70 and σ=30. Calculate each of the following probabilities:

a. $P(\bar{x} > 72.5)$ b. $P(\bar{x} < 73.6)$

c. $P(69.1 < \bar{x} < 74.0)$ d. $P(\bar{x} < 65.5)$

6.5 This part year, an elementary school began using a new method to teach arithmetic to first graders. A standardized test, administered at the end of the year, was used to measure the effectiveness of the new method. The relative frequency distribution of the test scores in past years had a mean of 75 and a standard deviation of 10. Consider the standardized test scores for a random sample of 36 first graders taught by the new method.

a. If the relative frequency distribution of test scores for first graders taught by the new method is no different from that of the old method, describe the sampling distribution of $\bar{x}$, the mean test score for random sample of 36 first graders.

b. If the sample mean test score was computed to be $\bar{x}$ = 79, what would you conclude about the effectiveness of the new method of teaching arithmetic? (Hint: Calculate $P(\bar{x} \geq 79)$ using the sampling distribution described in part a.)

Chapter 7 Estimation

CONTENTS

7.1 Introduction

In preceding chapters we learned that populations are characterized by numerical descriptive measures (parameters), and that inferences about parameter values are based on statistics computed from the information in a sample selected from the population of interest. In this chapter, we will demonstrate how to estimate population means, proportions, or variances, and how to estimate the difference between two population means or proportions. We will also be able to assess the reliability of our estimates, based on knowledge of the sampling distributions of the statistics being used.

Example 7.1 **Suppose we are interested in estimating the average number of children ever born to all 4,171 women in the VNDHS 1998 in Appendix A. Although we already know the value of the population mean, this example will be continued to illustrate the concepts involved in estimation. How could one estimate the parameter of interest in this situation?**

Solution **An intuitively appealing estimate of a population mean, μ, is the sample mean, $\bar{x}$, computed from a random sample of *n* observations from the target population. Assume, for example, that we obtain a random sample of size *n* = 30 from numbers of children ever born in Appendix A, and then compute the value of the sample mean to be $\bar{x}$ =*3.05* children. This value of $\bar{x}$ provides a** *point estimate* **of the population mean.**

Definition 7.1

A ***point estimate*** of a parameter is a statistic, a single value computed from the observations in a sample that is used to estimate the value of the target parameter.

How reliable is a point estimate for a parameter? In order to be truly practical and meaningful, an inference concerning a parameter must consist more than just a point estimate; that is, we need to be able to state how close our estimate is likely to be to the true value of the population. This can be done by using the characteristics of the sampling distribution of the statistic that was used to obtain the point estimate; the procedure will be illustrated in the next section.

7.2 Estimation of a population mean: Large-sample case

Recall from Section 6.4 that, for sufficient large sample size, the sampling distribution of the sample mean, $\bar{x}$*,* is approximately normal, as shown in Figure 7.1.

Example 7.2 **Suppose we plan to take a sample of *n* = 30 measurements from population of numbers of children ever born in Appendix A and construct interval**

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

where σ is the population standard deviation of the 4,171 numbers of children ever born and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ is the standard deviation of the sampling distribution of $\bar{x}$ (often called the *standard error* of $\bar{x}$*.)* In other word, we will construct an interval 1.96 standard deviations around the sample mean $\bar{x}$. What can we say about how likely is it is that this interval will contain the true value of the population mean, μ?

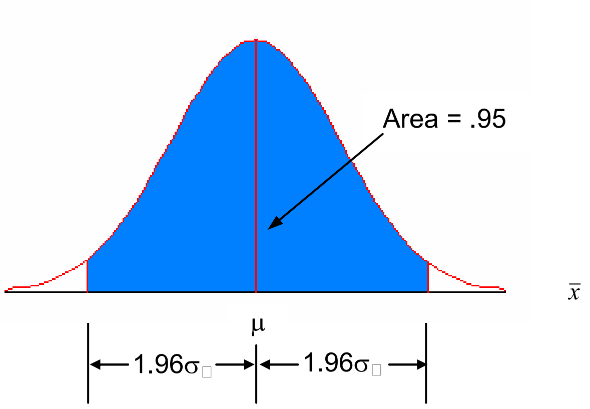


Figure 7.1 *Sample distribution of* $\bar{x}$

Solution **We arrive at a solution by the following three-step process:**

Step 1 First note that, the area beneath the sampling distribution of $\bar{x}$ between $\mu - 1.96\sigma_{\bar{x}}$ and $\mu + 1.96\sigma_{\bar{x}}$ is approximately .95. (This area colored green in

Figure 7.1, is obtained from Table 1 of Appendix C.) This applies that before the sample of measurements is drawn, the probability that $\bar{x}$ will fall within the interval $\mu \pm 1.96\sigma_{\bar{x}}$.

Step 2 If in fact the sample yields a value of $\bar{x}$ that falls within the interval $\mu \pm 1.96\sigma_{\bar{x}}$, then it is true that $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will contain μ, as demonstrated in Figure 7.2. For particular value of $\bar{x}$ that falls within the interval $\mu \pm 1.96\sigma_{\bar{x}}$, a distance of $1.96\sigma_{\bar{x}}$ is marked off both to the left and to the right of $\bar{x}$. You can see that the value of μ must fall within $\bar{x} \pm 1.96\sigma_{\bar{x}}$.

Step 3 Step 1 and Step 2 combined imply that, before the sample is drawn, the probability that the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will enclose μ is approximately .95.

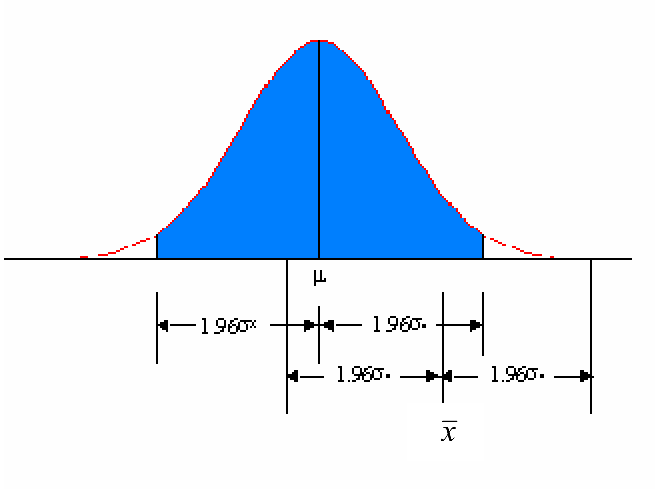


Figure 7.2 *Sample distribution of $\bar{x}$ in Example 7.2*

The interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ in Example 7.2 is called a large-sample 95% **confidence interval** for the population mean μ. The term *large-sample* refers to the sample being of a sufficiently large size that we can apply the Central Limit Theorem to determine the form of the sampling distribution of $\bar{x}$.

Definition 7.2

A ***confidence interval*** for a parameter is an interval of numbers within which we expect the true value of the population parameter to be contained. The endpoints of the interval are computed based on sample information.

Example 7.3 **Suppose that a random sample of observations from the population of three-year old children heights yield the following sample statistics:**

$\bar{x}$ = 88.62 *cm* and s = 4.09 *cm*

Construct a 95% confidence interval of μ, the population mean height, based on this sample.

Solution **A 95% confidence interval for μ, based on a sample of size = 30, is given by**

$$\bar{x}\pm 1.96\sigma_{\bar{x}}=\bar{x}\pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)=92.67\pm 1.96\left(\frac{\sigma}{\sqrt{30}}\right)$$

In most practical applications, the value of the population deviation σ will be unknown. However, for larger samples (*n* ≥ *30*), the sample standard deviation *s* provides a good approximation to σ, and may be used on the formula for the confidence interval. For this example, we obtain

$$88.62\pm 1.96\left(\frac{\sigma}{\sqrt{30}}\right)=88.62\pm 1.96\left(\frac{4.09}{\sqrt{30}}\right)=88.62\pm 1.46$$

or (87.16, 90.08). Hence, we estimate that the population mean height falls within the interval from 87.16 cm to 90.08 *cm*.

How much confidence do we have that μ, the true population mean height, lies within the interval (87.16, 90.08)? Although we cannot be certain whether the sample interval contain μ (unless we calculate the true value of μ for all 823 observations in Appendix B), we can be reasonably sure that it does. This confidence is based on the interpretation of the confidence interval procedure: If we were to select repeated random samples of size *n* = *30* heights, and from a 1.96 standard deviation interval around $\bar{x}$ for each sample, then approximately 95% of the intervals constructed in this manner would contain μ. Thus, we are 95% confident that the particular interval (89.93, 95.41) contains μ, and this is our measure of the reliability of the point estimate $\bar{x}$.

Example 7.4 **To illustrate the classical interpretation of a confidence interval, we generated 40 random samples, each of size *n* = 30, from the population of heights in Appendix B. For each sample, the sample mean and standard deviation are presented in Table 7.1. We then constructed the 95% confidence interval for μ, using the information from each sample. Interpret the results, which are shown in Table 7.2.**

Table 7.1 *Means and standard deviations for 40 random samples of 30 heights from* Appendix B

Sample	Mean	Standard Deviation	Sample	Mean	Standard Deviation
1	89.53	6.39	21	91.17	5.67
2	90.70	4.64	22	89.47	6.68
3	89.02	5.08	23	88.86	4.63
4	90.45	4.69	24	88.70	5.02
5	89.96	4.85	25	90.13	5.07

6	89.96	5.53	26	91.10	5.27
7	89.81	5.60	27	89.27	4.91
8	90.12	6.70	28	88.85	4.77
9	89.45	3.46	29	89.34	5.68
10	89.00	4.61	30	89.07	4.85
11	89.95	4.48	31	91.17	5.30
12	90.18	6.34	32	90.33	5.60
13	89.15	5.98	33	89.31	5.82
14	90.11	5.86	34	91.05	4.96
15	90.40	4.50	35	88.30	5.48
16	90.04	5.26	36	90.13	6.74
17	88.88	4.29	37	90.33	4.77
18	90.98	4.56	38	86.82	4.82
19	88.44	3.64	39	89.63	6.37
20	89.44	5.05	40	88.00	4.51

Table 7.2 *95% confidence intervals for μ for 40 random samples of 30 heights from Appendix B*

Sample	***LCL***	***UCL***	***Sample***	***LCL***	***UCL***
1	87.24	91.81	21	89.14	93.20
2	89.04	92.36	22	87.07	91.86
3	87.20	90.84	23	87.20	90.52
4	88.77	92.13	24	86.90	90.50
5	88.23	91.69	25	88.31	91.95
6	87.99	91.94	26	89.22	92.99
7	87.81	91.82	27	87.51	91.02
8	87.72	92.51	28	87.14	90.56
9	88.21	90.69	29	87.31	91.37
10	87.35	90.65	30	87.33	90.80
11	88.35	91.56	31	89.27	93.07
12	87.91	92.45	32	88.33	92.33
13	87.01	91.29	33	87.23	91.39
14	88.01	92.21	34	89.27	92.83
15	88.79	92.01	35	86.34	90.26
16	88.16	91.92	36	87.71	92.54
17	87.35	90.41	37	88.62	92.04
18	89.35	92.61	38	85.10	88.55
19	87.14	89.75	39	87.35	91.91
20	87.63	91.25	40	86.39	89.62

(Note: The green intervals don't contain μ = 89.67 cm)

Solution **For the target population of 823 heights, we have obtained the population mean value μ = *89.67 cm*. In the 40 repetitions of the confidence interval procedure described above, note that only two of the intervals (those based on samples 38 and 40, indicated**

by red color) do not contain the value of μ, where the remaining 38 intervals (or 95% of the 40 interval) do contain the true value of μ.

Note that, in actual practice, you would not know the true value of μ and you would not perform this repeated sampling; rather you would select a single random sample and construct the associated 95% confidence interval. The one confidence interval you form may or not contain μ, but you can be fairly sure it does because of your confidence in the statistical procedure, the basis for which was illustrated in this example.

Suppose you want to construct an interval that you believe will contain μ with some degree of confidence other than 95%; in other words, you want to choose a confidence coefficient other than .95.

Definition 7.3

The ***confidence coefficient*** is the proportion of times that a confidence interval encloses the true value of the population parameter if the confidence interval procedure is used repeatedly a very large number of times.

The first step in constructing a confidence interval with any desired confidence coefficient is to notice from Figure 7.1 that, for a 95% confidence interval, the confidence coefficient of 95% is equal to the total area under the sampling distribution (1.00), less .05 of the area, which is divided equally between the two tails of the distribution. Thus, each tail has an area of .025. Second, consider that the tabulated value of z (Table 1 of Appendix C) that cuts off an area of .025 in the right tail of the standard normal distribution is 1.96 (see Figure 7.3). The value z = 1.96 is also the distance, in terms of standard deviation, that $\bar{x}$ is from each endpoint of the 95% confidence interval. By assigning a confidence coefficient other than .95 to a confidence interval, we change the area under the sampling distribution between the endpoint of the interval, which in turn changes the tail area associated with z. Thus, this z-value provides the key to constructing a confidence interval with any desired confidence coefficient.

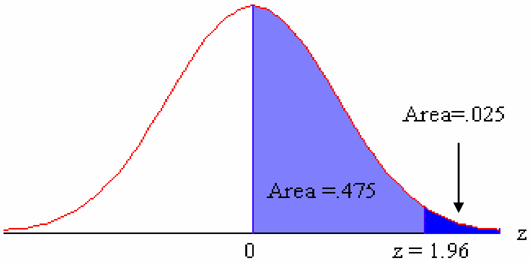


Figure 7.3 *Tabulated z-value corresponding to a tail area of .025*

Definition 7.4

We define $z_{\alpha/2}$ to be the z-value such that an area of $\alpha/2$ lies to its right (see Figure 7.4).

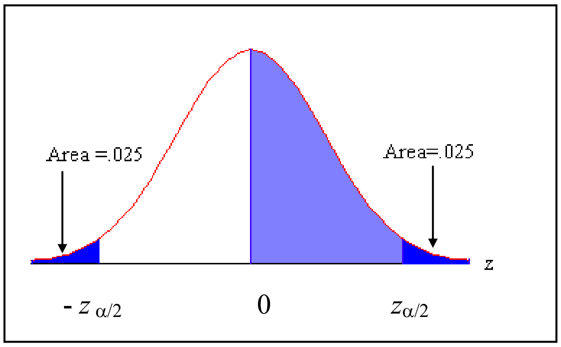


Figure 7.4 *Locating $z_{\alpha/2}$ on the standard normal curve*

Now, if an area of $\alpha/2$ lies beyond $z_{\alpha/2}$ in the right tail of the standard normal (z) distribution, then an area of $z_{\alpha/2}$ lies to the left of $-z_{\alpha/2}$ in the left tail (Figure 7.4) because of the symmetry of the distribution. The remaining area, $(1-\alpha)$, is equal to the confidence coefficient - that is, the probability that $\bar{x}$ falls within $z_{\alpha/2}$ standard deviation of μ is $(1-\alpha)$. Thus, a lager-sample confidence interval for μ, with confidence coefficient equal to $(1-\alpha)$, is given by

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$

Example 7.5 **In statistic problems using confidence interval techniques, a very common confidence coefficient is .90. Determine the value of $z_{\alpha/2}$ that would be used in constructing a 90% confidence interval for a population mean based on a large sample.**

Solution **For a confidence coefficient of .90, we have**

$$\begin{aligned} 1-\alpha &= .90 \\ \alpha &= .10 \\ \alpha/2 &= .05 \end{aligned}$$

and we need to obtain the value $z_{\alpha/2} = z_{.05}$ that locates an area of .05 in the upper tail of the standard normal distribution. Since the total area to the right of 0 is .50, $z_{.05}$ is the value such that the area between 0 and $z_{.05}$ is .50 - .05 = .45. From Table 1 of Appendix C, we find that $z_{.05}$ = 1.645 (see Figure 7.5). We conclude that a large-sample 90% confidence interval for a population mean is given by

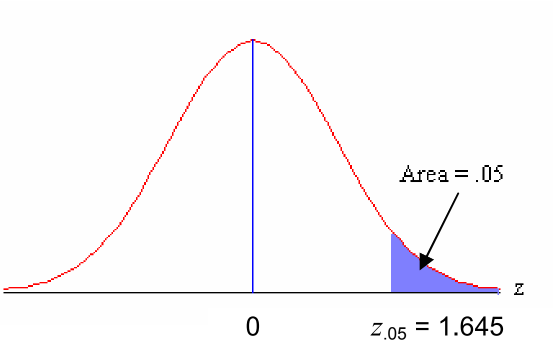
$$\bar{x} \pm 1.645\sigma_{\bar{x}}$$

In Table 7.3, we present the values of $z_{\alpha/2}$ for the most commonly used confidence coefficients.

Table 7.3 *Commonly used confidence coefficient*

Confidence Coefficient $(1-\alpha)$	$\alpha/2$	$z_{\alpha/2}$
.90	.050	1.645
.95	.025	1.960
.98	.010	2.330
.99	.005	2.58

Figure 7.5 *Location of $z_{\alpha/2}$ for Example 7.5*



A summary of the large-sample confidence interval procedure for estimating a population means appears in the next box.

Large-sample $(1-\alpha)$100% confidence interval for a population mean, μ

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the z-value that locates an area of $\alpha/2$ to its right, σ is the standard deviation of the population from which the sample was selected, n is the sample size, and $\bar{x}$ is the value of the sample mean.

Assumption: $n \geq 30$

[When the value of σ is unknown, the sample standard deviation s may be used to approximate σ in the formula for the confidence interval. The approximation is generally quite satisfactory when n $\geq$ 30.]

Example 7.6 **Suppose that in the previous year all graduates at a certain university reported the number of hours spent on their studies during a certain week; the average was 40 hours and the standard deviation was 10 hours. Suppose we want to investigate the problem whether students now are studying more than they used to. This year a random sample of *n* = 50 students is selected. Each student in the sample was interviewed about the number of hours spent on his/her study. This experiment produced the following statistics:**

$\bar{x}$ = 41.5 hours s = 9.2 hours

Estimate μ, the mean number of hours spent on study, using a 99% confidence interval. Interpret the interval in term of the problem.

Solution **The general form of a large-sample 99% confidence interval for μ is**

$$\bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{x} \pm 2.58\left(\frac{s}{\sqrt{n}}\right) = 41.5 \pm 2.58\left(\frac{9.2}{\sqrt{50}}\right) = 41.5 \pm 3.36$$

or (38.14, 44.86).

We can be 99% confident that the interval (38.14, 44.86) encloses the true mean weekly time spent on study this year. Since all the values in the interval fall above 38 hours and below 45 hours, we conclude that there is tendency that students now spend more than 6 hours and less than 7.5 hours per day on average (suppose that they don't study on Sunday).

Example 7.7 **Refer to Example 7.6.**

a. Using the sample information in Example 7.6, construct a 95% confidence interval for mean weekly time spent on study of all students in the university this year.

b. For a fixed sample size, how is the width of the confidence interval related to the confidence coefficient?

Solution

a. The form of a large-sample 95% confidence interval for a population mean μ is

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right) = 41.5 \pm 1.96\left(\frac{9.2}{\sqrt{50}}\right) = 41.5 \pm 2.55$$

or (38.95, 44.05).

b. The 99% confidence interval for μ was determined in Example 7.6 to be (38.14, 44.86). The 95% confidence interval, obtained in this example and based on the same sample information, is narrower than the 99% confidence interval. This relationship holds in general, as stated in the next box.

Relationship between width of confidence interval and confidence coefficient

For a given sample size, the width of the confidence interval for a parameter increases as the confidence coefficient increases. Intuitively, the interval must become wider for us to have greater confidence that it contains the true parameter value.

Example 7.8 **Refer to Example 7.6.**

a. Assume that the given values of the statistic $\bar{x}$ and *s* were based on a sample of size n = 100 instead of a sample size n = 50. Construct a 99% confidence interval for μ, the population mean weekly time spent on study of all students in the university this year.

b. For a fixed confidence coefficient, how is the width of the confidence interval related to the sample size?

Solution

a. Substitution of the values of the sample statistics into the general formula for a 99% confidence interval for μ yield

$$\bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \approx \bar{x} \pm 2.58\left(\frac{s}{\sqrt{n}}\right) = 41.5 \pm 2.58\left(\frac{9.2}{\sqrt{100}}\right) = 41.5 \pm 2.37$$

or (39.13, 43.87)

b. The 99% confidence interval based on a sample of size n = 100, constructed in part a., is narrower than the 99% confidence interval based on a sample of size n = 50, constructed in Example 7.6. This will also hold in general, as stated in the box.

Relationship between width of confidence interval and sample size

For a fixed confidence coefficient, the width of the confidence interval decreases as the sample size increases. That is, larger samples generally provide more information about the target population than do smaller samples.

In this section we introduced the concepts of point estimation of the population mean μ, based on large samples. The general theory appropriate for the estimation of μ also carries over to the estimation of other population parameters. Hence, in subsequent sections we will present only the point estimate, its sampling distribution, the general form of a confidence interval for the parameter of interest, and any assumptions required for the validity of the procedure.

7.3 Estimation of a population mean: small sample case

In the previous section, we discussed the estimation of a population mean based on large samples ($n \geq 30$). However, time or cost limitations may often restrict the number of sample observations that may be obtained, so that the estimation procedures of Section 7.2 would not be applicable.

With small samples, the following two problems arise:

1. Since the Central Limit Theorem applies only to large samples, we are not able to assume that the sampling distribution of $\bar{x}$ is approximately normal. For small samples, the sampling distribution of $\bar{x}$ depends on the particular form of the relative frequency distribution of the population being sampled.
2. The sample standard deviation s may not be a satisfactory approximation to the population standard deviation σ if the sample size is small.

Fortunately, we may proceed with estimation techniques based on small samples if we can make the following assumption:

Assumption required for estimating μ based on small samples (n < 30)

The population from which the sample is selected has an approximate normal distribution.

If this assumption is valid, then we may again use $\bar{x}$ as a point estimation for μ, and the general form of a small-sample confidence interval for μ is as shown next box.

Small-sample confidence interval for μ

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

where the distribution of t based on (n - 1) degrees of freedom.

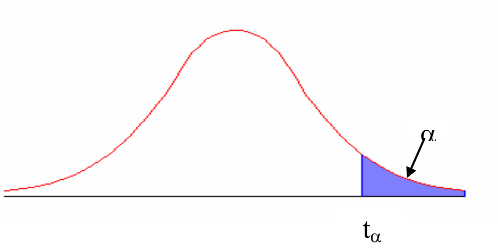
Upon comparing this to the large-sample confidence interval for μ, you will observe that the sample standard deviation s replaces the population standard deviation σ. Also, the sampling distribution upon which the confidence interval is based is known as a Student's t-distribution**.**

Consequently, we must replace the value of $z_{\alpha/2}$ used in a large-sample confidence interval by a value obtained from the t-distribution.

The t-distribution is very much like the z-distribution. In particular, both are symmetric, bell-shaped, and have a mean of 0. However, the distribution of t depends on a quantity called its degrees of freedom (df), which is equal to (n - *1*) when estimating a population mean based on a small sample of size n. Intuitively, we can think of the number of degrees of freedom as the amount of information available for estimating, in addition to μ, the unknown quantity σ^2. Table 2 of Appendix C, a portion of which is reproduced in Table 7.4, gives the value of t_α that located an area of α in the upper tail of the t-distribution for various values of α and for degrees of freedom ranging from 1 to 120.

Table 7.6

Some values for Student's t-distribution



Degrees of freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.102	3.852	4.221
14	1.345	1.760	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073

Example 7.9 **Using Table 7.4 to determine the t-value that would be used in constructing a 95% confidence interval for μ based on a sample of size n = 14.**

Solution For confidence coefficient of .95, we have

$$1-\alpha = .95$$

$$\alpha = .05$$

$$\alpha/2 = .025$$

We require the value of $t_{.025}$ for a t-distribution based on (n - 1) = (14 - 1) = 13 degrees of freedom. In Table 7.4, at intersection of the column labeled $t_{.025}$ and the row corresponding to df = 13, we find the entry 2.160 (see Figure 7.6). Hence, a 95% confidence interval for μ, based on a sample of size n = 13 observations, would be given by

$$\bar{x} \pm 2.160\left(\frac{s}{\sqrt{14}}\right)$$

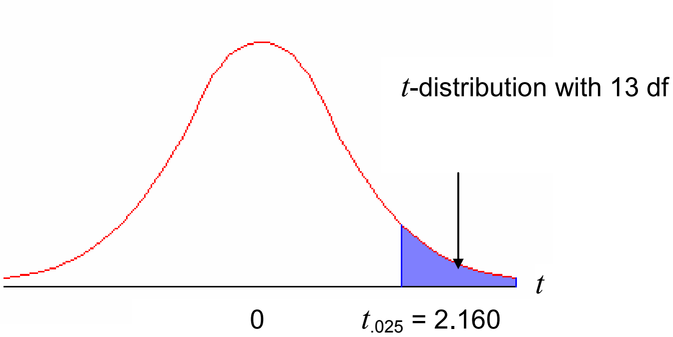


Figure 7.6 *Location of $t_{.025}$ for Example 7.9*

At this point, the reasoning for the arbitrary cutoff point of n = 30 for distinguishing between large and small samples may be better understood. Observe that the values in the last row of Table 2 in Appendix C (corresponding to df = ∞) are the values from the standard normal z-distribution. This phenomenon occurs because, as the sample size increases, the t distribution becomes more like the z distribution. By the time n reaches 30, i.e., df = 29, there is very little difference between tabulated values of t and z.

Before concluding this section, we will comment on the assumption that the sampled population is normally distributed. In the real world, we rarely know whether a sampled population has an exact normal distribution. However, empirical studies indicate that moderates departures from this assumption do not seriously affect the confidence coefficients for small-sample confidence intervals. As a consequence, the definition of the small-sample confidence given in this section interval is frequently used by experimenters when estimating the population mean of a non-normal distribution as long as the distribution is bell-shaped and only moderately skewed.

7.4 Estimation of a population proportion

We will consider now the method for estimating the binomial proportion of successes, that is, the proportion of elements in a population that have a certain characteristic. For example, a demographer may be interested in the proportion of a city residents who are married; a physician may be interested in the proportion of men who are smokers. How would you estimate a binomial proportion p based on information contained in a sample from a population.

Example 7.10 **A commission on crime is interested in estimation the proportion of crimes to firearms in an area with one of the highest crime rates in a country. The commission selects a random sample of 300 files of recently committed crimes in the area and determines that a firearm was reportedly used in 180 of them. Estimate the true**

proportion *p* of all crimes committed in the area in which some type of firearm was reportedly used.

Solution **A logical candidate for a point estimate of the population proportion p is the proportion of observations in the sample that have the characteristic of interest (called a "success"); we will call this sample proportion $\hat{p}$ (read "p hat"). In this example, the sample proportion of crimes related to firearms is given by**

$$\hat{p} = \frac{\text{Number of crimes in sample in which a firearm was reportedly used}}{\text{Total number of crimes in sapmle}} = 180/300 = .60$$

That is, 60% of the crimes in the sample were related to firearms; the value $\hat{p} = .60$ servers as our point estimate of the population proportion p.

To assess the reliability of the point estimate $\hat{p}$, we need to know its sampling distribution. This information may be derived by an application of the Central Limit Theorem. Properties of the sampling distribution of $\hat{p}$ are given in the next box.

Sampling distribution of $\hat{p}$

For sufficiently large samples, the sampling distribution of $\hat{p}$ is approximately normal, with

Mean: $\mu_{\hat{p}} = p$

and *Standard deviation:* $\sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}$

where q = q- p.

A large-sample confidence interval for p may be constructed by using a procedure analogous to that used for estimating a population mean.

Large-sample $(1-\alpha)$100% confidence interval for a population proportion, p

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} \approx \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p}$ is the sample proportion of observations with the characteristic of interest, and $\hat{q} = 1 - \hat{p}$.

Note that, we must substitute $\hat{p}$ and $\hat{q}$ into the formula for $\sigma_{\hat{p}} = \sqrt{pq/n}$ in order to construct the confidence interval. This approximation will be valid as long as the sample size n is sufficiently large.

Example 7.11 **Refer to Example 7.10. Construct a 95% confidence interval for p, the population proportion of crimes committed in the area in which some type of firearm is reportedly used.**

Solution **For a confidence interval of .95, we have** $1-\alpha = .95$**;** $\alpha = .05$**;** $\alpha/2 = .025$**; and the required** z**-value is** $z_{.025}$ **= 1.96. In Example 7.10, we obtained** $\hat{p}=180/300=.60$**. Thus,** $\hat{q}=1-\hat{p}=1-.60=.40$**. Substitution of these values into the formula for an approximate confidence interval for** p **yields**

$$\hat{p}\pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}=.60\pm 1.96\sqrt{\frac{(.60)(.40)}{300}}=.60\pm .06$$

or (.54, .66). Note that the approximation is valid since the interval does not contain 0 or 1.

We are 95% confident that the interval from .54 to .66 contains the true proportion of crimes committed in the area that are related to firearms. That is, in repeated construction of 95% confidence intervals, 95% of all samples would produce confidence interval that enclose p.

It should be noted that small-sample procedure are available for the estimation of a population proportion *p*. We will not discuss details here, however, because most surveys in actual practice use samples that are large enough to employ the procedure of this section.

7.5 Estimation of the difference between two population means: Independent samples

In Section 7.2, we learned how to estimate the parameter μ based on a large sample from a single population. We now proceed to a technique for using the information in two samples to estimate the difference between two population means. For example, we may want to compare the mean heights of the children in province No.18 and in province No.14 using the observations in Appendix B. The technique to be presented is a straightforward extension of that used for large-sample estimation of a single population mean.

Example 7.12 **To estimate the difference between the mean heights for all children of province No. 18 and province No. 14 use the following information**

1. A random sample of 30 heights of children in province No. 18 produced a sample mean of 91.72 *cm* and a standard deviation of 4.50 *cm*.
2. A random sample of 40 heights of children in province No. 14 produced a sample mean of 86.67 *cm* and a standard deviation of 3.88 *cm*.

Calculate a point estimate for the difference between heights of children in two provinces.

Solution **We will let subscript 1 refer to province No. 18 and the subscript 2 to province No. 14. We will also define the following notation:**

μ_1 = Population mean height of all children of province No. 18.

μ_2 = Population mean height of all children of province No. 14.

Similarly, lets $\bar{x}_1$ and $\bar{x}_2$ denote the respective means; s_1 and s_2, the respective sample standard deviations; and n_1 and n_2, the respective sample sizes. The given information may be summarized as in Table 7.5.

Table 7.5 *Summary information for Example 7.12*

	Province No. 18	Province No. 14
Sample size	n_1 = 30	n_2 = 40

Sample mean	$\bar{x}_1$ = 91.72 *cm*	$\bar{x}_2$ = 86.67 *cm*
Sample standard deviation	s_1 = 4.50 *cm*	s_2 = 3.88 *cm*

To estimate $(\mu_1 - \mu_2)$, it seems sensible to use the difference between the sample means $(\bar{x}_1 - \bar{x}_2)$ = (91.72 - 86.67) = 5.05 as our point estimate of the difference between two population means. The properties of the point estimate $(\bar{x}_1 - \bar{x}_2)$ are summarized by its sampling distribution shown in Figure 7.8.

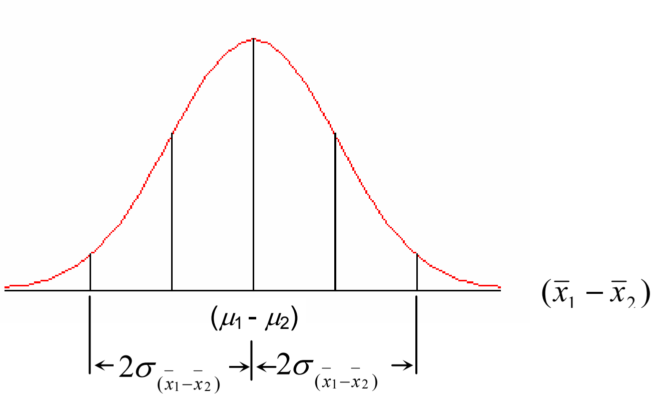

$(\bar{x}_1 - \bar{x}_2)$
$(\mu_1 - \mu_2)$
$\leftarrow 2\sigma_{(\bar{x}_1 - \bar{x}_2)} \rightarrow$
$\leftarrow 2\sigma_{(\bar{x}_1 - \bar{x}_2)} \rightarrow$

Figure 7.8 *Sampling distribution of* $(\bar{x}_1 - \bar{x}_2)$

Sampling distribution of $(\bar{x}_1 - \bar{x}_2)$

For sufficiently large sample size (*n₁* and *n₂* ≥ *30*), the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$, based on independent random samples from two population, is approximately normal with

Mean: $\mu_{(\bar{x}_1 - \bar{x}_2)} = (\mu_1 - \mu_2)$

Standard deviation: $\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

where σ_1^2 and σ_2^2 are standard deviations of two population from which the samples were selected.

As was the case with large-sample estimation of single population mean, the requirement of large sample size enables us to apply the Central Limit Theorem to obtain the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$; it also suffices use to s_1^2 and s_2^2 as approximation to the respective population variances, σ_1^2 and σ_2^2.

The procedure for forming a large-sample confidence interval for $(\mu_1 - \mu_2)$ appears in the accompanying box.

Large-sample (1 - α)100% confidence interval for $(\mu_1 - \mu_2)$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{(\bar{x}_1 - \bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\approx (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(*Note:* We have used the sample variances s_1^2 and s_2^2 as approximations to the corresponding population parameters.)

The assumptions upon which the above procedure is based are the following:

Assumptions required for large-sample estimation of $(\mu_1 - \mu_2)$

1. The two random samples are selected in an independent manner from the target populations. That is the choice of elements in one sample does not affect, and is not affected by, the choice of elements in the other sample.
2. The sample sizes n_1 and n_2 are sufficiently large. (at least 30)

Example 7.13 **Refer to Example 7.12. Construct a 95% confidence interval for $(\mu_1 - \mu_2)$, the difference between mean heights of all children in province No. 18 and province No. 14. Interpret the interval.**

Solution **The general form of a 95% confidence interval for $(\mu_1 - \mu_2)$ based on large samples from the target populations, is given by**

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Recall that $z_{.025}$ = 1.96 and use the information in Table 7.5 to make the following substitutions to obtain the desired confidence interval:

$$(91.72 - 86.67) \pm 1.96\sqrt{\frac{\sigma_1^2}{30} + \frac{\sigma_2^2}{40}}$$

$$\approx (91.72 - 86.67) \pm 1.96\sqrt{\frac{(4.50)^2}{30} + \frac{(3.88)^2}{40}}$$

$$\approx 5.05 \pm 2.01$$

or (3.04, 7.06).

The use of this method of estimation produces confidence intervals that will enclose $(\mu_1 - \mu_2)$, the difference between population means, 95% of the time. Hence, we can be reasonably sure that the mean height of children in province No. 18 was between 3.04 cm and 7.06 cm higher than the mean height of children in province No. 14 at the survey time.

When estimating the difference between two population means, based on small samples from each population, we must make specific assumptions about the relative frequency distributions of the two populations, as indicated in the box.

Assumptions required for small-sample estimation of $(\mu_1 - \mu_2)$

1. Both of the populations which the samples are selected have relative frequency distributions that are approximately normal.
2. The variances σ_1^2 and σ_2^2 of the two populations are equal.
3. The random samples are selected in an independent manner from two populations.

When these assumptions are satisfied, we may use the procedure specified in the next box to construct a confidence interval for $(\mu_1 - \mu_2)$, based on small samples (n_1 and $n_2 < 30$) from respective populations.

Small-sample *(1 - α)*100% confidence interval for $(\mu_1 - \mu_2)$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and the value of $t_{\alpha/2}$ is based on (n_1 + n_2 - 2) degrees of freedom.

Since we assume that the two populations have equal variances (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$), we construct an estimate of σ^2 based on the information contained in both samples. This pooled estimate is denoted by s_p^2 and is computed as in the previous box.

7.6 Estimation of the difference between two population means: Matched pairs

The procedure for estimating the difference between two population means presented in Section 7.5 were based on the assumption that the samples were randomly selected from the target populations. Sometimes we can obtain more information about the difference between population means $(\mu_1 - \mu_2)$, by selecting **paired observations.**

For example, suppose we want to compare two methods for teaching reading skills to first graders using sample of ten students with each method. The best method of sampling would be to match the first graders in pairs according to IQ and other factors that might affect reading achievement. For each pair, one member would be randomly selected to be taught by method 1; the other member would be assigned to class taught by method 2. Then the differences between matched pairs of achievement test scores should provide a clearer picture of the difference in achievement for the two reading methods because the matching would tend to cancel the effects of the factors that formed the basic of the matching.

In the following boxes, we give the assumptions required and the procedure to be used for estimating the difference between two population means based on matched-pairs data.

Assumptions required for estimation of $(\mu_1 - \mu_2)$: Matched pairs

1. The sample paired observations are randomly selected from the target population of paired observations.
2. The population of paired differences is normally distributed.

Small-sample $(1-\alpha)$ 100% confidence interval for $\mu_d = (\mu_1 - \mu_2)$

Let d_1, d_2, . . . d_n represent the differences between the pair-wise observations in a random sample of n matched pairs. Then the small-sample confidence interval for $\mu_d = (\mu_1 - \mu_2)$ is

$$\bar{d} \pm t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)$$

where $\bar{d}$ is the mean of n sample differences, s_d is their standard deviation, and $t_{\alpha/2}$ is based on $(n\text{-}1)$ degrees of freedom.

Example 7.14 **Suppose that the n = 10 pairs of achievement test scores were given in Table 7.7 . Find a 95% confidence interval for the difference in mean achievement, $\mu_d = (\mu_1 - \mu_2)$.**

Table 7.7 *Reading achievement test scores for Example 7.14*

	Student pair									
	1	2	3	4	5	6	7	8	9	10
Method 1 score	78	63	72	89	91	49	68	76	85	55
Method 2 score	71	44	61	84	74	51	55	60	77	39
Pair difference	7	19	11	5	17	-2	13	16	8	16

Solution **The differences between matched pairs of reading achievement test scores are computed as**

d = (method 1 score - method 2 score)

The mean, variance, and standard deviation of the differences are

$$\bar{d} = \frac{\sum d}{n} = \frac{110}{10} = 11.0$$

$$s_d^2 = \frac{\sum d^2 - \frac{\left(\sum d\right)^2}{n}}{n-1} = \frac{1{,}594 - \frac{(110)^2}{10}}{9} = \frac{1{,}594 - 1{,}210}{9} = 42.6667$$

$$s_d = \sqrt{42.67} = 6.53$$

The value of $t_{.025}$, based on (n -1) = 9 degrees of freedom, is given in Table 2 of Appendix C as $t_{.025}$ = 2.262. Substituting these values into the formula for the confidence interval, we obtain

$$\bar{d} \pm t_{.025}\left(\frac{s_d}{\sqrt{n}}\right)$$

$$= 11.0 \pm 2.262\left(\frac{6.53}{\sqrt{10}}\right) = 11.0 \pm 4.7$$

or (6.3, 15.7).

We estimate, with 95% confidence that the difference between mean reading achievement test scores for method 1 and 2 falls within the interval from 6.3 to 15.7. Since all the values within the interval are positive. method 1 seems to produce a mean achievement test score that substantially higher than the mean score for method 2.

7.7 Estimation of the difference between two population proportions

This section extends the method of Section 7.4 to the case in which we want to estimate the difference between two population proportions. For example, we may be interested in comparing the proportions of married and unmarried persons who are overweight.

Example 7.15 **Suppose that there were two surveys, one was carried out in 1990 and another in 1998. In both surveys, random samples of 1,400 adults in a country were asked whether they were satisfied with their life. The results of the surveys are reported in Table 7.8. Construct a point estimate for difference between the proportions of adults in the country in 1990 and in 1998 who were satisfied with their life.**

Table 7.8 *Proportions of two samples for Example 7.15*

	1990	***1998***
Number surveyed	n_1 = 1,400	n_2 = 1,400
Number in sample who said they were satisfied with their life	462	674

Solution **We define some notations:**

p_1 = Population proportion of adults who said that they were satisfied with their life in 1990.

p_2 = Population proportion of adults who said that they were satisfied with their life in 1998.

As a point estimate of (p_1 - p_2), we will use the difference between the corresponding sample proportions, $(\hat{p}_1 - \hat{p}_2)$, where

$$\hat{p}_1 = \frac{Number\ of\ adults\ \ in\ 1990\ who\ said\ that\ they\ were\ satisfied\ with\ their\ life}{Number\ of\ adults\ surveyed\ in\ 1990} = \frac{462}{1{,}400} = .33$$

and

$$\hat{p}_2 = \frac{Number\ of\ adults\ in\ 1998\ who\ said\ that\ they\ were\ satisfied\ with\ their\ life}{Number\ of\ adults\ surveyed\ in\ 1998} = \frac{674}{1{,}400} = .48$$

Thus, the point estimate of $(p_1 - p_2)$, is

$(\hat{p}_1 - \hat{p}_2) = .33 - .48 = -.15$

To judge the reliability of the point estimate $(\hat{p}_1 - \hat{p}_2)$, we need to know the characteristics of its performance in repeated independent sampling from two populations. This information is provided by the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$, shown in the next box.

Sampling distribution of $(\hat{p}_1 - \hat{p}_2)$

For sufficiently large sample size, n_1 and n_2, the sample distribution of $(\hat{p}_1 - \hat{p}_2)$, based on independent random samples from two populations, is approximately normal with

Mean: $\mu_{(\hat{p}_1 - \hat{p}_2)} = (\hat{p}_1 - \hat{p}_2)$

and

Standard deviation: $\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

It follows that a large-sample confidence interval for $(\hat{p}_1 - \hat{p}_2)$ may be obtained as shown in the box.

Large-sample $(1-\alpha)$ 100% confidence interval for $(\hat{p}_1 - \hat{p}_2)$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of observations with the characteristics of interest.

Assumption: The samples are sufficiently large so that the approximation is valid. As a general rule of thumb we will require that intervals

$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}}$ and $\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}}$ do not contain 0 or 1.

Example 7.16 **Refer to Example 7.15. Estimate the difference between the proportions of the adults in this country in 1990 and in 1998 who said that they were satisfied with their life, using a 95% confidence interval.**

Solution **From Example 7.15, we have $n_1 = n_2$ = 1,400, $\hat{p}_1$= .33 and $\hat{p}_2$ = .48.**

Thus, $\hat{q}_1 = 1 - .33 = .67$ and $\hat{q}_2 = 1 - .48 = .52$. Note that the intervals

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}} = .33 \pm 2\sqrt{\frac{(.33)(.67)}{1{,}400}} = .33 \pm .025$$

$$\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_2}} = .48 \pm 2\sqrt{\frac{(.48)(.67)}{1{,}400}} = .48 \pm .027$$

do not contain 0 and 1. Thus, we can apply the large-sample confidence interval for $(p_1 - p_2)$.

The 95% confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{.025}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = (.33 - .48) \pm 1.96\sqrt{\frac{(.33)(.67)}{1{,}400} + \frac{(.48)(.52)}{1{,}400}}$$

$$= -.15 \pm .036$$

or (-.186, -.114). Thus we estimate that the interval (-.186, -.114) enclose the difference $(p_1 - p_2)$ with 95% confidence. It appears that there were between 11.4% and 18.6% more adults in 1998 than in 1990 who said that they were satisfied with their life.

7.8 Choosing the sample size

Before constructing a confidence interval for a parameter of interest, we will have to decide on the number *n* of observations to be included in a sample. Should we sample n = 10 observations, n = 20, or n = 100? To answer this question we need to decide how wide a confidence interval we are willing to tolerate and measure of confidence - that is, the confidence coefficient- that we wish to place in it. The following example will illustrate the method for determining the appropriate sample size for estimating a population mean.

Example 7.17 **A mail-order house wants to estimate the mean length of time between shipment of an order and receipt by customer. The management plans to randomly sample n orders and determine, by telephone, the number of days between shipment and receipt for each order. If management wants to estimate the mean shipping time correct to within .5 day with probability equal to .95, how many orders should be sample?**

Solution **We will use $\bar{x}$, the sample mean of the *n* measurements, to estimate μ, the mean shipping time. Its sampling distribution will be approximately normal and the probability that $\bar{x}$ will lie within**

$$1.96\sigma_{\bar{x}} = 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

of the mean shipping time, μ, is approximately .95 (see Figure 7.9). Therefore, we want to choose the sample size n so that $1.96\sigma/\sqrt{n}$ equals .5 day:

$$1.96\left(\frac{\sigma}{\sqrt{n}}\right)=.5$$

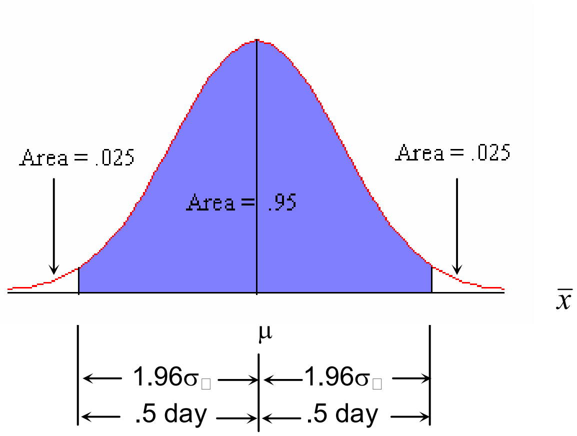


Figure 7.9 *Sampling distribution of the sample mean,* $\bar{x}$

To solve the equation $1.96\sigma/\sqrt{n}$ = .5, we need to know that value of σ , a measure of variation of the population of all shipping times. Since σ is unknown, we must approximate its value using the standard deviation of some previous sample data or deduce an approximate value from other knowledge about the population. Suppose, for example, that we know almost all shipments will delivered within 7 days. Then the population of shipping times might appear as shown in Figure 7.10.

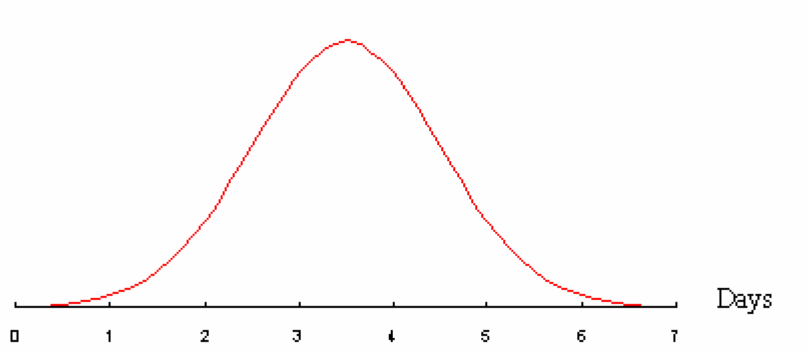


Figure 7.10 *Hypothetical relative frequency distribution of population of shipping times for Example 7.17.*

Figure 7.9 provides the information we need to find an approximation for σ . Since the Empirical Rule tells us that almost all the observations in a data set will fall within the interval $\mu \pm 3\sigma$, it follows that the range of a population is approximately 6σ . If the range of population of shipping times is 7 days, then

6σ = 7 days

and σ is approximately equal to 7/6 or 1.17 days.

The final step in determining the sample size is to substitute this approximate value of σ into the equation obtained previously and solve for *n*.

Thus, we have

$$1.96\left(\frac{1.17}{\sqrt{n}}\right)=.5 \text{ or } \sqrt{n}=\frac{1.96(1.17)}{.5}=4.59$$

Squaring both sides of this equation yields: n = 21.07.

we will follows the usual convention of rounding the calculated sample size upward. Therefore, the mail-order house needs to sample approximately n = 22 shipping times in order to estimate the mean shipping time correct to within .5 day with probability equal .95.

In Example 7.17, we wanted our sample estimate to lie within .5 day of the true mean shipping time, μ, with probability .95, where .95 represents the confidence coefficient. We could calculate the sample size for a confidence coefficient other than .95 by changing the z-value in the equation. In general, if we want $\bar{x}$ to lie within a distance *d* of μ with probability $(1-\alpha)$, we would solve for *n* in the equation

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)=d$$

where the value of $z_{\alpha/2}$ is obtained from Table 1 of Appendix C. The solution is given by

$$n=\left(\frac{z_{\alpha/2}\sigma}{d}\right)^2$$

For example, for a confidence coefficient of .90, we would require a sample size of

$$n=\left(\frac{1.64\sigma}{d}\right)^2$$

Choosing the sample size for estimating a population mean μ to within d units with probability $(1-\alpha)$

$$n=\left(\frac{z_{\alpha/2}\sigma}{d}\right)^2$$

(*Note*: The population standard deviation σ will usually have to be approximated.)

The procedures for determining the sample sizes needed to estimate a population proportion, the difference between two population means, or the difference between two population proportions are analogous to the procedure for the determining the sample size for estimating a population mean.

Choosing the sample size for estimating a population proportion p to within d units with probability $(1-\alpha)$

$$n = \left(\frac{z_{\alpha/2}\sigma}{d} \right)^2 pq$$

where *p* is the value of the population proportion that we are attempting to estimate and q = 1 - p.

(*Note*: This technique requires previous estimates of p and q. If none are available, use p = q = .5 for a conservative choice of n.)

7.9 Estimation of a population variance

In the previous sections, we considered interval estimates for population means or proportions. In this optional section, we discuss a confidence interval for a population variance, σ^2. Intuitively, it seems reasonable to use the sample variance s^2 to estimate σ^2 and to construct our confidence interval around this value. However, unlike sample means and sample proportions, the sampling distribution of the sample variances does not possess a normal z-distribution or a t-distribution.

Rather, when certain assumptions are satisfied, the sampling distribution of s^2 possesses approximately a **chi-square (χ^2) distribution**. The chi-square probability distribution, like the t-distribution, is characterized by a quantity called the degrees of freedom associated with the distribution. Several chi-square probability distributions with different degrees of freedom are shown in Figure 7.11. Unlike z- and t-distributions, the chi-square distribution is not symmetric about 0.

Throughout this section we will use the words *chi-square* and the Greek symbol χ^2 interchangeably.

Example 7.18 **Tabulated values of the χ^2 distribution are given in Table 3 of Appendix C; a partial reproduction of this table is shown in Table 7.9. Entries in the table give an upper-tail value of χ^2, call it χ^2_α , such that $P(\chi^2 > \chi^2_\alpha) = \alpha$. Find the tabulated value of χ^2 corresponding to 9 degrees of freedom that cuts off an upper-tail area of .05.**

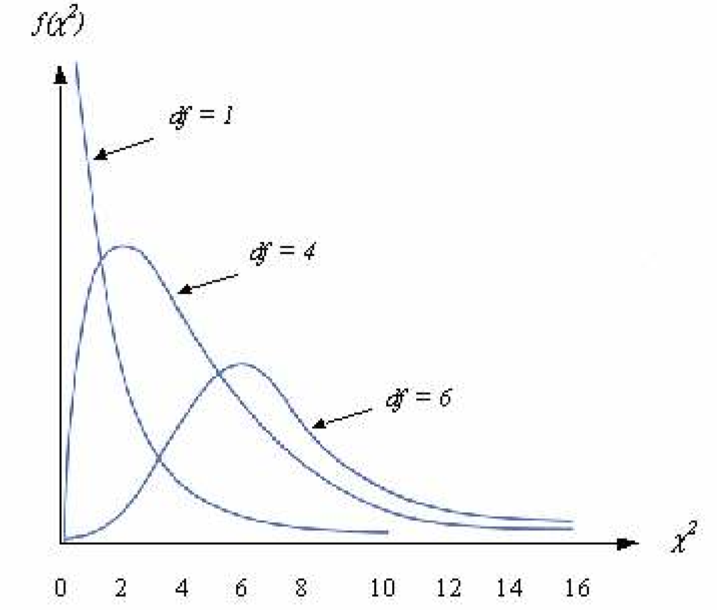


Figure 7.11 *Several chi-square probability distribution*

Solution **The value of χ^2 that we seek appears (shaded) in the partial reproduction of Table 3 of Appendix C given in Table 7.9. The columns of the table identify the value of α associated with the tabulated value of χ^2_α and the rows correspond to the degrees of freedom. For this example, we have df = 9 and α = .05. Thus, the tabulated value of χ^2 corresponding to 9 degrees of freedom is**

$$\chi^2_{.05} = 16.9190$$

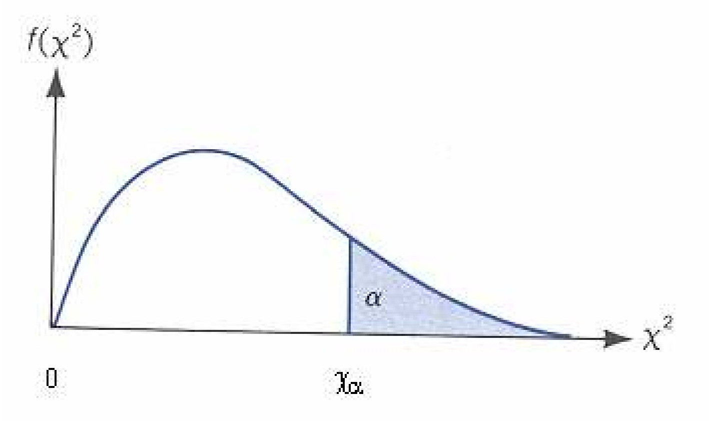


Table 7.9 *Reproduction of part of Table 3 of Appendix C*

We use the tabulated values of χ^2 to construct a confidence interval for σ^2 as the next example.

Example 7.19 **There was a study of contaminated fish in a river. Suppose it is important for the study to know how stable the weights of the contaminated fish are. That is, how large is the variance σ^2 in the fish weights? The 144 samples of fish in the study produced the following summary statistics:**

$$\bar{x} = 1{,}049.7 \text{ grams}, s = 376.6 \text{ grams}.$$

Use this information to construct a 95% confidence interval for the true variation in weights of contaminated fish in the river.

Degrees of freedom		$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.59660
3	6.25139	7.81473	9.34840	11.34490	12.83810
4	7.77944	9.48773	11.14330	13.27670	14.86020
5	9.23635	11.07050	12.83250	15.08630	16.74960
6	10.64460	12.59160	14.44940	16.81190	18.54760
7	12.01700	14.06710	16.01280	18.47530	20.27770
8	13.36160	15.50730	17.53460	20.09020	21.95500

9	14.68370	16.91900	19.02280	21.66600	23.58930
10	15.98710	18.30700	20.48310	23.20930	25.18820
11	17.27500	19.67510	21.92000	24.72500	26.75690
12	18.54940	21.02610	23.33670	26.21700	28.29950
13	19.81190	22.36210	24.73560	27.68830	29.81940
14	21.06420	23.68480	26.11900	29.14130	31.31930
15	22.30720	24.99580	27.48840	30.57790	32.80130
16	23.54180	26.29620	28.84540	31.99990	34.26720
17	24.76900	27.58710	30.19100	33.40870	35.71850
18	25.98940	28.86930	31.52640	34.80530	37.15640
19	27.20360	30.14350	32.85230	36.19080	38.58220

Solution **A (1 - α)100% confidence interval for σ^2 depends on the quantities s^2, (n - 1), and critical values of χ^2 as shown in the box. Note that (n - 1) represents the degrees of freedom associated with the χ^2 distribution. To construct the interval, we first locate the critical values $\chi^2_{1-\alpha/2}$, and $\chi^2_{\alpha/2}$. These are the values of χ^2 that cut off an area of α/2 in the lower and upper tails, respectively, of the chi-square distribution (see Figure 7.11).**

A (1 - α)100% confidence interval for a population variance, σ^2

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

where $\chi^2_{1-\alpha/2}$, and $\chi^2_{\alpha/2}$ are values of χ^2 that locate an area of α/2 to the right and α/2 to the left, respectively, of a chi-square distribution based on (n - 1) degrees of freedom.

Assumption: The population from which the sample is selected has an approximate normal distribution.

For a 95% confidence interval, (1 - α) = .95 and α/2 = .05/2 = .025. There- fore, we need the tabulated values $\chi^2_{.025}$, and $\chi^2_{.975}$ for (n - 1) = 143 df. Looking in the df = 150 row of Table 3 of Appendix C (the row with the df values closest to 143), we find $\chi^2_{.025}$ = 185.800 and $\chi^2_{.975}$ = 117.985. Substituting into the formula given in the box, we obtain

$$\frac{(144-1)(376.6)^2}{185.800} \le \sigma^2 \le \frac{(144-1)(376.6)^2}{117.985}$$

We are 95% confident that the true variance in weights of contaminated fish in the river falls between 109,156.8 and 171,898.4.

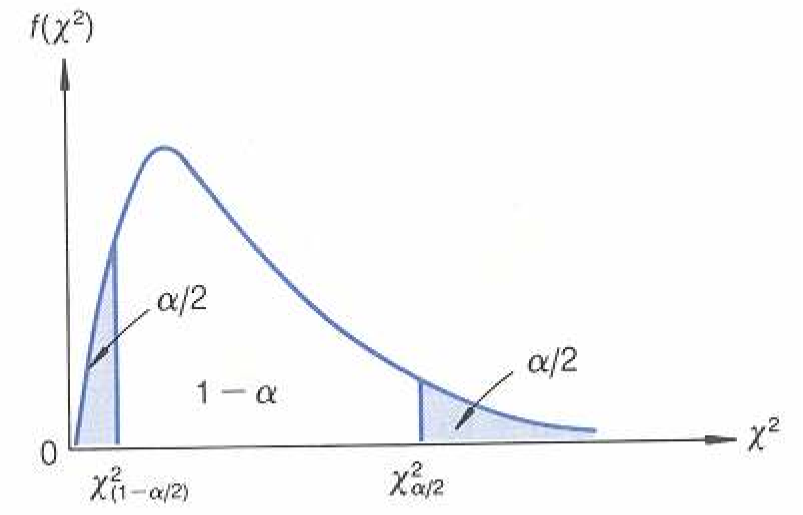


Figure 7.11 *The location of* $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ for a chi-square distribution

Example 7.20 **Refer to Example 7.19. Find a 95% confidence interval for σ, the true standard deviation of the fish weights.**

Solution **A confidence interval for σ is obtained by taking the square roots of the lower and upper endpoints of a confidence interval for σ^2. Thus, the 95% confidence interval is**

$$\sqrt{109{,}156.8} \le \sigma^2 \le \sqrt{171{,}898.4}$$
$$330.4 \le \sigma^2 \le 414.6$$

Thus, we are 95% confident that the true standard deviation of the fish weights is between 330.4 grams and 414.6 grams.

Note that the procedure for calculating a confidence interval for σ^2 in Example 7.19 (and the confidence interval for a in Example 7.20) requires an assumption regardless of whether the sample size n is large or small (see box). We must assume that the population from which the sample is selected has an approximate normal distribution. It is reasonable to expect this assumption to be satisfied in Examples 7.19 and 7.20 since the histogram of the 144 fish weights in the sample is approximately normal.

7.10 Summary

This chapter presented the technique of estimation - that is, using sample information to make an inference about the value of a population parameter, or the difference between two population parameters. In each instance, we presented the point estimate of the parameter of interest, its sampling distribution, the general form of a confidence interval, and any assumptions required for the validity of the procedure. In addition, we provided techniques for determining the sample size necessary to estimate each of these parameters.

7.11 Exercises

7.1. Use Table 1 of Appendix C to determine the value of $z_{\alpha/2}$ that would be used to construct a large-sample confidence interval for μ, for each of the following confidence coefficients:

a) .85
b) .95

c) .975

7.2. Suppose a random sample of size n = 100 produces a mean of $\bar{x}$ =81 and a standard deviation of s = 12.

a) Construct a 90% confidence interval for μ.
b) Construct a 95% confidence interval for μ.
c) Construct a 99% confidence interval for μ.

7.3. Use Table 2 of Appendix C to determine the values of $t_{\alpha/2}$ that would used in the construction of a confidence interval for a population mean for each of the following combinations of confidence coefficient and sample size:

a) Confidence coefficient .99, n = 18.
b) Confidence coefficient .95, n = 10.
c) Confidence coefficient .90, n = 15.

7.4. A random sample of n = 10 measurements from a normally distributed population yields $\bar{x}$ = 9.4 and s = 1.8.

a) Calculate a 90% confidence for μ.
b) Calculate a 95% confidence for μ.
c) Calculate a 99% confidence for μ.

7.5. The mean and standard deviation of n measurements randomly sampled from a normally distributed population are 33 and 4, respectively. Construct a 95% confidence interval for μ when:

a) n = 5 b) n = 15 c) n = 25

7.6. Random samples of n measurements are selected from a population with unknown proportion of successes p. Compute an estimate of $\sigma_{\hat{p}}$ for each of the following situations:

a) n = 250, $\hat{p}$ = .4 b) n = 500, $\hat{p}$ = .85 c) n = 95, $\hat{p}$ = .25

7.7. A random sample of size 150 is selected from a population and the number of successes is 60.

a) Find $\hat{p}$.
b) Construct a 90% confidence interval for p.
c) Construct a 95% confidence interval for p.
d) Construct a 99% confidence interval for p.

7.8. Independent random samples from two normal population produced the sample means and variances listed in the following table.

Sample from population 1	Sample from population 2
n_1 = 14	n_2 = 7
$\bar{x}_1$ = 53.2	$\bar{x}_1$ = 43.4
s_1^2 = 96.8	s_2^2 = 102.0

a) Find a 90% confidence interval for (μ_1 - μ_2).

b) Find a 95% confidence interval for (μ_1 - μ_2).
c) Find a 99% confidence interval for (μ_1 - μ_2).

7.9. A random sample of ten paired observations yielded the following summary information:

$\overline{d}$ = 2.3 s_d = 2.67

a) Find a 90% confidence interval for μ_d.
b) Find a 95% confidence interval for μ_d.
c) Find a 99% confidence interval for μ_d.

Chapter 8 Hypothesis Testing

CONTENTS

8.1 Introduction

In this chapter we will study another method of inference-making: ***hypothesis testing***. The procedures to be discussed are useful in situations where we are interested in making a decision about a parameter value, rather than obtaining an estimate of its value. It is often desirable to know whether some characteristics of a population is larger than a specified value, or whether the obtained value of a given parameter is less than a value hypothesized for the purpose of comparison.

8.2 Formulating Hypotheses

When we set out to test a new theory, we first formulate a ***hypothesis***, or a claim, which we believe to be true. For example, we may claim that the mean number of children born to urban women is less than the mean number of children born to rural women.

Since the value of the population characteristic is unknown, the information provided by a sample from the population is used to answer the question of whether or not the population quantity is larger than the specified or hypothesized value. In statistical terms, a statistical *hypothesis* is a statement about the value of a population parameter. The hypothesis that we try to establish is called the ***alternative hypothesis*** and is denoted by H_a. To be paired with the alternative hypothesis is the ***null hypothesis***, which is "opposite" of the alternative hypothesis, and is denoted by H_0. In this way, the null and alternative hypotheses, both stated in terms of the appropriate parameters, describe two possible states of nature that cannot simultaneously be true. When a researcher begins to collect information about the phenomenon of interest, he or she generally tries to present evidence that lends support to the alternative hypothesis. As you will subsequently learn, we take an indirect approach to obtaining support for the alternative hypothesis: Instead of trying to show that the alternative hypothesis is true, we attempt to produce evidence to show that the null hypothesis is false.

It should be stressed that researchers frequently put forward a null hypothesis in the hope that they can discredit it. For example, consider an educational researcher who designed a new way to teach a particular concept in science, and wanted to test experimentally whether this new method worked better than the existing method. The researcher would design an experiment comparing the two methods. Since the null hypothesis would be that there is no difference between the two methods, the researcher would be hoping to reject the null hypothesis and conclude that the method he or she developed is the better of the two.

The null hypothesis is typically a hypothesis of no difference, as in the above example where it is the hypothesis that there is no difference between population means. That is why the word "null" in "null hypothesis" is used – it is the hypothesis of no difference.

Example 8.1 **Formulate appropriate null and alternative hypotheses for testing the demographer's theory that the mean number of children born to urban women is less than the mean number of children born to rural women.**

Solution **The hypotheses must be stated in terms of a population parameter or parameters. We will thus define**

μ_1 = Mean number of children born to urban women, and

μ_2 = Mean number of children ever born of the rural women.

The demographer wants to support the claim that μ_1 is less than μ_2; therefore, the null and alternative hypotheses, in terms of these parameters, are

H_0: $(\mu_1 - \mu_2) = 0$ (i.e., $\mu_1 = \mu_2$; there is no difference between the mean numbers of children born to urban and rural women)

H_a: $(\mu_1 - \mu_2) < 0$ (i.e., $\mu_1 < \mu_2$; the mean number of children born to urban women is less than that for the rural women)

Example 8.2 **For many years, cigarette advertisements have been required to carry the following statement: "Cigarette smoking is dangerous to your health." But, this waning is often located in inconspicuous corners of the advertisements and printed in small type. Consequently, a researcher believes that over 80% of those who read cigarette advertisements fail to see the warning. Specify the null and alternative hypotheses that would be used in testing the researcher's theory.**

Solution **The researcher wants to make an inference about p, the true proportion of all readers of cigarette advertisements who fail to see the warning. In particular, he wishes to collect evidence to support the claim that p is greater than .80; thus, the null and alternative hypotheses are**

H_0: $p = .80$

H_a: $p > .80$

Observe that the statement of H_0 in these examples and in general, is written with an equality (=) sign. In Example 8.2, you may have been tempted to write the null hypothesis as H_0: $p \leq .80$. However, since the alternative of interest is that $p > .80$, then any evidence that would cause you to reject the null hypothesis H_0: $p = .80$ in favor of H_a: $p > .80$ would also cause you to reject H_0: $p = p'$, for any value of p' that is less than .80. In other words, H_0: $p = .80$ represents the worst possible case, from the researcher's point of view, when the alternative hypothesis is not correct. Thus, for mathematical ease, we combine all possible situations for describing the opposite of H_a into one statement involving equality.

Example 8.3 **A metal lathe is checked periodically by quality control inspectors to determine if it is producing machine bearings with a mean diameter of .5 inch. If the mean diameter of the bearings is larger or smaller than .5 inch, then the process is out of control and needs to be adjusted. Formulate the null and alternative hypotheses that could be used to test whether the bearing production process is out of control.**

Solution **We define the following parameter:**

μ = True mean diameter (in inches) of all bearings produced by the lathe

If either $\mu > .5$ or $\mu < .5$, then the metal lathe's production process is out of control. Since we wish to be able to detect either possibility, the null and alternative hypotheses would be

H_0: μ = .5 (i.e., the process is in control)

H_a: $\mu \neq$ .5 (i.e., the process is out of control)

An alternative hypothesis may hypothesize a change from H_0 in a particular direction, or it may merely hypothesize a change without specifying a direction. In Examples 8.1 and 8.2, the researcher is interested in detecting departure from H_0 in one particular direction. In Example 8.1, the interest focuses on whether the mean number of children born to the urban women is less than the mean number of children born to rural women. The interest focuses on whether the proportion of cigarette advertisement readers who fail to see the warning is greater than .80 in Example 8.2. These two tests are called one-tailed tests. In contrast, Example 8.3 illustrates a two-tailed test in which we are interested in whether the mean diameter of the machine bearings differs in either direction from .5 inch, i.e., whether the process is out of control.

8.3 Types of errors for a Hypothesis Test

The goal of any hypothesis testing is to make a decision. In particular, we will decide whether to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_a. Although we would like always to be able to make a correct decision, we must remember that the decision will be based on sample information, and thus we are subject to make one of two types of error, as defined in the accompanying boxes.

Definition 8.1

A *Type I error* is the error of rejecting the null hypothesis when it is true. The probability of committing a Type I error is usually denoted by α.

Definition 8.2

A *Type II error* is the error of accepting the null hypothesis when it is false. The probability of making a Type II error is usually denoted by β.

The null hypothesis can be either true or false further, we will make a conclusion either to reject or not to reject the null hypothesis. Thus, there are four possible situations that may arise in testing a hypothesis (see Table 8.1).

Table 8.1 *Conclusions and consequences for testing a hypothesis*

		Conclusions	
		Do not reject Null Hypothesis	*Reject Null Hypothesis*
True "State of Nature"	*Null Hypothesis*	Correct conclusion	Type I error
	Alternative Hypothesis	Type II error	Correct conclusion

The kind of error that can be made depends on the actual state of affairs (which, of course, is unknown to the investigator). Note that we risk a Type I error only if the null hypothesis is rejected, and we risk a Type II error only if the null hypothesis is not rejected. Thus, we may make no error, or we may make either a Type I error (with probability α), or a Type II error (with probability β), but not both. We don't know which type of error corresponds to actuality and so would like to keep the probabilities of both types of errors small. There is an intuitively appealing relationship between the probabilities for the two types of error: As α increases, β

decreases, similarly, as β increases, a decreases. The only way to reduce α and β simultaneously is to increase the amount of information available in the sample, i.e., to increase the sample size.

Example 8.4 **Refer to Example 8.3. Specify what Type I and Type II errors would represent, in terms of the problem.**

Solution **A Type I error is the error of incorrectly rejecting the null hypothesis. In our example, this would occur if we conclude that the process is out of control when in fact the process is in control, i.e., if we conclude that the mean bearing diameter is different from .5 inch, when in fact the mean is equal to .5 inch. The consequence of making such an error would be that unnecessary time and effort would be expended to repair the metal lathe.**

A Type II error that of accepting the null hypothesis when it is false, would occur if we conclude that the mean bearing diameter is equal to .5 inch when in fact the mean differs from .5 inch. The practical significance of making a Type II error is that the metal lathe would not be repaired, when in fact the process is out of control.

The probability of making a Type I error (α) can be controlled by the researcher (how to do this will be explained in Section 8.4). α is often used as a measure of the reliability of the conclusion and called the *level of significance* (or *significance level*) for a hypothesis test.

You may note that we have carefully avoided stating a decision in terms of "accept the null hypothesis H_0." Instead, if the sample does not provide enough evidence to support the alternative hypothesis H_a, we prefer a decision "not to reject H_0." This is because, if we were to "accept H_0," the reliability of the conclusion would be measured by β, the probability of Type II error. However, the value of β is not constant, but depends on the specific alternative value of the parameter and is difficult to compute in most testing situations.

In summary, we recommend the following procedure for formulating hypotheses and stating conclusions.

Formulating hypotheses and stating conclusions

1. State the hypothesis as the alternative hypothesis H_a.
2. The null hypothesis, H_0, will be the opposite of H_a and will contain an equality sign.
3. If the sample evidence supports the alternative hypothesis, the null hypothesis will be rejected and the probability of having made an incorrect decision (when in fact H_0 is true) is α, a quantity that can be manipulated to be as small as the researcher wishes.
4. If the sample does not provide sufficient evidence to support the alternative hypothesis, then conclude that the null hypothesis cannot be rejected on the basis of your sample. In this situation, you may wish to collect more information about the phenomenon under study.

Example 8.5 **The logic used in hypothesis testing has often been likened to that used in the courtroom in which a defendant is on trial for committing a crime.**

a. Formulate appropriate null and alternative hypotheses for judging the guilt or innocence of the defendant.

b. Interpret the Type I and Type II errors in this context.

c. If you were the defendant, would you want α to be small or large? Explain.

Solution

a. Under a judicial system, a defendant is "innocent until proven guilty." That is, the burden of proof is not on the defendant to prove his or her innocence; rather, the court must collect sufficient evidence to support the claim that the defendant is guilty. Thus, the null and alternative hypotheses would be

H_0: Defendant is innocent

H_a: Defendant is guilty

b. The four possible outcomes are shown in Table 8.2. A Type I error would be to conclude that the defendant is guilty, when in fact he or she is innocent; a Type II error would be to conclude that the defendant is innocent, when in fact he or she is guilty.

Table 8.2 *Conclusions and consequences inn Example 8.5*

		Decision of Court	
		Defendant is innocent	Defendant is guilty
True State of Nature	Defendant is innocent	*Correct decision*	*Type II error*
	Defendant is guilty	Type I error	Correct decision

c. Most would probably agree that the Type I error in this situation is by far the more serious. Thus, we would want α, the probability of committing a Type I error, to be very small indeed.

A convention that is generally observed when formulating the null and alternative hypotheses of any statistical test is to state H_0 so that the possible error of incorrectly rejecting H_0 (Type I error) is considered more serious than the possible error of incorrectly failing to reject H_0 (Type II error). In many cases, the decision as to which type of error is more serious is admittedly not as clear-cut as that of Example 8.5; experience will help to minimize this potential difficulty.

8.4 Rejection Regions

In this section we will describe how to arrive at a decision in a hypothesis-testing situation. Recall that when making any type of statistical inference (of which hypothesis testing is a special case), we collect information by obtaining a random sample from the populations of interest. In all our applications, we will assume that the appropriate sampling process has already been carried out.

Example 8.6 **Suppose we want to test the hypotheses**

H_0: $\mu = 72$

H_a: $\mu > 72$

What is the general format for carrying out a statistical test of hypothesis?

Solution **The first step is to obtain a random sample from the population of interest. The information provided by this sample, in the form of a sample statistic, will help us decide whether to reject the null hypothesis. The sample statistic upon which we base our decision is called the *test statistic*.**

The second step is to determine a test statistic that is reasonable in the context of a given hypothesis test. For this example, we are hypothesizing about the value of the population mean μ. Since our best guess about the value of μ is the sample mean $\bar{x}$ (see Section 7.2), it seems reasonable to use $\bar{x}$ as a test statistic. We will learn how to choose the test statistic for other hypothesis-testing situations in the examples that follow.

The third step is to specify the range of possible computed values of the test statistic for which the null hypothesis will be rejected. That is, what specific values of the test statistic will lead us to reject the null hypothesis in favor of the alternative hypothesis? These specific values are known collectively as the *rejection region* for the test. For this example, we would need to specify the values of $\bar{x}$ that would lead us to believe that H_a is true, i.e., that μ is greater than 72. We will learn how to find an appropriate rejection region in later examples.

Once the rejection region has been specified, the fourth step is to use the data in the sample to compute the value of the test statistic. Finally, we make our decision by observing whether the computed value of the test statistic lies within the rejection region. If in fact the computed value falls within the rejection region, we will reject the null hypothesis; otherwise, we do not reject the null hypothesis.

An outline of the hypothesis-testing procedure developed in Example 8.6 is given followings.

Outline for testing a hypothesis

1. Obtain a random sample from the population(s) of interest.
2. Determine a test statistic that is reasonable in the context of the given hypothesis test.
3. Specify the rejection region, the range of possible computed values of the test statistic for which the null hypothesis will be rejected.
4. Use the data in the sample to compute the value of the test statistic.
5. Observe whether the computed value of the test statistic lies within the rejection region. If so, reject the null hypothesis; otherwise, do not reject the null hypothesis.

Recall that the null and alternative hypotheses will be stated in terms of specific population parameters. Thus, in step 2 we decide on a test statistic that will provide information about the target parameter.

Example 8.7 **Refer to Example 8.1, in which we wish to test**

H_0: $(\mu_1 - \mu_2) = 0$

H_a: $(\mu_1 - \mu_2) < 0$

where μ_1, and μ_2, are the population mean numbers of children born to urban women and rural women, respectively. Suggest an appropriate test statistic in the context of this problem.

Solution **The parameter of interest is $(\mu_1 - \mu_2)$, the difference between the two population means. Therefore, we will use $(\bar{x}_1 - \bar{x}_2)$, the difference between the corresponding sample means, as a basis for deciding whether to reject H_0. If the difference between the sample means, $(\bar{x}_1 - \bar{x}_2)$, falls greatly below the hypothesized value of $(\mu_1 - \mu_2) = 0$, then we have evidence that disagrees with the null hypothesis. In fact, it would support the alternative hypothesis that $(\mu_1 - \mu_2) < 0$. Again, we are using the point estimate of the target parameter as the test statistic in the hypothesis-testing approach. In general, when the hypothesis test involves a specific population parameter, the test statistic to be used is the conventional point estimate of that parameter.**

In step 3, we divide all possible values of the test into two sets: the *rejection region* and its complement. If the computed value of the test statistic falls within the rejection region, we reject the null hypothesis. If the computed value of the test statistic does not fall within the rejection region, we do not reject the null hypothesis.

Example 8.8 **Refer to Example 8.6. For the hypothesis test**

H_0: μ = 72

H_a: μ > 72

indicate which decision you may make for each of the following values of the test statistic:

a. $\bar{x}=110$ b. $\bar{x}=59$ c. $\bar{x}=73$

Solution

a. If $\bar{x}=110$, then much doubt is cast upon the null hypothesis. In other words, if the null hypothesis were true (i.e., if μ is in fact equal to 72), then it is very unlikely that we would observe a sample mean $\bar{x}$ as large as 110. We would thus tend to reject the null hypothesis on the basis of information contained in this sample.

b. Since the alternative of interest is μ > 72, this value of the sample mean, $\bar{x}=59$, provides no support for H_a. Thus, we would not reject H_0 in favor of H_a: μ > 72, based on this sample.

c. Does a sample value of $\bar{x}=73$ cast sufficient doubt on the null hypothesis to warrant its rejection? Although the sample mean $\bar{x}=73$ is larger than the null hypothesized value of μ =72, is this due to chance variation, or does it provide strong enough evidence to conclude in favor of H_a? We think you will agree that the decision is not as clear-cut as in parts a and b, and that we need a more formal mechanism for deciding what to do in this situation.

We now illustrate how to determine a rejection region that takes into account such factors as the sample size and the maximum probability of a Type I error that you are willing to tolerate.

Example 8.9 **Refer to Example 8.8. Specify completely the form of the rejection region for a test of**

H_0: $\mu = 72$

H_a: $\mu > 72$

at a significance level of α = .05.

Solution **We are interested in detecting a directional departure from H_0; in particular, we are interested in the alternative that *μ is greater than 72*. Now, what values of the sample mean $\bar{x}$ would cause us to reject H_0 in favor of H_a? Clearly, values of $\bar{x}$ which are "sufficiently greater" than 72 would cast doubt on the null hypothesis. But how do we decide whether a value, $\bar{x}=73$ is "sufficiently greater" than 72 to reject H_0? A convenient measure of the distance between $\bar{x}$ and 72 is the z-score, which "standardizes" the value of the test statistic $\bar{x}$:**

$$z=\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}=\frac{\bar{x}-72}{\sigma/\sqrt{n}}\approx\frac{\bar{x}-72}{s/\sqrt{n}}$$

The z-score is obtained by using the values of $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ that would be valid if the null hypothesis were true, i.e., if μ = 72. The z-score then gives us a measure of how many standard deviations the observed $\bar{x}$ is from what we would expect to observe if H_0 were true.

We examine Figure 8.1a and observe that the chance of obtaining a value of $\bar{x}$ more than 1.645 standard deviations above 72 is only .05, when in fact the true value of μ is 72. We are assuming that the sample size is large enough to ensure that the sampling distribution of $\bar{x}$ is approximately normal. Thus, if we observe a sample mean located more than 1.645 standard deviations above 72, then either H_0, is true and a relatively rare (with probability .05 or less) event has occurred, or H_a is true and the population mean exceeds 72. We would tend to favor the latter explanation for obtaining such a large value of $\bar{x}$, and would then reject H_0.

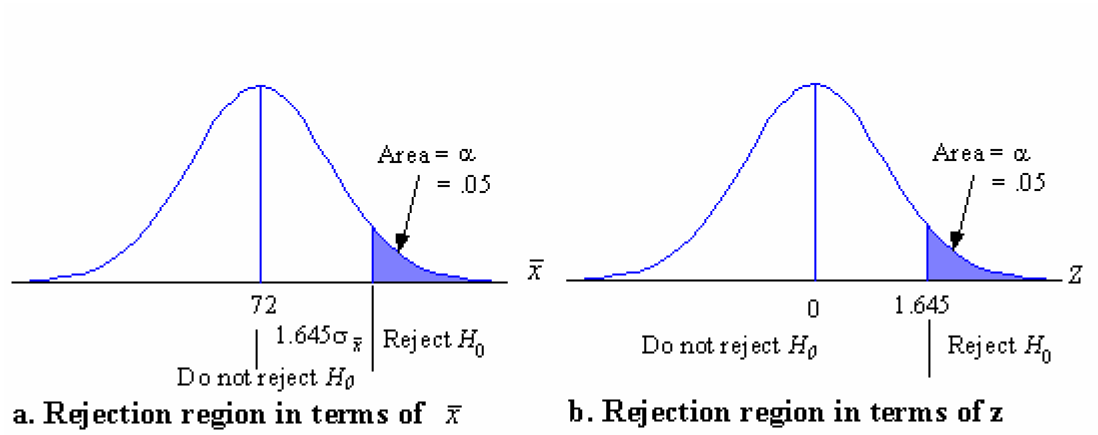


Figure 8.1 *Location of rejection region of Example 8.9*

In summary, our rejection region for this example consists of all values of z that are greater than 1.645 (i.e., all values of $\bar{x}$ that are more than 1.645 standard deviations above 72). The value at the boundary of the rejection region is called the *critical value*. The *critical value* 1.645 is shown in Figure 8.1b. In this situation, the probability of a Type I error – that is, deciding in favor of H_a if in fact H_0 is true – is equal to a α =.05.

Example 8.10 **Specify the form of the rejection region for a test of**

H_0: μ = 72

H_a: μ < 72

at significance level α = .01.

Solution **Here, we want to be able to detect the directional alternative that *μ is less than 72*; in this case, it is "sufficiently small" values of the test statistic $\bar{x}$ that would cast doubt on the null hypothesis. As in Example 8.9, we will standardize the value of the test statistic to obtain a measure of the distance between $\bar{x}$ and the null hypothesized value of 72:**

$$z = \frac{(\bar{x} - \mu_{\bar{x}})}{\sigma_{\bar{x}}} = \frac{\bar{x} - 72}{\sigma / \sqrt{n}} \approx \frac{\bar{x} - 72}{s / \sqrt{n}}$$

This z-value tells us how many standard deviations the observed $\bar{x}$ is from what would be expected if H_0 were true. Here, we have also assumed that $n \geq 30$ so that the sampling distribution of $\bar{x}$ will be approximately normal. The appropriate modifications for small samples will be indicated in Chapter 9.

Figure 8.2a shows us that, when in fact the true value of μ is 72, the chance of observing a value of $\bar{x}$ more than 2.33 standard deviations below 72 is only .01. Thus, at significance level (probability of Type I error) equal to .01, we would reject the null hypothesis for all values of z that are less than - 2.33 (see Figure 8.2b), i.e., for all values of $\bar{x}$ that lie more than 2.33 standard deviations below 72.

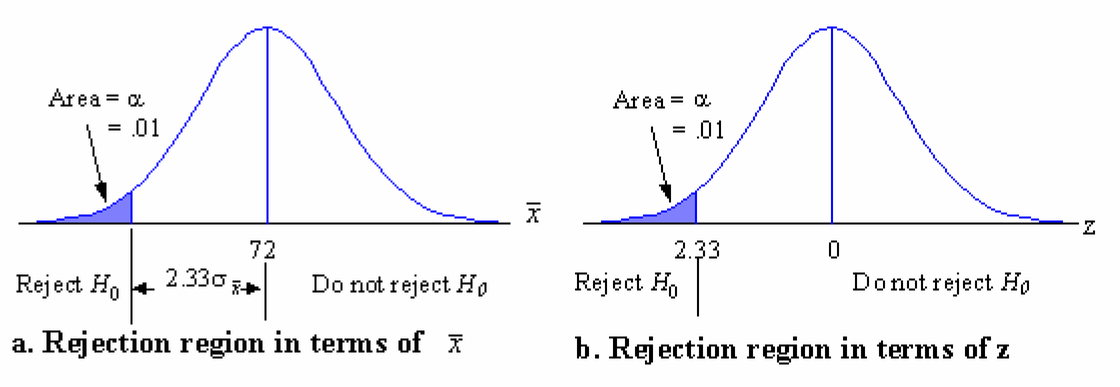


Figure 8.2 *Location of rejection region of Example 8.10*

Example 8.11 **Specify the form of the rejection region for a test of**

H_0: $\mu = 72$

H_a: $\mu \neq 72$

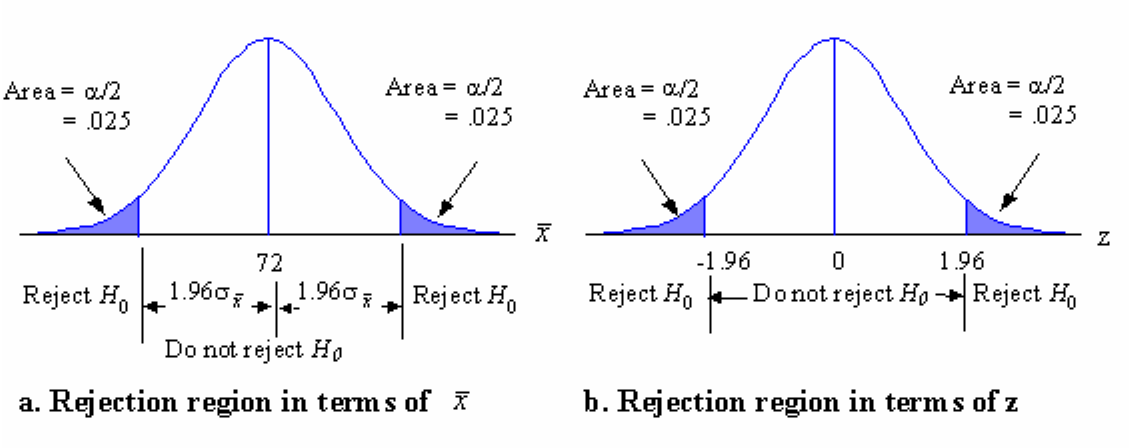
where we are willing to tolerate a .05 chance of making a Type I error.

Solution **For this two-sided (non-directional) alternative, we would reject the null hypothesis for "sufficiently small" or "sufficiently large" values of the standardized test statistic**

$$z \approx \frac{\bar{x}-72}{s/\sqrt{n}}$$

Now, from Figure 8.3a, we note that the chance of observing a sample mean $\bar{x}$ more than 1.96 standard deviations below 72 or more than 1.96 standard deviations above 7 2, when in fact H_0 is true, is only $\alpha = .05$. Thus, the rejection region consists of two sets of values: We will reject H_0 if z is either less than -1.96 or greater than 1.96 (see Figure 8.3b). For this rejection rule, the probability of a Type I error is .05.

The three previous examples all exhibit certain common characteristics regarding the rejection



region, as indicated in the next paragraph.

Figure 8.3 Location of rejection region of Example 8.11

Guidelines for Step 3 of Hypothesis Testing

1. The value of α, the probability of a Type I error; is specified in advance by the researcher. It can be made as small or as large as desired; typical values are α = .01, .02, .05, and .10. For a fixed sample size, the size of the rejection region decreases as the value of a decreases (see Figure 8.4). That is, for smaller values of α, more extreme departures of the test statistic from the null hypothesized parameter value are required to permit rejection of H_0.
2. For testing means or proportions, the test statistic (i.e., the point estimate of the target parameter) is standardized to provide a measure of how great is its departure from the null hypothesized value of the parameter. The standardization is based on the sampling distribution of the point estimate, assuming H_0 is true. (It is through standardization that the rejection rule takes into account the sample sizes.)

$$\text{Standard test statistic} = \frac{\text{Point estimate - Hypothesized value}}{\text{Standard deviation of point estimate}}$$

3. The location of the rejection region depends on whether the test is one-tailed or two-tailed, and on the pre-specified significance level, α.
   a. For a one-tailed test in which the symbol ">" occurs in H_0, the rejection region consists of values in the upper tall of the sampling distribution of the standardized test statistic. The critical value is selected so that the area to its right is equal to α.
   b. For a one-tailed test in which the symbol "<" appears in H_a, the rejection region consists of values in the lower tail of the sampling distribution of the standardized test statistic. The critical value is selected so that the area to its left is equal to α.
   c. For a two-tailed test, in which the symbol "$\neq$" occurs in H_a, the rejection region consists of two sets of values. The critical values are selected so that the area in each tail of the sampling distribution of the standardized test statistic is equal to $\alpha/2$.

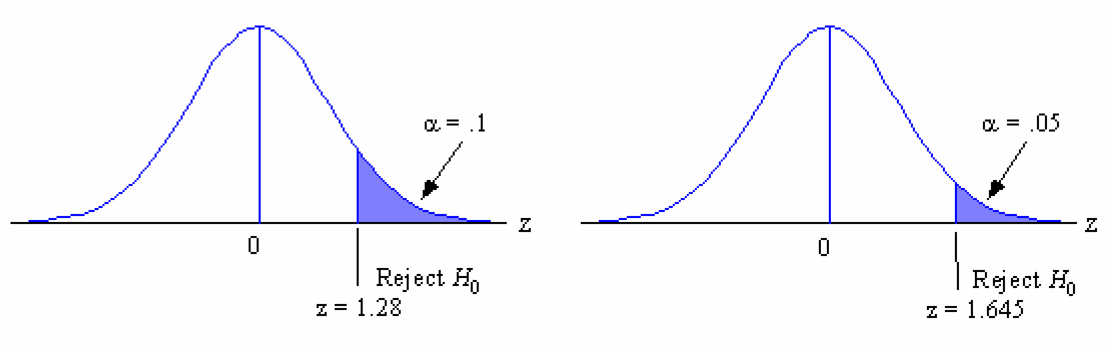


Figure 8.4 Size of the upper-tail rejection region for different values of α

Steps 4 and 5 of the hypothesis-testing approach require the computation of a test statistic from the sample information. Then we determine if the standardized of the test statistic value lies within the rejection region in order to make a decision about whether to reject the null hypothesis.

Example 8.12 **Refer to Example 8.9. Suppose the following statistics were calculated based on a random sample of n = 30 measurements: $\bar{x}$ = 73, s = 13. Perform a test of**

H_0: $\mu = 72$

H_a: $\mu > 72$

at a significance level of $\alpha = .05$.

Solution In Example 8.9, we determined the following rejection rule for the given value of α and the alternative hypothesis of interest:

Reject H_0 if $z > 1.645$.

The standardized test statistic, computed assuming H_0 is true, is given by

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - 72}{\sigma / \sqrt{n}} \approx \frac{\bar{x} - 72}{s / \sqrt{n}} = \frac{73 - 72}{13 / \sqrt{30}} = .42$$

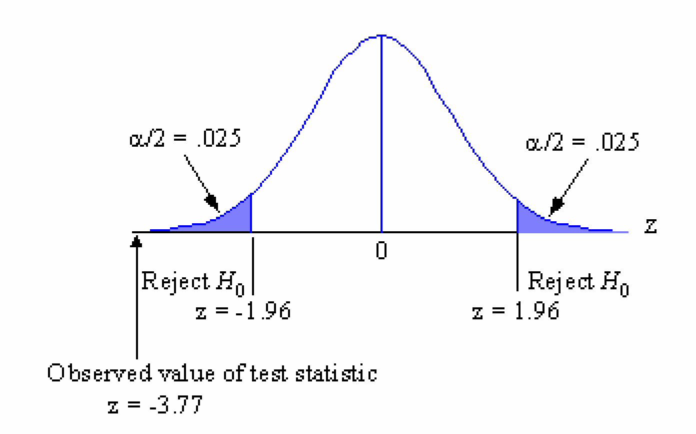


Figure 8.5 *Location of rejection region of Example 8.12*

Since this value does not lie within the rejection region (shown in Figure 8.5), we fail to reject H_0 and conclude there is insufficient evidence to support the alternative hypothesis, H_a: $\mu > 72$. (Note that we do not conclude that H_0 is true; rather, we state that we have insufficient evidence to reject H_0.)

8.5 Summary

In this chapter, we have introduced the logic and general concepts involved in the statistical procedure of hypothesis testing. The techniques will be illustrated more fully with practical applications in Chapter 9.

8.6 Exercises

8.1. A medical researcher would like to determine whether the proportion of males admitted to a hospital because of heart disease differs from the corresponding proportion of females. Formulate the appropriate null and alternative hypotheses and state whether the test is one-tailed or two-tailed.

8.2. Why do we avoid stating a decision in terms of "accept the null hypothesis H_0"?

8.3. Suppose it is desired to test

H_0: $\mu = 65$

H_a: $\mu \neq 65$

at significance level $\alpha = .02$. Specify the form of the rejection region. (Hint: assuming that the sample size will be sufficient to guarantee the approximate normality of the sampling distribution of $\bar{x}$.)

8.4. Indicate the form of the rejection region for a test of

H_0: $(p_1 - p_2) = 0$

H_a: $(p_1 - p_2) > 0$

Assume that the sample size will be appropriate to apply the normal approximation to the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$, and that the maximum tolerable probability of committing a Type I error is .05.

8.5. For each of the following rejection region, determine the value of α, the probability of a Type I error:

a) $z < -1.96$ b) $z > 1.645$ c) $z < -2.58$ or $z > 2.58$

Chapter 9 Applications of Hypothesis Testing

9.1 Introduction

In this chapter, we will present applications of the hypothesis-testing logic developed in Chapter 8. Among the population parameters to be considered are $(\mu_1 - \mu_2)$, p, and $(p_1 - p_2)$.
The concepts of a hypothesis test are the same for all these parameters; the null and alternative hypotheses, test statistic, and rejection region all have the same general form (see Chapter 8). However, the manner in which the test statistic is actually computed depends on the parameter of interest. For example, in Chapter 7 we saw that the large-sample test statistic for testing a hypothesis about a population mean μ is given by

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \text{(see also Example 8.9)}$$

while the test statistic for testing a hypothesis about the parameter p is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

The key to correctly diagnosing a hypothesis test is to determine first the parameter of interests. In this section, we will present several examples illustrating how to determine the parameter of interest. The following are the key words to look for when conducting a hypothesis test about a population parameter.

Determining the parameter of interest

PARAMETER	DESCRIPTION
μ	Mean; average
$(\mu_1 - \mu_2)$	Difference in means or averages; mean difference; comparison of means or averages
p	Proportion; percentage; fraction; rate
$(p_1 - p_2)$	Difference in proportion, percentage, fraction, or rates; comparison of proportions, percentages, fractions, or rates
σ^2	Variance; variation; precision
$\dfrac{\sigma_1^2}{\sigma_2^2}$	Ratio of variances; difference in variation; comparison of variances

In the following sections we will present a summary of the hypothesis-testing procedures for each of the parameters listed in the previous box.

9.2 Hypothesis test about a population mean

Suppose that in the last year all students at a certain university reported the number of hours spent on their studies during a certain week; the average was 40 hours. This year we want to determine whether the mean time spent on studies of all students at the university is in excess of 40 hours per week. That is, we will test

H_0: $\mu = 40$

H_a: $\mu > 40$

where

μ = Mean time spent on studies of all students at the university.

We are conducting this study in an attempt to gather support for H_a; we hope that the sample data will lead to the rejection of H_0. Now, the point estimate of the population mean μ is the sample mean $\bar{x}$. Will the value of $\bar{x}$ that we obtain from our sample be large enough for us to conclude that μ is greater than 40? In order to answer this question, we need to perform each step of the hypothesis-testing procedure developed in Chapter 8.

Tests of population means using large samples

The following box contains the elements of a large-sample hypothesis test about a population mean, μ. Note that for this case, the only assumption required for the validity of the procedure is that the sample size is in fact large ($n \geq 30$).

Large-sample test of hypothesis about a population mean

ONE -TAILED TEST	TWO -TAILED TEST
H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$
H_a: $\mu > \mu_0$ (or H_a: $\mu < \mu_0$)	H_a: $\mu \neq \mu_0$

Test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \approx \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Rejection region:	*Rejection region:*
$z > z_\alpha$ (or $z < - z_\alpha$)	$z < -z_{\alpha/2}$ (or $z > z_{\alpha/2}$)

where z_α is the z-value such that $P(z > z_\alpha) = \alpha$; and $z_{\alpha/2}$ is the z-value such that $P(z > z_{\alpha/2}) = \alpha/2$. [*Note:* μ_0 is our symbol for the particular numerical value specified for μ in the null hypothesis.]

Assumption: The sample size must be sufficiently large (say, $n \geq 30$) so that the sampling distribution of $\bar{x}$ is approximately normal and that s provides a good approximately to σ.

Example 9.1 **The mean time spent on studies of all students at a university last year was 40 hours per week. This year, a random sample of 35 students at the university was drawn. The following summary statistics were computed:**

$\bar{x} = 42.1$ hours; $s = 13.85$ hours

Test the hypothesis that μ, the population mean time spent on studies per week is equal to 40 hours against the alternative that μ is larger than 40 hours. Use a significance level of $\alpha = .05$.

Solution **We have previously formulated the hypotheses as**

H_0: $\mu = 40$

H_a: $\mu > 40$

Note that the sample size n = 35 is sufficiently large so that the sampling distribution of $\bar{x}$ is approximately normal and that s provides a good approximation to σ. Since the required assumption is satisfied, we may proceed with a large-sample test of hypothesis about μ.

Using a significance level of $\alpha = .05$, we will reject the null hypothesis for this one-tailed test if

$z > z_{\alpha/2} = z_{.05}$

i.e., if $z > 1.645$. This rejection region is shown in Figure 9.1.

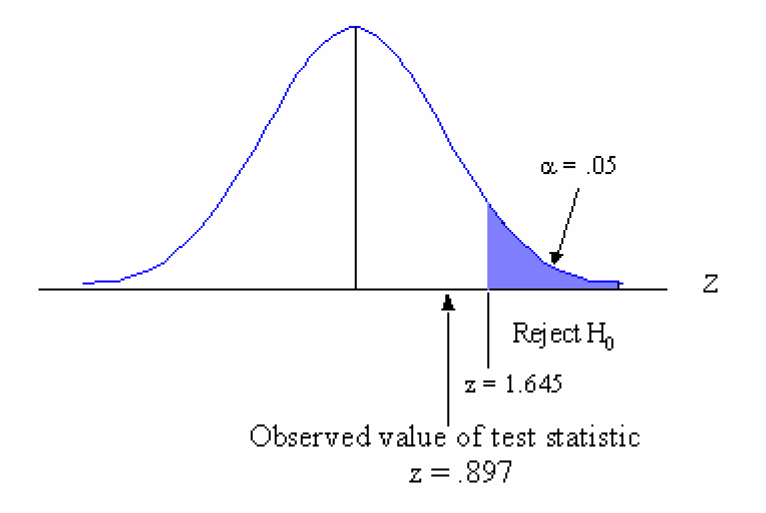


Figure 9.1 *Rejection region for Example 9.1*

Computing the value of the test statistic, we obtain

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{42.1 - 40}{13.85/\sqrt{35}} = .897$$

Since this value does not fall within the rejection region (see Figure 9.1), we do not reject H_0. We say that there is insufficient evidence (at α = .05) to conclude that the mean time spent on studies per week of all students at the university this year is greater than 40 hours. We would need to take a larger sample before we could detect whether μ > 40, if in fact this were the case.

Example 9.2 **A sugar refiner packs sugar into bags weighing, on average 1 kilogram. Now the setting of machine tends to drift i.e. the average weight of bags filled by the machine sometimes increases sometimes decreases. It is important to control the average weight of bags of sugar. The refiner wish to detect shifts in the mean weight of bags as quickly as possible, and reset the machine. In order to detect shifts in the mean weight, he will periodically select 50 bags, weigh them, and calculate the sample mean and standard deviation. The data of a periodical sample as follows:**

$\bar{x} = 1.03\ kg \qquad s = .05\ kg$

Test whether the population mean μ is different from 1 kg at significance level $\alpha = .01$.

Solution **We formulate the following hypotheses:**

H_0: $\mu = 1$

H_a: $\mu \neq 1$

The sample size (50) exceeds 30, we may proceed with the larger sample test about μ. Because shifts in μ in either direction are important, so the test is two-tailed.

At significance level α = .01, we will reject the null hypothesis for this two tail test if

$z < - z_{\alpha/2} = - z_{.005}$ or $z > z_{\alpha/2} = z_{.005}$

i.e., if $z < - 2.576$ or $z > 2.576$.

The value of the test statistic is computed as follows:

$$z \approx \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{1.03 - 1}{.05 / \sqrt{50}} = 4.243$$

Since this value is greater than the upper-tail critical value (2.576), we reject the null hypothesis and accept the alternative hypothesis at the significance level of 1%. We would conclude that the overall mean weight was no longer 1 kg, and would run a less than 1% chance of committing a Type I error.

Example 9.3 **Prior to the institution of a new safety program, the average number of on-the-job accidents per day at a factory was 4.5. To determine if the safety program has been effective in reducing the average number of accidents per day, a random sample of 30 days is taken after the institution of the new safety program and the number of accidents per day is recorded. The sample mean and standard deviation were computed as follows:**

$\bar{x} = 3.7 \qquad s = 1.3$

a. Is there sufficient evidence to conclude (at significance level .01) that the average number of on-the-job accidents per day at the factory has decreased since the institution of the safety program?

b. What is the practical interpretation of the test statistic computed in part a?

Solution

a. In order to determine whether the safety program was effective, we will conduct a large-sample test of

H_0: $\mu = 4.5$ (i.e., no change in average number of on-the-job accidents per day)

H_a: $\mu < 4.5$ (i.e., average number of on-the-job accidents per day has decreased)

where μ represents the average number of on-the-job accidents per day at the factory after institution of the new safety program. For a significance level of $\alpha = .01$, we will reject the null hypotheses if

$z < - z_{.01} = - 2.33$ (see Figure 9.2)

The computed value of the test statistic is

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{3.7 - 4.5}{1.3 / \sqrt{30}} = 3.37$$

Since this value does fall within the rejection region, there is sufficient evidence (at α =.01) to conclude that the average number of on-the-job accidents per day at the factory has decreased since the institution of the safety program. It appears that the safety program was effective in reducing the average number of accidents per day.

b. If the null hypothesis is true, μ = 4.5. Recall that for large samples, the sampling distribution of $\bar{x}$ is approximately normal, with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. Then the z-score for $\bar{x}$, under the assumption that H_0 is true, is given by

$$z = \frac{\bar{x} - 4.5}{\sigma / \sqrt{n}}$$

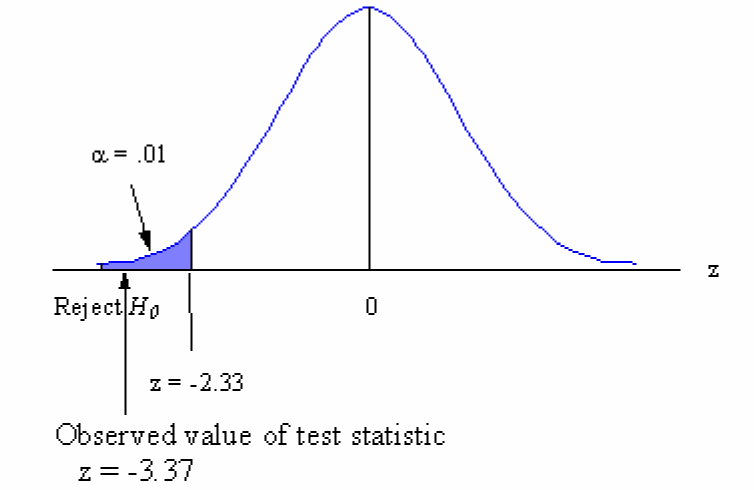


Figure 9.2 *Location of rejection region of Example 9.3*

You can see that the test statistic computed in part a is simply the z-score for the sample mean $\bar{x}$, if in fact μ = 4.5. A calculated z-score of -3.37 indicates that the value of $\bar{x}$ computed from the sample falls a distance of 3.37 standard deviations below the hypothesized mean of μ = 4.5. Of course, we would not expect to observe a z-score this extreme if in fact μ = 4.5.

Tests of population means using small samples

When the assumption required for a large-sample test of hypothesis about μ is violated, we need a hypothesis-testing procedure that is appropriate for use with small samples. Because if we use methods of the large-sample test, we will run into trouble on two accounts. Firstly, our small sample will underestimate the population variance, so our test statistic will be wrong. Secondly, the means of small samples are not normally distributed, so our critical values will be wrong. We have learnt that the means of small samples have a t-distribution, and the appropriate t-distribution will depend on the number of degrees of freedom in estimating the population variance. If we use large samples to test a hypothesis, then the critical values we use will depend upon the type of test (one or two tailed). But if we use small samples, then the critical values will depend upon the degrees of freedom as well as the type of test.

A hypothesis test about a population mean, μ, based on a small sample (n < 30) consists of the elements listed in the accompanying box.

Small-sample test of hypothesis about a population mean

ONE-TAILED TEST	TWO-TAILED TEST
H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$
H_a: $\mu > \mu_0$ (or H_a: $\mu < \mu_0$)	H_a: $\mu \neq \mu_0$

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Rejection region:	*Rejection region:*
$t > t_\alpha$ (or $t < -t_\alpha$)	$t < -t_{\alpha/2}$ (or $t > t_{\alpha/2}$)

where the distribution of t is based on (n – 1) degrees of freedom; t_α is the t-value such that $P(t > t_\alpha) = \alpha$; and $t_{\alpha/2}$ is the t-value such that $P(t > t_{\alpha/2}) = \alpha/2$.

Assumption: The relative frequency distribution of the population from Which the sample was selected is approximately normal.

As we noticed in the development of estimation procedures, when we are making inferences based on small samples, more restrictive assumptions are required than when making inferences from large samples. In particular, this hypothesis test requires the assumption that the population from which the sample is selected is approximately normal.

Notice that the test statistic given in the box is a t statistic and is calculated exactly as our approximation to the large-sample test statistic, z, given earlier in this section. Therefore, just like z, the computed value of t indicates the direction and approximate distance (in units of standard deviations) that the sample mean, $\bar{x}$, is from the hypothesized population mean, μ_0.

Example 9.4 **The expected lifetime of electric light bulbs produced by a given process was 1500 hours. To test a new batch a sample of 10 was taken which showed a mean lifetime of 1410 hours. The standard deviation is 90 hours. Test the hypothesis that the mean lifetime of the electric light bulbs has not changed, using a level of significance of α = .05.**

Solution This question asks us to test that the mean has not changed, so we must employ a two-tailed test:

H_0: $\mu = 1500$

H_a: $\mu \neq 1500$

Since we are restricted to a small sample, we must make the assumption that the lifetimes of the electric light bulbs have a relative frequency distribution that is approximately normal. Under

this assumption, the test statistic will have a
t-distribution with $(n - 1) = (10 - 1) = 9$ degrees of freedom. The rejection rule is then to reject the null hypothesis for values of t such that

$t < - t_{\alpha/2}$ or $t > t_{\alpha/2}$ with $\alpha/2 = .05/2 = .025$.

From Table 7.6 in Chapter 7 (or Table 2 of Appendix C) with 9 degrees of freedom, we find that $t_{.025} = 2.262$.

The value of test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1410 - 1500}{90/\sqrt{10}} = -2.999$$

The computed value of the test statistic, t = - 2.999, falls below the critical value of - 2.262. We reject H_0 and accept H_1 at significance level of .05, and conclude that there is some evidence to suggest that the mean lifetime of all light bulbs has changed.

9.3 Hypothesis tests of population proportions

Tests involving sample proportions are extremely important in practice. Many market researchers express their results in terms of proportions, e.g. "40% of the population clean their teeth with brand A toothpaste ". It will be useful to design tests that will detect changes in proportions. For example, we may want to test the null hypothesis that the true proportion of people who use brand A is equal to .40 (i.e., H_0: p = .40) against the alternative H_a: p > .40.

The procedure described in the next box is used to test a hypothesis about a population proportion, p, based on a large sample from the target population. (Recall that p represents the probability of success in a Bernoulli process.)

In order that the procedure to be valid, the sample size must be sufficiently large to guarantee approximate normality of the sampling distribution of the sample proportion, p. A general rule of thumb for determining whether n is "sufficiently large" is that the interval $\hat{p} \pm 2\sqrt{\hat{p}\hat{q}/n}$ does not include 0 or 1.

Large-sample test of hypothesis about a population proportion

ONE -TAILED TEST	TWO -TAILED TEST
H_0: $p = p_0$	H_0: $p = p_0$
H_a: $p > p_0$ (or H_a: $p < p_0$)	H_a: $p \neq p_0$

Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

Rejection region:	*Rejection region:*
$z > z_\alpha$ (or $z < - z_\alpha$)	$z < -z_{\alpha/2}$ (or $z > z_{\alpha/2}$)
where $q_0 = 1 - p_0$	where $q_0 = 1 - p_0$

Assumption: The interval $\hat{p} \pm 2\sqrt{\hat{p}\hat{q}/n}$ does not contain 0 and 1.

Example 9.5 **Suppose it is claimed that in a very large batch of components, about 10% of items contain some form of defect. It is proposed to check whether this proportion has**

increased, and this will be done by drawing randomly a sample of 150 components. In the sample, 20 are defectives. Does this evidence indicate that the true proportion of defective components is significantly larger than 10%? Test at significance level α = .0 5.

Solution **We wish to perform a large-sample test about a population proportion, p:**

H_0: p = .10 (i.e., no change in proportion of defectives)

H_a: p > .10 (i.e., proportion of defectives has increased)

where p represents the true proportion of defects.

At significance level α = .05, the rejection region for this one-tailed test consists of all values of z for which

$z > z_{.05} = 1.645$

The test statistic requires the calculation of the sample proportion, $\hat{p}$, of defects:

$$\hat{p} = \frac{\text{Number of sampled components}}{\text{Number of defective components in the sample}}$$

$$= \frac{20}{150} = .133$$

Noting that $q_0 = 1 - p_0 = 1 - .10 = .90$, we obtain the following value of the test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{.133 - .10}{\sqrt{(.10)(.90)/150}} = 1.361$$

This value of z lies out of the rejection region; so we would conclude that the proportion defective in the sample is not significant. We have no evidence to reject the null hypothesis that the proportion defective is .01 at the 5% level of significance. The probability of our having made a Type II error (accepting H_0 when, in fact, it is not true) is β = .05.

[Note that the interval

$$\hat{p} \pm 2\sqrt{\hat{p}\hat{q}/n} = .133 \pm 2\sqrt{(.133)(1-.133)/150} = .133 \pm .056$$

does not contain 0 or 1. Thus, the sample size is large enough to guarantee that

validity of the hypothesis test.]

Although small-sample procedures are available for testing hypotheses about a population proportion, the details are omitted from our discussion. It is our experience that they are of limited utility since most surveys of binomial population performed in the reality use samples that are large enough to employ the techniques of this section.

9.4 Hypothesis tests about the difference between two population means

There are two brands of coffee, A and B. Suppose a consumer group wishes to determine whether the mean price per pound of brand A exceeds the mean price per pound of brand B. That is, the consumer group will test the null hypothesis

H_0: $(\mu_1 - \mu_2) = 0$ against the alternative $((\mu_1 - \mu_2) > 0$. The large-sample procedure described in the box is applicable testing a hypothesis about $(\mu_1 - \mu_2)$, the difference between two population means.

Large-sample test of hypothesis about $(\mu_1 - \mu_2)$

ONE -TAILED TEST	TWO -TAILED TEST
H_0: $(\mu_1 - \mu_2) = D_0$	H_0: $(\mu_1 - \mu_2) = D_0$
H_a: $(\mu_1 - \mu_2) > D_0$ (or H_a: $(\mu_1 - \mu_2) < D_0$)	H_a: $(\mu_1 - \mu_2) \neq D_0$

Test statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} \approx \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Rejection region:	*Rejection region:*
$z > z_\alpha$ (or $z < -z_\alpha$)	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

[Note: In many practical applications, we wish to hypothesize that there is no difference between the population means; in such cases, $D_0 = 0$]

Assumptions:

1. The sample sizes n_1 and n_2 are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$).
2. The samples are selected randomly and independent from the target populations.

Example 9.6 **A consumer group selected independent random samples of suppermarkets located throughout a country for the purpose of comparing the retail prices per pound of coffee of brands A and B. The results of the investigation are summarized in Table 9.1. Does this evidence indicate that the mean retail price per pound of brand A coffee is significantly higher than the mean retail price per pound of brand B coffee? Use a significance level of $\alpha = .01$.**

Table 9.1 *Coffee prices for Example 9.6*

Brand A	*Brand B*
$n_1 = 75$	$n_2 = 64$
$\bar{x}_1 = \$3.00$	$\bar{x}_2 = \$2.95$
$s_1 = \$.11$	$s_2 = \$.09$

Solution **The consumer group wants to test the hypotheses**

H_0: $(\mu_1 - \mu_2) = 0$ (i.e., no difference between mean retail prices)

H_a: $(\mu_1 - \mu_2) > 0$ (i.e., mean retail price per pound of brand A is higher than that of brand B)

where

μ_1 = Mean retail price per pound of brand A coffee at all super-markets

μ_2 = Mean retail price per pound of brand B coffee at all super-markets

This one-tailed, large-sample test is based on a z statistic. Thus, we will reject H_0 if $z > z_\alpha = z_{.01}$. Since $z_{.01} = 2.33$, the rejection region is given by $z > 2.33$ (see Fig. 9.3)

We compute the test statistic as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}}} = \frac{(3.00 - 2.95) - 0}{\sqrt{\frac{(.11)^2}{75} + \frac{(.09)^2}{64}}} = 2.947$$

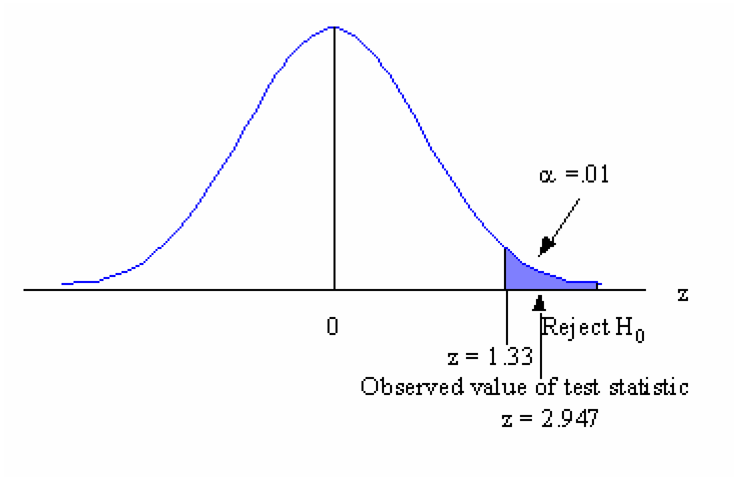


Figure 9.3 *Rejection region for Example 9.6*

Since this computed value of z = 2.947 lies in the rejection region, there is sufficient evidence (at α = .01) to conclude that the mean retail price per pound of brand A coffee is significantly higher than the mean retail price per pound of brand B coffee. The probability of our having committed a Type I error is α = .01.

When the sample sizes n_1 and n_2 are inadequate to permit use of the large-sample procedure of Example 9.9, we have made some modifications to perform a small-sample test of hypothesis about the difference between two population means. The test procedure is based on assumption that are more restrictive than in the large-sample case. The elements of the hypothesis test and required assumption are listed in the next box.

Small-sample test of hypothesis about (μ_1 - μ_2)

ONE -TAILED TEST	TWO -TAILED TEST
H_0: $(\mu_1 - \mu_2) = D_0$	H_0: $(\mu_1 - \mu_2) = D_0$
H_a: $(\mu_1 - \mu_2) > D_0$ (or H_a: $(\mu_1 - \mu_2) < D_0$)	H_a: $(\mu_1 - \mu_2) \neq D_0$

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Rejection region:

$t > t_\alpha$ (or $t < -t_\alpha$)

Rejection region:

$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and the distribution of t is based on (n_1 + n_2 - 2) degrees of freedom.

Assumptions:

1. The population from which the samples are selected both have approximately normal relative frequency distributions.
2. The variances of the two populations are equal.
3. The random samples are selected in an independent manner from the two populations.

Example 9.7 **There was a research on the weights at birth of the children of urban and rural women. The researcher suspects there is a significant difference between the mean weights at birth of children of urban and rural women. To test this hypothesis, he selects independent random samples of weights at birth of children of mothers from each group, calculates the mean weights and standard deviations and summarizes in Table 9.2. Test the researcher's belief, using a significance of α = .02.**

Table 9.2 *Weight at birth data for Example 9.7*

Urban mothers	Rural mothers
$n_1 = 15$	$n_2 = 14$
$\bar{x}_1 = 3.5933\,kg$	$\bar{x}_2 = 3.2029\,kg$
$s_1 = .3707\ kg$	$s_2 = .4927\ kg$

Solution **The researcher wants to test the following hypothesis:**

H_0: (μ_1 - μ_2) = 0 (i.e., no difference between mean weights at birth)

H_a: (μ_1 - μ_2) ≠ 0 (i.e., mean weights at birth of children of urban and rural women are different)

where μ_1 and μ_2 are the true mean weights at birth of children of urban and rural women, respectively.

Since the sample sizes for the study are small (n_1 = 15, n_2 = 14), the following assumptions are required:

1. The populations of weights at birth of children both have approximately normal distributions.

2. The variances of the populations of weights at birth of children for two groups of mothers are equal.
3. The samples were independently and randomly selected.

If these three assumptions are valid, the test statistic will have a *t*-distribution with (n_1 + n_2 - 2) = (15 + 14 - 2) = 27 degree of freedom with a significance level of α = .02, the rejection region is given by

$t < -t_{.01} = -2.473$ or $t > t_{.01} = 2.473$ (see Figure 9.4)

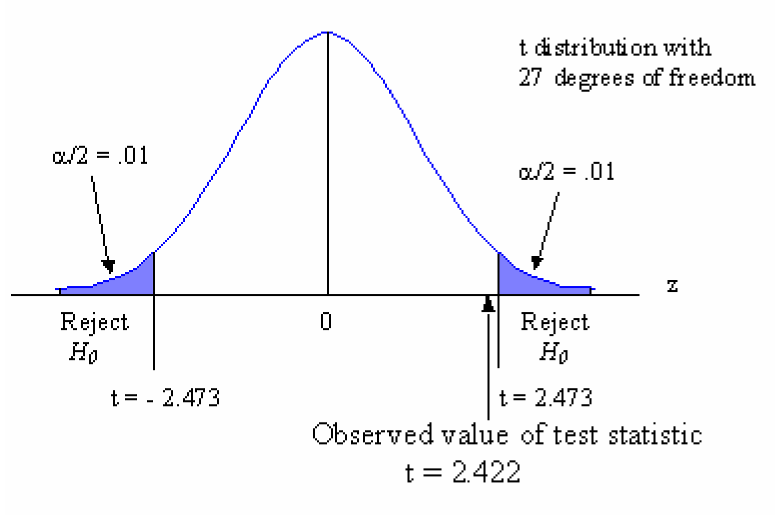


Figure 9.4 *Rejection region of Example 9.7*

Since we have assumed that the two populations have equal variances (i.e. that $\sigma_1^2 = \sigma_2^2 = \sigma$), we need to compute an estimate of this common variance. Our pooled estimate is given by

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(15-1)(.3707)^2 + (14-1)(.4927)^2}{15+14-2} = 0.1881$$

Using this pooled sample variance in the computation of the test statistic, we obtain

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(3.5933 - 3.2029) - D_0}{\sqrt{.1881\left(\frac{1}{15} + \frac{1}{14}\right)}} = 2.422$$

Now the computed value of *t* does not fall within the rejection region; thus, we fail to reject the null hypothesis (at α = .02) and conclude that there is insufficient evidence of a difference between the mean weights at birth of children of urban and rural women.

In this example, we can see that the computed value of *t* is very closed to the upper boundary of the rejection region. This region is specified by the significance level and the degree of freedom. How is the conclusion about the difference between the mean weights at births affected if the significance level is α = .05? We will answer the question in the next example.

Example 9.8 **Refer Example 9.7. Test the investigator's belief, using a significance level of $\alpha = .05$.**

Solution **With a significance level of $\alpha = .05$, the rejection region is given by**
$t < - t_{.025} = - 2.052$ or $t > t_{.025} = 2.052$ (see Figure 9.5)

Since the sample sizes are not changed, therefore test statistic is the same as in Example 9.10, $t = 2.422$.

Now the value of t falls in the rejection region; and we have sufficient evidence at a significance level of $\alpha = .05$ to conclude that the mean weight at birth of children of urban women differs significantly (or we can say that is higher than) from the mean weight at birth of children of rural women. But you should notice that the probability of our having committed a Type I error is $\alpha = .05$.

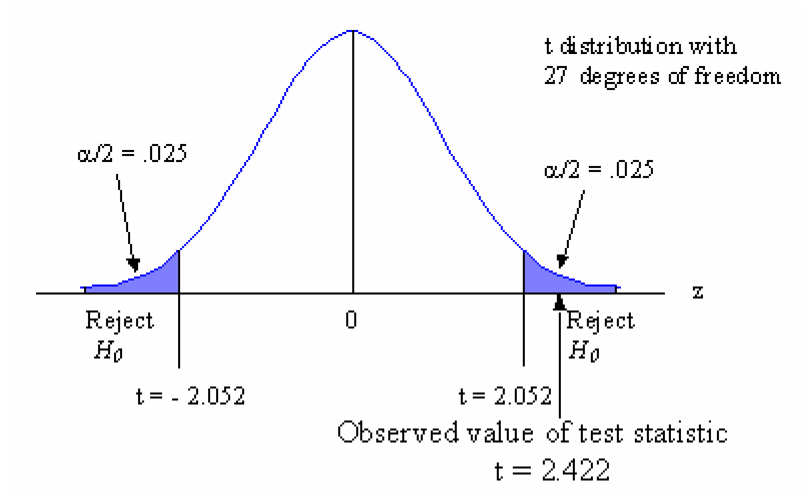


Figure 9.5 *Rejection region of Example 9.8*

9.5 Hypothesis tests about the difference between two proportions

Suppose we are interested in comparing p_1, the proportion of a population with p_2, the proportion of other population. Then the target parameter about which we will test a hypothesis is $(p_1 - p_2)$. Recall that p_1, and p_2 also represent the probabilities of success for two binomial experiments. The method for performing a large-sample test of hypothesis about $(p_1 - p_2)$, the difference between two binomial proportions, is outlined in the following box.

Large-sample test of hypothesis about $(p_1 - p_2)$

ONE -TAILED TEST	TWO -TAILED TEST
H_0: $(p_1 - p_2) = D_0$	H_0: $(p_1 - p_2) = D_0$
H_a: $(p_1 - p_2) > D_0$ or (H_a: $(p_1 - p_2) < D_0$)	H_a: $(p_1 - p_2) \neq D_0$

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

Rejection region: $z > z_\alpha$ (or $z < -z_\alpha$)

Rejection region: $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

where $\sigma_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

when $D_0 \neq 0$, calculate $\sigma_{(\hat{p}_1-\hat{p}_2)}$ using $\hat{p}_1$ and $\hat{p}_2$:

$$\sigma_{(\hat{p}_1-\hat{p}_2)} \approx \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$.

For the special case where D_0 = 0, calculate

$$\sigma_{(\hat{p}_1-\hat{p}_2)} \approx \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

when the total number of successes in the combined samples is (x_1 + x_2) and

$$\hat{p}_1 = \hat{p}_2 = \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Assumption: The intervals $\hat{p}_1 \pm 2\sqrt{\hat{p}_1\hat{q}_1 / n_1}$ and $\hat{p}_2 \pm 2\sqrt{\hat{p}_2\hat{q}_2 / n_2}$ do not contain 0 and 1.

When testing the null hypothesis that (p_1 - p_2) equals some specified difference D_0, we make a distinction between the case D_0 = 0 and the case $D_0 \neq 0$. For the special case D_0 = 0, i.e., when we are testing H_0: (p_1 - p_2) = 0 or, equivalently, H_0: $p_1 = p_2$, the best estimate of $p_1 = p_2 = p$ is found by dividing the total number of successes in the combined samples by the total number of observations in the two samples. That is, if x_1 is the number of successes in sample 1 and x_2 is the number of successes in sample 2, then

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

In this case, the best estimate of the standard deviation of the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is found by substituting $\hat{p}$ for both $\hat{p}_1$ and $\hat{p}_2$:

$$\sigma_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

For all cases in which $D_0 \neq 0$ [for example, when testing H_0: (p_1 - p_2)=.2], we use $\hat{p}_1$ and $\hat{p}_2$ in the formula for $\sigma_{(\hat{p}_1-\hat{p}_2)}$.

However, in most practical situations, we will want to test for a difference between proportions - that is, we will want to test H_0: (p_1 - p_2) = 0.

The sample sizes n_1 and n_2, must be sufficiently large to ensure that the sampling distribution of $\hat{p}_1$ and $\hat{p}_2$, and hence of the difference $(\hat{p}_1 - \hat{p}_2)$ are approximately normal. The rule of thumb given in the previous box may be used to determine if the sample sizes are "sufficiently large."

Example 9.9 **Two types of needles, the old type and the new type, used for injection of medical patients with a certain substance. The patients were allocated at random to two group, one to receive the injection from needle of the old type, the other to receive the injection from needles of the new type. Table 9.3 shows the number of patients showing reactions to the injection. Does the information support the belief that the proportion of patients giving reactions to needles of the old type is less than the corresponding proportion patients giving reactions to needles of the new type? Test at significance level of α = .01.**

Table 9.3 *Data on the patients' reactions in Example 9.9*

	Injected by old type needles	Injected by new type needles
Number of sampled patients	100	100
Number in sample with reactions	37	56

Solution **We wish to perform a test of**

H_0: $(p_1 - p_2) = 0$

H_a: $(p_1 - p_2) < 0$

where

p_1 = Proportion of patients giving reactions to needles of the old type.

p_2 = Proportion of patients giving reactions to needles of the new type.

For this large-sample, one-tailed test, the null hypothesis will be rejected if

$z < -z_{.01}$, = -2.33 (see Figure 9.6)

The sample proportions p_1 and p_2 are computed for substitution into the formula for the test statistic:

$\hat{p}_1 =$ Sample proportion of patients giving reactions with needles of the old type

$$= \frac{37}{100} = .37$$

$\hat{p}_2 =$ Sample proportion of patients giving reactions with needles of the new type

$$= \frac{56}{100} = .56$$

Hence,

$\hat{q}_1 = 1 - \hat{p}_1 = 1 - .37 = .63$ and $\hat{q}_2 = 1 - \hat{p}_2 = 1 - .56 = .44$

Since D_0 = 0 for this test of hypothesis, the test statistic is given by

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\hat{p} = \frac{\text{Total number of patients giving reactions with needles of both types}}{\text{Total number of patients sampled}}$$

$$= \frac{37 + 56}{100 + 100} = .465$$

Then we have

$$z = \frac{(.37 - .56) - 0}{\sqrt{(.465)(.535)\left(\frac{1}{100} + \frac{1}{100}\right)}} = -2.69$$

This value falls below the critical value of - 2.33. Thus, at α = .01, we reject the null hypothesis; there is sufficient evidence to conclude that the proportion of patients giving reactions to needles of the old type is significantly less than the corresponding proportion of patients giving reactions to needles of the new type, i.e., $p_1 < p_{2.}$

The inference derived from the test in Example 9.12 is valid only if the sample sizes, n_1 and n_2, are sufficiently large to guarantee that the intervals

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}} \quad \text{and} \quad \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_1}}$$

do not contain 0 and 1. This requirement is satisfied for Example 9.12:

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}} = .37 \pm 2\sqrt{\frac{(.37)(.63)}{100}} = .37 \pm .097 \text{ or } (.273,.467)$$

$$\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n_1}} = .56 \pm 2\sqrt{\frac{(.56)(.44)}{100}} = .56 \pm .099 \text{ or } (.467,.659)$$

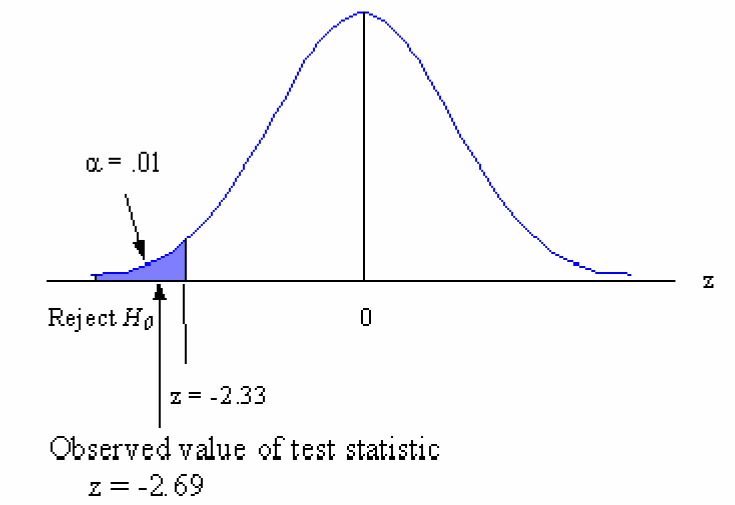


Figure 9.6 *Rejection region of Example 9.9*

9.6 Hypothesis test about a population variance

Hypothesis tests about a population variance σ^2 are conducted using the chi-square (χ^2) distribution introduced in Section 7.9. The test is outlined in the box. Note that the assumption of a normal population is required regardless of whether the sample size n is large or small.

Example 9.10 **A quality control supervisor in a cannery knows that the exact amount each can contains will vary, since there are certain uncontrollable factors that affect the amount of fill. The mean fill per can is important, but equally important is the variation σ^2**

of the amount of fill. If σ^2 is large, some cans will contain too little and others too much. Suppose regulatory agencies specify that the standard deviation of the amount of fill should be less than .1 ounce. The quality control supervisor sampled n = 10 cans and calculated s = .04. Does this value of s provide sufficient evidence to indicate that the standard deviation σ of the fill measurements is less than .1 ounce?

Test of hypothesis about a population variance σ^2

ONE -TAILED TEST

H_0: $\sigma^2 = \sigma_0^2$

H_a: $\sigma^2 > \sigma_0^2$ or (H_a: $\sigma^2 < \sigma_0^2$)

TWO -TAILED TEST

H_0: $\sigma^2 = \sigma_0^2$

H_a: $\sigma^2 \neq \sigma_0^2$

Test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0{}^2}$$

Rejection region:

$\chi^2 > \chi^2_\alpha$ (or $\chi^2 < \chi^2_{1-\alpha}$)

Rejection region:

$\chi^2 < \chi^2_{1-\alpha/2}$ or $\chi^2 > \chi^2_{\alpha/2}$

where χ^2_∞ and $\chi^2_{1-\infty}$ are values of χ^2 that locate an area of α to the right and α to the left, respectively, of a chi-square distribution based on (n -1) degrees of freedom.

[*Note*: σ_0^2 is our symbol for the particular numerical value specified for σ^2 in the null hypothesis.]

Assumption: The population from which the random sample is selected has an approximate normal distribution.

Solution **Since the null and alternative hypotheses must be stated in terms of σ^2 (rather than σ), we will want to test the null hypothesis that σ^2 = .01 against the alternative that σ^2 < .01. Therefore, the elements of the test are**

H_0: $\sigma^2 = .01$

H_a: $\sigma^2 < .01$

Assumption: The population of "amounts of fill" of the cans are approximately normal.

$$\text{Test statistic}: \chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Rejection region: The smaller the value of s^2 we observe, the stronger the evidence in favor of H_a. Thus, we reject H_0 for "small values" of the test statistic. With α = .05 and 9 df, the χ^2 value for rejection is found in Table 3, Appendix C and pictured in Figure 9.7. We will reject H_0 if χ^2 < 3.32511.

Remember that the area given in Table 3 of Appendix C is the area to the right of the numerical value in the table. Thus, to determine the lower-tail value that has α = .05 to its left, we use the $\chi^2_{.95}$ column in Table 3 of Appendix C.

Since

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9(.04)^2}{.01} = 1.44$$

is less than 3.32511, the supervisor can conclude that the variance of the population of all amounts of fill is less than .01 ($\sigma < 0.1$) with 95 % confidence. As usual, the confidence is in the procedure used - the χ^2 test. If this procedure is repeatedly used, it will incorrectly reject H_0 only 5% of the time. Thus, the quality control supervisor is confident in the decision that the cannery is operating within the desired limits of variability.

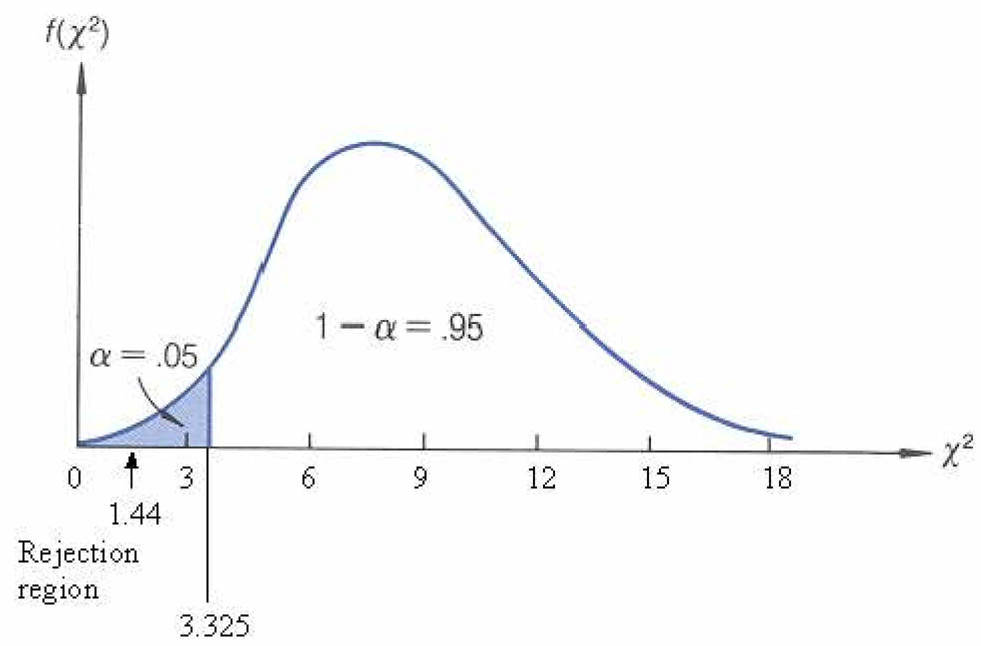


Figure 9.7 *Rejection region of Example 9.10*

9.7 Hypothesis test about the ratio of two population variances

In this section, we present a test of hypothesis for comparing two population variances, σ_1^2 and σ_2^2. Variance tests have broad applications in business. For example, a production manager may be interested in comparing the variation in the length of eye-screws produced on each of two assembly lines. A line with a large variation produces too many individual eye-screws that do not meet specifications (either too long or too short), even though the mean length may be satisfactory. Similarly, an investor might want to compare the variation in the monthly rates of return for two different stocks that have the same mean rate of return. In this case, the stock with the smaller variance may be preferred because it is less risky - that is, it is less likely to have many very low and very high monthly return rates.

Test of hypothesis for the ratio of two population variances, σ_1^2/σ_2^2	
ONE -TAILED TEST	TWO -TAILED TEST
H_0: $\sigma_1^2/\sigma_2^2 = 1$ $\quad (i.e.\ \sigma_1^2 = \sigma_2^2)$	H_0: $\sigma_1^2/\sigma_2^2 = 1$ $\quad (i.e.\ \sigma_1^2 = \sigma_2^2)$

H_a: $\sigma_1^2 / \sigma_2^2 > 1$ $\quad (i.e. \sigma_1^2 > \sigma_2^2)$ or
[H_a: $\sigma_1^2 / \sigma_2^2 < 1$ $\quad (i.e.\ \sigma_1^2 < \sigma_2^2)$]

Test statistic:

$$F = \frac{s_1^2}{s_2^2} \text{ or } F = \frac{s_2^2}{s_1^2}$$

Rejection region:

$F > F_\alpha$

H_a: $\sigma_1^2 / \sigma_2^2 \neq 1$ $\quad (i.e. \sigma_1^2 \neq \sigma_2^2)$

Test statistic:

$$F = \frac{\text{Larger sample variance}}{\text{Smaller sample variance}} \text{ i.e.}$$

$$F = \begin{cases} \dfrac{s_1^2}{s_2^2} & \text{when } s_1^2 > s_2^2 \\ \dfrac{s_2^2}{s_1^2} & \text{when } s_2^2 > s_1^2 \end{cases}$$

Rejection region:

$F > F_{\alpha/2}$

where F_α, and $F_{\alpha/2}$ are values that locate an area α and $\alpha/2$, respectively, in the upper tail of the F-distribution with ν_1 = numerator degrees of freedom (i.e., the df for the sample variance in the numerator) and ν_2 = denominator degrees of freedom (i.e., the df for the sample variance in the denominator).

Assumptions: 1. Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.

2. The random samples are selected in an independent manner from the two populations.

Variance tests can also be applied prior to conducting a small-sample t test for (μ_1 - μ_2), discussed in Section 9.4. Recall that the t test requires the assumption that the variances of the two sampled populations are equal. If the two population variances are greatly different, any inferences derived from the t test are suspect. Consequently, it is important that we detect a significant difference between the two variances, if it exists, before applying the small-sample t test.

The common statistical procedure for comparing two population variances, σ_1^2 and σ_2^2, makes an inference about the ratio σ_1^2 / σ_2^2. This is because the sampling

distribution of the estimator for σ_1^2 / σ_2^2 is well known when the samples are randomly and independently selected from two normal populations.

The elements of a hypothesis test for the ratio of two population variances, σ_1^2 / σ_2^2, are given in the preceding box.

Example 9.11 **A class of 31 students were randomly divided into an experimental set of size n_1 = 18 that received instruction in a new statistics unit and a control set of size n_2 =**

13 that received the standard statistics instruction. All students were given a test of computational skill at the end of the course. A summary of the results appears in Table 9.4. Do the data provide sufficient evidence to indicate a difference in the variability of this skill in the hypothetical population of students who might be given the new instruction and the population of students who might be given the standard instruction? Test using α = .01.

Table 9.4 *Data on students' scores in Example 9.11*

	Control set	Experimental set
Sample size	18	13
Standard deviation	1.93	3.10

Solution **Let**

σ_1^2 = Variance of test scores of the experimental population

σ_2^2 = Variance of test scores of the control population

The hypotheses of interest are

H_0: $\sigma_1^2 / \sigma_2^2 = 1 \qquad (\sigma_1^2 = \sigma_2^2)$

H_a: $\sigma_1^2 / \sigma_2^2 \neq 1 \qquad (\sigma_1^2 \neq \sigma_2^2)$

According to the box, the test statistic for this two-tailed test is

$$F = \frac{\text{Larger } s^2}{\text{Smaller } s^2} = \frac{s_2^2}{s_1^1} = \frac{(3.10)^2}{(1.93)^2} = 2.58$$

To find the appropriate rejection region, we need to know the sampling distribution of the test statistic. Under the assumption that both samples of test scores come from normal populations, the F statistic, $F = s_2^2 / s_1^2$, possesses an F distribution with $\nu_1 = (n_2 - 1)$ numerator degrees of freedom and $\nu_2 = (n_1 - 1)$ denominator degrees of freedom.

Unlike the z and t-distributions of the preceding sections, an F-distribution can be symmetric about its mean, skewed to the left, or skewed to the right; its exact shape depends on the degrees of freedom associated with s_2^2 and s_1^2, in this example, $(n_2 - 1) = 12$ and $(n_2 - 1) = 17$, respectively. An F-distribution with $\nu_1 = 12$ numerator df and $\nu_2 = 17$ denominator df is shown in Figure 9.8. You can see that this particular F-distribution is skewed to the right.

Upper-tail critical values of F are found in Table 4 of Appendix C. Table 9.5 is partially reproduced from this table. It gives F values that correspond to α = .05 upper-tail areas for different pairs of degrees of freedom. The columns of the table correspond to various numerator degrees of freedom, while the rows correspond to various denominator degrees of freedom.

Thus, if the numerator degrees of freedom are 12 and the denominator degrees of freedom are 17, we find the F value,

$F_{.05}$ = 2.38

As shown in Figure 9.8, $\alpha/2$ = .0 5 is the tail area to the right of 2.38 in the F-distribution with 12 numerator df and 17 denominator df. Thus, the probability that the F statistic will exceed 2.38 is $\alpha/2$ = .05.

Given this information on the F-distribution, we are now able to find the rejection region for this test. Since the test is two-tailed, we will reject H_0 if $F > F_{\alpha/2}$. For α = .10, we have $\alpha/2$ = .05 and $F_{.05}$ = 2.38 (based on ν_1 = 12 and ν_2 = 17 df). Thus, the rejection region is

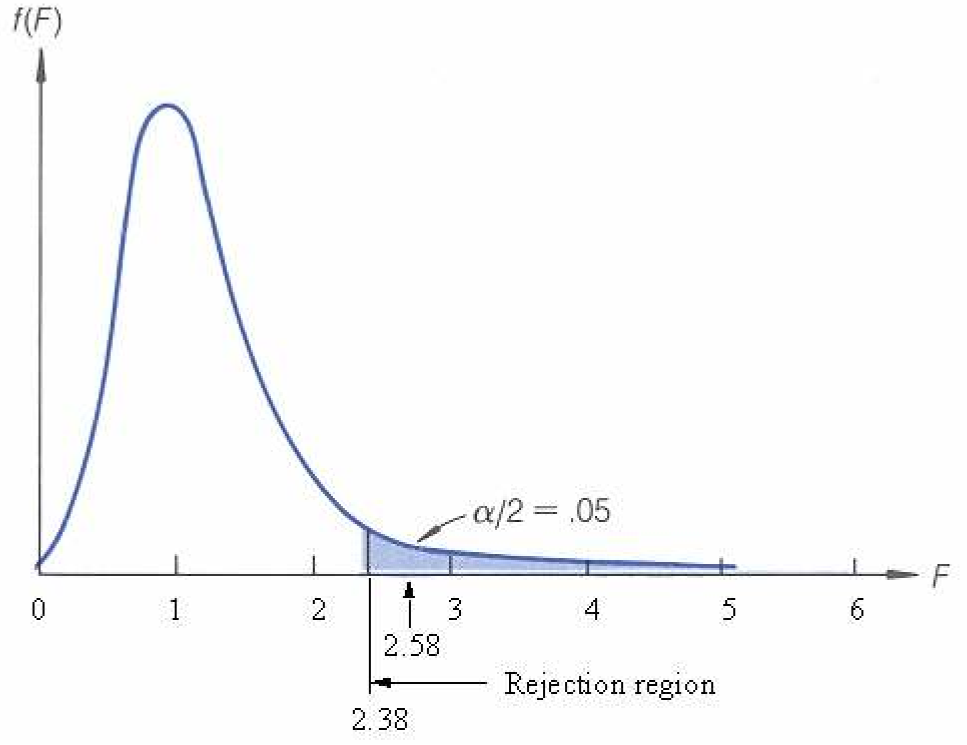


Figure 9.8 *Rejection region of Example 9.11*

Rejection region: Reject H_0 if $F > 2.38$.

Since the test statistic, F = 2.58, falls in the rejection region (see Figure 9.8), we reject H_0. Therefore, at α = .10, the data provide sufficient evidence to indicated that the population variances differ. It appears that the new statistics instruction results in a greater variability in computational skill.

Example 9.11 illustrates the technique for calculating the test statistic and rejection region for a two-tailed F test. The reason we place the larger sample variance in the numerator of the test statistic is that only upper-tail values of F are shown in the F table of Appendix C - no lower-tail values are given. By placing the larger sample variance in the numerator, we make certain that only the upper tail of the rejection region is used. The fact that the upper-tail area is $\alpha/2$ reminds us that the test is two-tailed.

The problem of not being able to locate an F value in the lower tail of the F-distribution is easily avoided in a one-tailed test because we can control how we specify the ratio of the population variances in H_0 and H_a. That is, we can always make a one-tailed test an upper-tailed test.

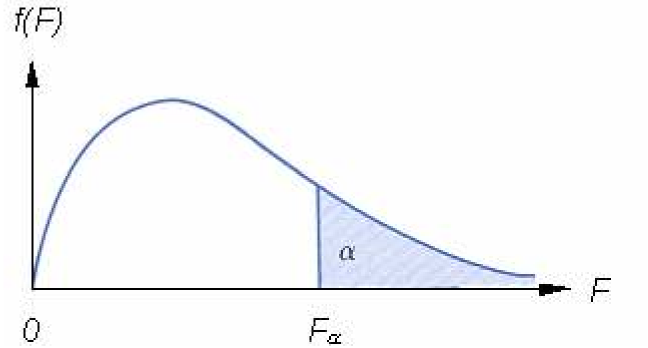


Table 9.5 *Reproduction of part of Table 4 from Appendix C; α = .05*

	ν_1	Numerator degrees of freedom									
ν_2		10	12	15	20	24	30	40	60	120	∞
	1	241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.33	254.30
	2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
Denominator	3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
degrees	5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
of	6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
freedom	8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
	12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
	14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
	15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
	17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96

9.8 Summary

In this chapter we have learnt the procedures for testing hypotheses about various population parameters. Often the comparison focuses on the means. As we note with the estimation techniques of Chapter 7, fewer assumptions about the sampled populations are required when the sample sizes are large. It would be emphasized that statistical significance differs from practical significance, and the two must not be confused. A reasonable approach to hypothesis testing blends a valid application of the formal statistical procedures with the researcher's knowledge of the subject matter.

9.9 Exercises

9.1. A random sample of n observation is selected from a population with unknown mean μ and variance σ^2. For each of the following situations, specify the test statistic and reject region.

a. $H_0 : \mu=40,\ H_a : \mu>40;\ n=35,\ \bar{x}=60,\ s^2=64;\quad \alpha=.05$
b. $H_0 : \mu=120,\ H_a : \mu\neq 120;\ n=40,\ \bar{x}=140.5,\ s=9.6;\ \alpha=.01$
c. $H_0 : \mu=11,\ H_a : \mu<11;\ n=48,\ \bar{x}=9.5,\ s=.6;\quad \alpha=.10$

9.2. A random sample of 51 measurements produced the following sums:

$$\sum x=50.3 \qquad \sum x^2=68$$

a. Test the null hypothesis that μ = 1.18 against the alternative that μ < 1.18. Use α = .01.
b. Test the null hypothesis that μ = 1.18 against the alternative that μ < 1.18. Use α = .10.

9.3. A random sample of n observations is selected from a binominal population. For each of the following situations, specify the rejection region, test statistic value, and conclusion:

a. $H_0 : p=.25,\ H_a : p>.25,\ \hat{p}=.28,\quad n=200,\quad \alpha=.10$
b. $H_0 : p=.05,\ H_a : p<.05,\ \hat{p}=.04,\quad n=1{,}000,\quad \alpha=.05$
c. $H_0 : p=.85,\ H_a : p\neq .85,\ \hat{p}=.80,\quad n=60,\quad \alpha=.01$

9.4. Two independent random samples are selected from populations with means μ_1 and μ_2, respectively. The sample sizes, means, and standard deviations are shown in the table.

Sample 1	Sample 2
$\bar{x}=7.5$	$\bar{x}=6.5$
$s=3.0$	$s=1.0$
$n=45$	$n=55$

a. Test the null hypothesis H_0: $(\mu_1 - \mu_2) = 0$ against the alternative hypothesis H_a: $(\mu_1 - \mu_2) \neq 0$ at α = .05.
b. Test the null hypothesis H_0: $(\mu_1 - \mu_2) = .5$ against the alternative hypothesis H_a: $(\mu_1 - \mu_2) \neq .5$ at α = .05.

9.5. Independent random samples selected from two binomial populations produced the results given in the table

	Sample 1	Sample 2
Number of successes	80	74
Sample sizes	100	100

a. Test $H_0 : (p_1 - p_2) = 0,\ H_a : (p_1 - p_2) > 0,$ at $\alpha = .10$

b. Suppose $n_1 = n_2 = 1{,}000$, but the sample estimates $\hat{p}_1, \hat{p}_2$, and $\hat{p}$ remain the same as in part a. Test $H_0 : (p_1 - p_2) = 0,\ H_a : (p_1 - p_2) > 0,$ at $\alpha = .05.$

9.6. A random sample of n = 10 observations yields $\bar{x} = 231.7$ and $s^2 = 15.5.$ Test the null hypothesis H_0: σ^2 = 20 against the alternative hypothesis
H_a: σ^2 < 20. Use α = .05. What assumptions are necessary for the test to be valid.

9.7. The following measurements represent a random sample of n = 5 observations from a normal population: 10, 2, 7, 9, 14. Is this sufficient evidence to conclude that $\sigma^2 \neq 2$. Test using α = .10.

9.8. Calculate the value of the test statistic for testing H_0: σ_1^2/σ_2^2 in each of following cases:

a. $H_a : \sigma_1^2 / \sigma_2^2 > 1;\ s_1^2 = 1.75,\ s_2^2 = 1.23$

b. $H_a : \sigma_1^2 / \sigma_2^2 < 1;\ s_1^2 = 1.52,\ s_2^2 = 5.90$

c. $H_a : \sigma_1^2 / \sigma_2^2 \neq 1;\ s_1^2 = 1{,}750,\ s_2^2 = 2{,}235$

Chapter 10 Categorical data analysis and analysis of variance

CONTENTS

10.1 Introduction

In this chapter we present some methods for treatment of categorical data. The methods involve the comparison of a set of observed frequencies with frequencies specified by some hypothesis to be tested. A test of such a hypothesis is called a *test of goodness of fit*.

We will show how to test the hypothesis that two categorical variables are independent. The test statistics discussed have sampling distributions that are approximated by chi-square distributions. The tests are called *chi-square tests*. These tests are useful in analyzing more than two population means.

In this chapter we will discuss the procedures for selecting sample data and analyzing variances. The objective of these sections is to introduce some aspects of experimental design and analysis of data from such experiments using an analysis of variance.

10.2 Tests of goodness of fit

We know that observations of a qualitative variable can only be categorized. For example, consider the highest level of education attained by each in a group of women in a rural region. "Level of education" is a qualitative variable and each woman would fall into one and only one of the following three categories: can read/write degree; primary degree; and secondary and above degree. The result of the categorization would be a count of the numbers of rural women falling in the respective categories. When the qualitative variable results in one of the two responses (yes or no, success or failure, favor or do not favor, etc.) the data (i.e., the counts) can be analyzed using the binomial probability distribution. However, qualitative variables such as "level of education" that allow for more than two categories for a response are much more common, and these must be analyzed using a different method called *test of goodness of fit.* A test of goodness of fit tests whether a given distribution fits a set of data. It is based on comparison of an observed frequency distribution with the hypothesized distribution.

Example 10.1 **Level of education attained by the women from a rural region is divided into three categories: can read/write degree; primary degree; secondary and above degree. A demographer estimates that 28% of them have can read/write degree, 61% have primary degree and 11% have higher secondary degree. In order to verify these**

percentages, a random sample of n = 100 women at the region were selected and their level of education recorded. The number of the women whose level of education falling into each of the three categories is shown in Table 10.1.

Table 10.1 *Categories corresponding to level of education*

Level of education			
Primary degree	Secondary degree	Higher secondary	Total
22	64	14	100

Do the data given in Table 10.1 disagree with the percentages of 28%, 61%, and 11% estimated by the demographer? As a first step in answering this question, we need to find the number of women in the sample of 100 that would be expected to fall in each of the three educational categories of Table 10.1, assuming that the demographer's percentages are accurate.

Solution **Each woman in the sample was assigned to one and only one of the three educational categories listed in Table 10.1. If the demographer's percentages are correct, then the probabilities that a education level will fall in the three educational categories are as shown in Table 10.2.**

Table 10.2 *Categories probabilities based on the demographer's percentages*

	Level of education			
	Can read/write	Primary	Secondary and above	Total
Cell number	1	2	3	
Cell probability	p_1 = .28	p_2=.61	p_3 =.11	1.00

Consider first the "Can read/write" cell of Table 10.2. If we assume that the level of education of any woman independent of the level of education of any other, then the observed number O_1, of responses falling into cell 1 is a binomial random variable and its expected value is

$e_1 = np_1 = (100)(.28) = 28$

Similarly, the expected observed numbers of responses in cells 2 and 3 (categories 2 and 3) are

$e_2 = np_2 = (100)(.61) = 61$

and

$e_3 = np_3 = (100)(.11) = 11$

The observed numbers of responses and the corresponding expected numbers (in parentheses) are shown in Table 10.3.

Table 10.3 *Observed and expected numbers of responses falling in the cell categories for Example 10.1*

	Level of education			
	Can read/write	Primary	Secondary and above	Total
Observed numbers	22	64	14	100

Expected numbers	(28)	(61)	(11)	100

Formula for calculating expected cell counts

$e_i = np_i$

where

e_i = Expected count for cell i

n = Sample size

p_i = Hypothesized Probability that an observation will fall in cell i.

Do the observed responses for the sample of 100 women disagree with the category probabilities based on the demographer's estimates? If they do, we say that the theorized demographer probabilities do not fit the data or, alternatively, that a lack of fit exists. The relevant null and alternative hypotheses are:

H_0: The category (cell) probabilities are p_1= .28, p_2= .61, p_3= .11

H_a: At least two of the probabilities, p_1, p_2, p_3, differ from the values specified in the null hypothesis

To find the value of the test statistic, we first calculate

$$\frac{(\text{Observed cell count - Expected cell count})^2}{\text{Expected cell count}} = \frac{(O_i - e_i)^2}{e_i}$$

for each of the cells, i = 1, 2, 3. The sum of these quantities is the test statistic used for the goodness-of-fit test:

$$\chi^2 = \frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2} + \frac{(O_3 - e_3)^2}{e_3} = \sum_{i=1}^{3} \frac{(O_i - e_i)^2}{e_i}$$

Substituting the values of the observed and expected cell counts from Table 10.3 into the formula for calculating χ^2, we obtain

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - e_i)^2}{e_i} = \frac{(22-28)^2}{28} + \frac{(64-61)^2}{61} + \frac{(14-11)^2}{11}$$
$$= 1.29 + .15 + .82 = 2.26$$

Example 10.2 **Specify the rejection region for the test described in the preceding discussion. Use α = .05. Test to determine whether the sample data disagree with the demographer's estimated percentages.**

Solution **Since the value of chi-square increases as the differences between the observed and expected cell counts increase, we will reject**

H_0: p_1 = .28, p_2 = .61, p_3 = .11

for values of chi-square larger than some critical value, say χ^2_α, i.e.,

Rejection region: $\chi^2 > \chi^2_\alpha$

The critical values of the χ^2 distribution are given in Table 3 of Appendix C. The degrees of freedom for the chi-square statistic used to test the goodness of fit of a set of cell probabilities will always be 1 less than the number of cells. For example, if k cells were used in the categorization of the sample data, then

Degrees of freedom: df = k - 1

For our example, df = (k - 1) = (3 - 1) = 2 and α = .05. From Table 3 of Appendix C, the tabulated value of χ^2_α, corresponding to df = 2 is 5.99147.

The rejection region for the test, $\chi^2 > \chi^2_\alpha$, is illustrated in Figure 10.1. We will reject H_0 if χ^2 > 5.99147. Since the calculated value of the test statistic, $\chi^2 = 2.26$, is less than $\chi^2_{.05}$, we can not reject H_0. There is insufficient information to indicate a lack of fit of the sample data to the percentages estimated by the demographer.

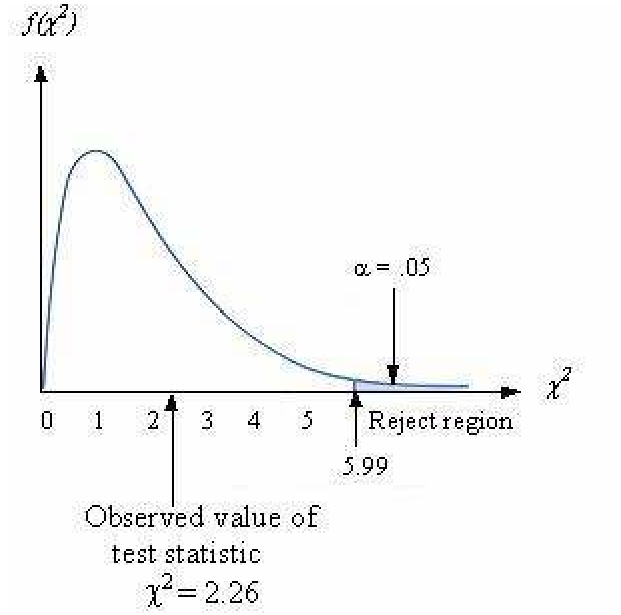


Figure 10.1 *Rejection region for Example 10.2*

Summary of a goodness of fit test for specified values of the Cell probabilities

H_0: The k cell probabilities are $p_1, \ p_2, \ldots, \ p_k$

H_a: At least two of the cell probabilities differ from the values specified in H_0

Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$

where

k = Number of cells in the categorization table

O_i = Observed count for cell *i*

e_i = Expected count for cell *i*

n = Sample size = $0_1 + 0_2 + \ldots + 0_k$

Rejection region: $\chi^2 > \chi^2_\alpha$

At the start, we assumed that each of *n* observations could fall into one of k categories (or cells), that the probability that an observation would fall in cell 1 was p_i, $i = 1, 2, \ldots, k$, and that the outcome for any one observation was independent of the outcome for any others. These characteristics define a multinomial experiment. The binomial experiment is a multinomial experiment with $k = 2$.

Properties of the underlying distribution of response data for a chi-square goodness of fit test

1. The experiment consists of n identical trials.
2. There are k possible outcomes to each trial.
3. The probabilities of the k outcomes, denoted by $p_1, \ p_2, \ldots, \ p_k$ remain the same from trial to trial, where $p_1 + p_2 + \ldots + p_k = 1.$
4. The trials are independent.
5. The (estimated) expected number of responses for each of the k cells should be at least 5.

Because it is widely used, the chi-square test is also one of the most abused statistical procedures. The user should always be certain that the experiment satisfies the assumptions before proceeding with the test. In addition, the chi- square test should be avoided when the estimated expected cell counts are small, because in this case the chi-square probability distribution gives a poor approximation to the sampling distribution of the χ^2 statistic. The estimated expected number of responses for each of the k cells should be at least 5. In this case the chi-square distribution can be used to determine an approximate critical value that specifies the rejection region.

In the sections that follow, we will present a method for analyzing data that have been categorized according to two qualitative variables. The objective is to determine whether a dependency exists between the two qualitative variables – the qualitative variable analogue to a correlation analysis for two quantitative random variables. As you will see subsequently, these methods are also based on the assumption that the sampling satisfies the requirements for one or more multinomial experiments.

10.3 The analysis of contingency tables

Qualitative data are often categorized according to two qualitative variables. As a practical example of a two-variable classification of data, we will consider a 2 × 3 table.

Suppose that a random sample of men and women indicated their view on a certain proposal as shown in Table 10.4.

Table 10.4 *Contingency table for views of women and men on a proposal*

	In favour	*Opposed*	*Undecided*	*Total*
Women	118	62	25	205
Men	84	78	37	199
Total	202	140	62	404

We are to test the statement that there is no difference in opinion between men and women, i.e. the response is independent of the sex of the person interviewed, and we adopt this as our null hypothesis. Now if the statement is not true, then the response will depend on the sex of the person interviewed, and the table will enable us to calculate the degree of dependence. A table constructed in this way (to indicate dependence or association) is called a *contingency table*. "Contingency" means dependence − many of you will be familiar with the terms "contingency planning"; i.e. plans that will be put into operation *if* certain things happen. Thus, the purpose of a contingency table analysis is to determine whether a dependence exists between the two qualitative variables.

We adopt the null hypothesis that there is no association between the response and the sex of person interviewed. On this basis we may deduce that the proportion of the sample who are female is 205/404, and as 202 people are in favour of the proposal, the expected number of women in favour of proposal is 205/404 × 202 = 102.5. Therefore, the estimated expected number of women (row 1) in favour of the proposal (column 1) is

$$e_{11} = \left(\frac{205}{404}\right) \times 202 = \left(\frac{\text{Row 1 total}}{n}\right) \times (\text{Column 1 total}) = 102.5$$

Also, as 140 people are against the proposal, the expected number of women against the proposal is (row 1, column 2)

$$e_{12} = \left(\frac{205}{404}\right) \times 140 = \left(\frac{\text{Row 1 total}}{n}\right) \times (\text{Column 2 total}) = 71$$

And the expected number of undecided women is (row 1, column 3)

$$e_{13} = \left(\frac{205}{404}\right) \times 62 = \left(\frac{\text{Row 1 total}}{n}\right) \times (\text{Column 3 total}) = 31.5$$

We now move to row 2 for men and note that the row total is 199. Therefore, we would expect the proportion of the sample who are male is 199/404 for all three types of opinion. The estimated expected cell counts for columns of row 2 are

$$e_{21} = \left(\frac{119}{404}\right) \times 202 = \left(\frac{\text{Row 2 total}}{n}\right) \times (\text{Column 1 total}) = 99.5$$

$$e_{22} = \left(\frac{119}{404}\right) \times 140 = \left(\frac{\text{Row 2 total}}{n}\right) \times (\text{Column 2 total}) = 69$$

$$e_{21} = \left(\frac{119}{404}\right) \times 62 = \left(\frac{\text{Row 2 total}}{n}\right) \times (\text{Column 3 total}) = 30.5$$

The formula for calculating any estimated expected value can be deduced from the values calculated above. Each estimated expected cell count is equal to the product of its respective row and column totals divided by the total sample size *n*:

$$e_{ij} = \frac{R_i \times C_j}{n}$$

where e_{ij} = Estimated expected counts for the cell in row i and column j

R_i = Row total corresponding to row i

C_j = Column total corresponding to column j

n = Sample size

The observed and estimated expected cell counts for the herring gull contingency table are shown in Table 10.5.

Table 10.5 *Observed and expected (in parentheses) counts for response of women and men*

	In-favour	*Opposed*	*Undecided*
Women	118 (102.5)	62 (71)	25 (31.5)
Men	84 (99.5)	78 (69)	37 (30.5)

In this example, the chi-square test statistic , χ^2, is calculated in the same manner as shown in Example 10.1.

$$\chi^2 = \frac{(O_{11} - e_{11})^2}{e_{11}} + \frac{(O_{12} - e_{12})^2}{e_{12}} + \frac{(O_{13} - e_{13})}{e_{13}} + \ldots + \frac{(O_{23} - e_{23})^2}{e_{23}}$$

$$= \frac{(118 - 102.5)^2}{102.5} + \frac{(62 - 71)^2}{71} + \frac{(25 - 31.5)^2}{31.5} + \ldots + \frac{(37 - 30.5)^2}{30.5} = 9.87$$

The appropriate degrees of freedom for a contingency table analysis will always be (r - 1) × (c -1), where r is the number of rows and c is the numbers of columns in the table. In this example, we have two degrees of freedom in calculating the expected values. Consulting Table 3 of Appendix C, we see that the critical values for χ^2 are 5.99 at a significance level of α = .05 and 9.21 at level of α = .01. In both cases, the computed test statistic is lager than these critical values. Hence, we would reject the null hypothesis accepting the alternative hypothesis that men and women think differently with 99% confidence.

General form of a chi-square test for independence of two directions of classification

H_0: The two direction of classification in the contingency table are independent

H_a: The two direction of classification in the contingency table are dependent

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$

where

r = Number of rows in the table
c = Number of columns in the table
O_{ij} = Observed number of responses in the cell in row *i* and column *j*
e_{ij} = Estimated expected number of responses in the cell(ij) = $(R_i \times C_j) / n$

Rejection region: $\chi^2 > \chi^2_\alpha$

where χ^2_α is tabulated value of the chi-square distribution based on *(r -1)* × *(c-1)* degrees of freedom such that $P(\chi^2 > \chi^2_\alpha) = \alpha$

10.4 Contingency tables in statistical software packages

In all statistical software packages there are procedures for analysis of categorical data. Following are printouts of the procedure "Crosstabs" of SPSS for creating the contingency table and computing value of the χ^2 statistic to test dependence of the education level on living region of women interviewed in the DHS Survey 1988 in Vietnam (data of the survey is given in Appendix A).

```
CROSSTABS

  /TABLES=urban  BY gd1
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ CC PHI
  /CELLS= COUNT EXPECTED ROW .
```

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
URBAN * Education Level	4172	100.0%	0	.0%	4171	100.0%

URBAN * Education Level Cross tabulation

			Education Level			
			Can read/write	Primary	Secondary and above	Total
URBAN	Urban	Count	163	299	266	728
		Expected Count	197.5	415.5	115.0	728.0
		% within URBAN	22.4%	41.1%	36.5%	100.0%
	Rural	Count	969	2082	393	3444
		Expected Count	934.5	1965.5	544.0	3444.0
		% within URBAN	28.1%	60.5%	11.4%	100.0%
Total		Count	1132	2381	659	4172
		Expected Count	1132.0	2381.0	659.0	4172.0
		% within URBAN	27.1%	57.1%	15.8%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	287.084[a]	2	.000
Likelihood Ratio	241.252	2	.000
Linear-by-Linear Association	137.517	1	.000
N of Valid Cases	4172		

[a] *0 cells (.0%) have expected count less than 5. The minimum expected count is 114.99.*

Before changing to discuss about analysis of variance, we make some remarks on methods for treating categorical data.

- Surveys that allow for more than two categories for a single response (a one-way table) can be analyzed using the chi-square goodness of fit test. The appropriate test statistic, called χ^2 statistic, has a sampling distribution approximated by the chi-square probability distribution and measures the amount of disagreement between the observed number of responses and the expected number of responses in each category.
- A contingency table analysis is an application of the χ^2 test for a two-way (or two-variable) classification of data. The test allows us to determine whether the two directions of classification are independent.

10.5 Introduction to analysis of variance

As we have seen in the preceding chapters, the solutions to many statistical problems are based on inferences about population means. Next sections extend the methods of Chapters 7 - 9 to the comparison of more than two means.

When the data have been obtained according to certain specified sampling procedures, they are easy to analyze and also may contain more information pertinent to the population means than could be obtained using simple random sampling. The procedure for selecting sample data is called the *design of the experiment* and the statistical procedure for comparing the population means is called an *analysis of variance*.

We will introduce some aspects of experimental design and the analysis of the data from such experiments using an analysis of variance.

10.6 Design of experiments

The process of collecting sample data is called an *experiment* and the variable to be measured in the experiment is called the *response*. The planning of the sampling procedure is called the *design* of the experiment. The object upon which the response measurement is taken is called an *experimental unit*.

Variables that may be related to a response variable are called *factors*. The value – that is, the intensity setting – assumed by a factor in an experiment is called a *level*. The combinations of levels of the factors for which the response will be observed are called *treatments*.

The process of the design of an experiment can be divided into four steps as follows:

1. Select the factors to be included in the experiment and identify the parameters that are the object of the study. Usually, the target parameters are the population means associated with the factor level.
2. Choose the treatments to be included in the experiment.
3. Determine the number of observations (sample size) to be made for each treatment.
4. Decide how the treatments will be assigned to the experimental units.

Once the data for a designed experiment have been collected, we will want to use the sample information to make inferences about the population means associated with the various treatments. The method used to compare the treatment means is known as *analysis of variance*, or **ANOVA**. The concept behind an analysis of variance can be explained using the following simple example.

Example 10.3 **A elementary school teacher wants to try out three different reading workbooks. At the end of the year the 18 children in the class will take a test in reading achievement. These test scores will be used to compare the workbooks. Table 10.6 gives reading achievement scores. Each set of scores of the 6 children using a type of workbook is considered as a sample from the hypothetical population of all kindergarten children who might use that type of workbook. The scores are plotted as line plots in Figure 10.2.**

Table 10.6 *Reading scores of 18 children using three different workbooks*

	Workbook 1	Workbook 2	Workbook 3
	2	9	4
	4	10	5
	3	10	6
	4	7	3
	5	8	7
	6	10	5
Sums	24	54	30
Sample means	4	9	5
Total of 3 samples: 108; mean of 3 samples: 6			

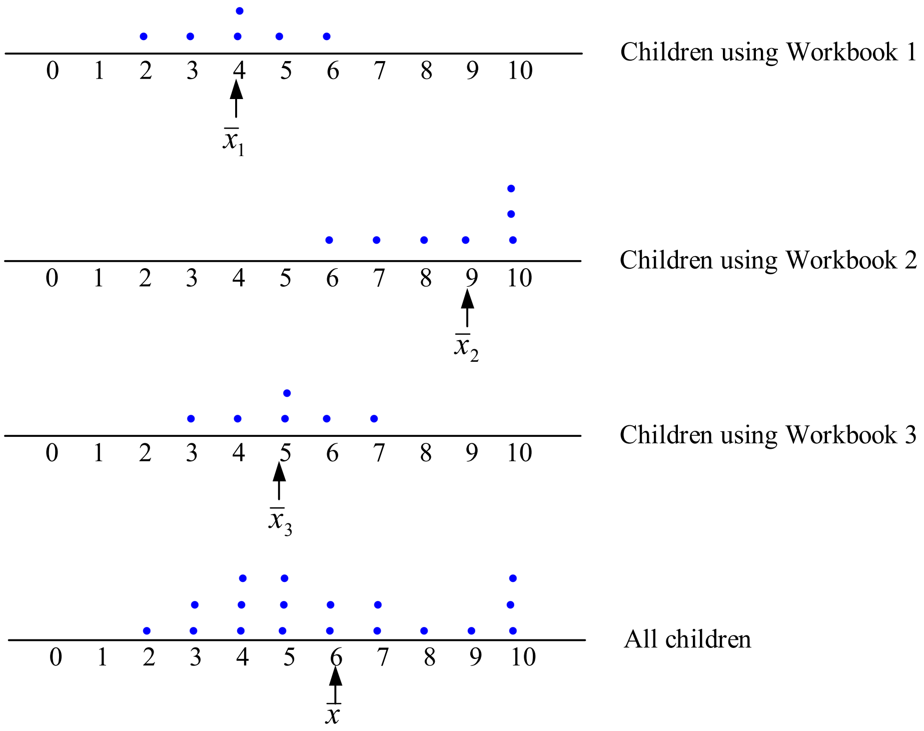


Figure 10.2 *Reading scores by workbook used and for combined sample*

The means of the three samples are 4, 9, and 5, respectively. Figure 10.2 shows these as the centers of the three samples; there is clearly variability from group to group. The variability in the entire pooled sample of 18 is shown by the last line.

In contrast to this rather typical allocation, we consider Tables 10.7 and 10.8 as illustrations of extreme cases. In Table 10.7 every observation in Group A is 3, every observation in Group B is 5, and every observation in Group C is 8. There is no variation within groups, but there is variation between groups.

Table 10.7 *No variation within groups*

	Group	
A	B	C
3	5	8
3	5	8
3	5	8
3	5	8

Means	3	5	8

In Table 10.8 the mean of each group is 3. There is no variation among the group means, although there is variability within each group. Neither extreme can be expected to occur in an actual data set. In actual data, one needs to make an assessment of the relative sizes of the between-groups and within-groups variability. It is to this assessment that the term "analysis of variance" refers.

Table 10.8 *No variation between groups*

		Group	
	A	B	C
	3	3	1
	5	6	4
	1	2	3
	3	1	4
Means	3	3	3

In Example 10.3, the overall mean, $\bar{x}$, is the sum of all the observations divided by the total number of observations:

$$\bar{x} = \frac{(2+4+\ldots+5)}{18} = \frac{108}{18} = 6$$

The sum of squared deviations of all 18 observations from mean of the combined sample is a measure of variability of the combined sample. This sum is called *Total Sum of Squares* and is denoted by SS(Total).

$$\begin{aligned} \text{SS(Total)} = \; & (2 - 6)^2 + (4 - 6)^2 + (3 - 6)^2 + (4 - 6)^2 + (5 - 6)^2 + (6 - 6)^2 + \\ & (9 - 6)^2 + (10- 6)^2 + (10 - 6)^2 + (7 - 6)^2 + (8 - 6)^2 + (10 - 6)^2 + \\ & (4 - 6)^2 + (5 - 6)^2 + (6 - 6)^2 + (3 - 6)^2 + (7 - 6)^2 + (5 - 6)^2 \\ = \; & 34 + 62 + 16 = 112 \end{aligned}$$

Next we measure the variability within samples. We calculate the sum of squared deviation of each of 18 observations from their respective group means. This sum is called the *Sum of Squares Within Groups* (or *Sum of Squared Errors*) and is denoted by SS(Within Groups) (or SSE).

$$\begin{aligned} \text{SSE} = \; & (2 - 4)^2 + (4 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 + \\ & (9 - 9)^2 + (10- 9)^2 + (10 - 9)^2 + (7 - 9)^2 + (8 - 9)^2 + (10 - 9)^2 + \\ & (4 - 5)^2 + (5 - 5)^2 + (6 - 5)^2 + (3 - 5)^2 + (7 - 5)^2 + (5 - 5)^2 \\ = \; & 10 + 8 + 10 = 28 \end{aligned}$$

Now let us consider the group means of 4, 9, and 5. The sum of squared deviation of the group means from the pooled mean of 6 is

$$(4 - 6)^2 + (9 - 6)^2 + (5 - 6)^2 = 4 + 9 + 1 =14.$$

However, this sum is not comparable to the sum of squares within groups because the sampling variability of means is less than that of individual measurements. In fact, the mean of a sample of 6 observations has a sampling of 1/6 the sampling variance of a single observation. Hence, to put the sum of squared deviations of group mean on a basis that can be compared with SS(Within Groups), we must multiply it by 6, the number of observation in each sample, to obtain 6 × 14 = 84. This is called the *Sum of Squares Between Groups* (or *Sum of the Squares for Treatment*) and is denoted by SS(Between Groups) (or *SST*).

Now we have three sums that can be compared: SS(Between Groups), SS(Within Groups), and SS(Total). They are given in Table 10.9. Observe that addition of the first two sum of squares gives the last sum. This demonstrates what we mean by the allocation of the total variability to the variability due to differences between means of groups and variability of individuals within groups.

Table 10.9 *Sums of Squares for Example 10.3*

SS(Between Groups)	84
SS(Within Groups)	28
SS(Total)	112

In this example we notice that the variability between groups is a large proportion of the total variability. However, we have to adjust the numbers in Table 10.9 in order to take account of the number of pieces of information going into each sum of squares. That is, we want to use the sums of squares to calculate sample variances. The sum of squares between groups has 3 deviations about the mean of combined sample. Therefore its number of degrees of freedom is 3 - 1 = 2 and the sample variance based on this sum of squares is

$$Between - Group\,Variation = \frac{SS(Between\,Groups)}{3-1} = \frac{84}{2} = 42$$

This quantity is also called *Mean Square for Treatments* (MST).

The sum of squares within groups is made up of 3 sample sums of squares. Each involves 6 squared deviations, and hence, each has 6 -1 = 5 degrees of freedom. Therefore 3 samples have 18 - 3 = 15 degrees of freedom. The sample variance based on this sum is

$$Within - Group\,Variation = \frac{SS(Within\,Groups)}{18-3} = \frac{28}{15} = 1.867$$

This variation is also called *Mean Square for Error* (MSE).

The two estimates of variation MST, measuring variability among groups and MSE, measuring variability within groups, are now comparable. Their ratio is

$$F = \frac{MST}{MSE} = \frac{42}{1.867} = 22.50\,.$$

The fact that MST is 22.5 times MSE seems to indicate that the variability among groups is much greater than that within groups. However, we know that such a ratio computed for different triplets of random samples would vary from triplet to triplet, even if the population means were the same. We must take account of this sampling variability. This is done by referring to the F-tables depending on the desired significance level as well as on the number of degrees of freedom of MST, which is 2 here, and the number of degrees of freedom of MSE, which is 15 here. The value in the F-table for a significance level of .01 is 6.36. Thus we would consider the calculated ratio of 22.50 as very significant. We conclude that there are real differences in average reading readiness due to the use of different workbooks.

The results of computation are set out in Table 10.10.

Table 10.10 *Analysis of Variance Table for Example 10.3*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean of Squares	F
Between groups	84	2	42	22.50
Within groups	28	15	1.867	

In next sections we will consider the analysis of variance for the general problem of comparing k population means for three special types of experimental designs.

10.7 Completely randomized designs

The most common experimental design employed in practice is called a completely randomized design. This experiment involves a comparison of the means for a number, say k, of treatments, based on independent random samples of $n_1, n_2, \ldots, n_k$ observations, drawn from populations associated with treatments $1, 2, \ldots, k,$ respectively.

After collecting the data from a completely randomized design, our goal is to make inferences about *k* population means where μ_i is the mean of the population of measurements associated with treatment *i*, for $i = 1, 2, \ldots, k$. The null hypothesis to be tested is that the *k* treatment means are equal, i.e.,

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

and the alternative hypothesis is that at least two of the treatment means differ.

An analysis of variance provides an easy way to analyze the data from a completely randomized design. The analysis partitions SS(Total) into two components, SST and SSE. These two quantities are defined in general term as follows:

$$SST = \sum_{j=1}^{k} (\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)$$

Recall that the quantity SST denotes the sum of squares for treatments and measures the variation explained by the differences between the treatment means. The sum of squares for error, SSE, is a measure of the unexplained variability, obtained by calculating a pooled measure of the variability within the k samples. If the treatment means truly differ, then SSE should be substantially smaller than SST. We compare the two sources of variability by forming an F statistic:

$$F = \frac{Between\text{ - }sample\ variation}{Within\text{ - }sample\ variation} = \frac{SST/(k-1)}{SSE/(n-k)} = \frac{MST}{MSE}$$

where n is the total number of measurements. Under certain conditions, the F statistic has a repeated sampling distribution known as the F-distribution. Recall from Section 9.6 that the F distribution depends on ν_1 numerator degrees of freedom and ν_2, denominator degrees of freedom. For the completely randomized design, F is based on $\nu_1 = (k - 1)$ and $\nu_2 = (n - k)$ degrees of freedom. If the computed value of F exceeds the upper critical value, F_∞ we reject H_0 and conclude that at least two of the treatment means differ.

Test to Compare k Population Means for a Completely Randomized Design

H_0: $\mu_1 = \mu_2 \ldots = \mu_k$ [i.e., there is no difference in the treatment (population) means]

H_a: At least two treatment means differ

Test statistic: F = MST/MSE

Rejection region: $F > F_\alpha$

where the distribution of F is based on (k - 1) numerator df and (n - k) denominator df, and F_α is the F value found in Table 4 of Appendix C such that $P(F > F_\alpha) = \alpha$.

Assumptions: 1. All k population probability distributions are normal.

2. The k population variances are equal.

3. The samples from each population are random and independent.

The results of an analysis of variance are usually summarized and presented in an analysis of variance (ANOVA) table. Such a table shows the sources of variation, their respective degrees of freedom, sums of squares, mean squares, and computed F statistic. The results of the analysis of variance for Example 10.3 are given in Table 10.9, and the general form of the ANOVA table for a completely randomized design is shown in Table 10.11.

Table 10.11 *Analysis of Variance Table for Completely Random Design*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean of Squares	F
Between groups	SST	k - 1	MST/(k – 1)	F = MST/MSE
Within groups	SSE	n - k	SSE/(n - k)	
Total	SS(Total)	n -1		

Example 10.4 **Consider the problem of comparing the mean number of children born to women in 10 provinces numbered from 1 to 10. Numbers of children born to 3448 women from these provinces are randomly selected from the column heading CEB of Appendix A. The women selected from 10 provinces are considered to be the only ones of interest. This ensure the assumption of equality between the population variances. Now, we want to compare the mean numbers of children born to all women in these provinces, i.e., we wish to test**

H_0: $\mu_1 = \mu_2 \ldots = \mu_{10}$

H_a: At least two population means differ

Solution **We will use the SPSS package to make an analysis of variance. Following are the syntax and the print out of the procedure "One-Way ANOVA" of SPSS for analysis of CEB by province.**

```
ONEWAY
  ceb BY province
  /STATISTICS DESCRIPTIVES
  /MISSING ANALYSIS .
```

ONEWAY

Descriptives

Children ever born

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	228	2.40	1.55	.10	2.19	2.60	0	10
2	323	2.84	2.30	.13	2.59	3.09	0	11
3	302	3.15	2.09	.12	2.91	3.39	0	12
4	354	2.80	2.00	.11	2.59	3.01	0	10
5	412	2.53	1.61	7.93E-02	2.37	2.68	0	9
6	366	3.08	1.99	.10	2.88	3.29	0	11
7	402	3.26	1.83	9.13E-02	3.08	3.44	0	10
8	360	3.45	2.21	.12	3.23	3.68	0	11
9	297	3.87	2.66	.15	3.56	4.17	0	12
10	403	3.75	2.52	.13	3.51	4.00	0	12
Total	3448	3.13	2.15	3.66E-02	3.06	3.20	0	12

ANOVA

Children born

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	702.326	9	78.036	17.621	.000
Within Groups	15221.007	3437	4.429		
Total	15923.333	3446			

From the printout we can see that the SPSS One-Way ANOVA procedure presents the results in the form of an ANOVA table. Their corresponding sums of squares and mean squares are:

SST = 702.326

SSE = 15221.007

MST = 78.036

MSE = 4.429

The computed value of the test statistic, given under the column heading F is

$F = 17.621$

with degrees of freedom between provinces is $\nu_1 = 9$ and degrees of freedom within provinces is $\nu_2 = 3437$.

To determine whether to reject the null hypothesis

H_0: $\mu_1 = \mu_2 \ldots = \mu_{10}$

in favor of the alternative

H_a: at least two population means are different

we may consult Table 4 of Appendix C for tabulated values of the F distribution corresponding to an appropriately chosen significance level α. However, since the SPSS printout gives the observed significance level (under the column heading Sig.) of the test, we will use this quantity to assist us in reaching a conclusion. This quality is the probability of obtaining F statistic at least as large as the one calculated when all population means are equal. If this probability is small enough, the null hypothesis (all population means are equal) is rejected. In this example, the observed significance level is approximately .0001. It implies that H_0 will be rejected at any chosen level of α lager than .0001. Thus, there is very strong evidence of a difference among the mean numbers of children ever born of women in 10 provinces. The probability that this procedure will lead to a Type I error is .0001.

Before ending our discussion of completely randomized designs, we make the following comment. The proper application of the ANOVA procedure requires that certain assumptions be satisfied, i.e., all k populations are approximately normal with equal variances. If you know, for example, that one or more of the populations are non-normal (e.g., highly skewed), then any inferences derived from the ANOVA of the data are suspect. In this case, we can apply a non-parametric technique.

10.8 Randomized block designs

Example 10.5 **Three methods of treating beer cans are being compared by a panel of 5 people. Each person samples beer from each type of can and scores the beer with a number (integer) between 0 and 6, 6 indicating a strong metallic taste and 0 meaning no metallic taste. It is obvious that different people will use the scale somewhat differently, and we shall take this into account when we compare the different types of can.**

The data are reported in Table 10.12. This is an example of a situation in which the investigator has data pertaining to *k* treatments (k = 3 types of can) in *b* blocks (b = 5 persons) . We let x_{gj} denote the observation corresponding to the *g*th treatment and the jth block, $\bar{x}_{g.}$ denote the mean of the *b* observations for the *g*th treatment, $\bar{x}_{.j}$ the mean of the k observations in the jth block, and $\bar{x}$ the overall mean of all n = kb observations. When this particular design is used, the three types of can are presented to the individuals in random order.

An experimental design of this type is called a *randomized blocks design*. In agricultural experiments the k treatments might correspond, for example, to k different fertilizers; the field would be divided into blocks of presupposed similar fertility; and every fertilizer was used in each block so that differences in fertility of the soil in different parts of the field (blocks) would not bias the comparison of the fertilizers. Each block would be subdivided into k sub-blocks, called "plots." The k fertilizers would be randomly assigned to the plots in each block; hence the name, "randomized blocks."

Table 10.12 *Scores of three types of can on "metallic" scale*

			Person			
Type of Can	P1	P2	P3	P4	P5	Sums
A	6	5	6	4	3	24
B	2	3	2	2	1	10
C	6	4	4	4	3	21
Sums	14	12	12	10	7	55

In general terms, we can define that a ***randomized block design*** as a design in which k treatments are compared within each of b blocks. Each block contains k matched experimental units and the k treatments are randomly assigned, one to each of the units within each block.

Table 10.13 shows the pattern of a data set resulting from a randomized blocks design; it is a two-way table with single measurements as entries. In the example people correspond to blocks and cans to treatments. The observation x_{gj} is called the response to treatment g in block j.

The treatment mean $\bar{x}_{g.}$, estimates the population mean μ_g, for treatment g (averaged out over people). An objective may be to test the hypothesis that treatments make no difference,

H_0: $\mu_1 = \mu_2 = . . . = \mu_k$

Table 10.13 *Randomized Blocks Design*

	Blocks			
Treatments	1	2	. . .	b
1	x_{11}	x_{12}	. . .	x_{1b}
2	x_{21}	x_{22}	. . .	x_{2b}
.	.	.		.
.	.	.		.
.	.	.		.
k	x_{k1}	x_{k2}	. . .	x_{kb}

Each observation x_{gj} can be written as a sum of meaningful terms by means of the identity

$$x_{gj} = \overline{x} + (\overline{x}_{g.} - \overline{x}) + (\overline{x}_{.j} - \overline{x}) + (x_{gj} - \overline{x}_{g.} - \overline{x}_{.j} + \overline{x}) .$$

In word, the

$$\begin{pmatrix} Observed\ value\ for \\ gth\ treatment\ in \\ jth\ block \end{pmatrix} = \begin{pmatrix} overal \\ mean \end{pmatrix} + \begin{pmatrix} deviation \\ due\ to\ gth \\ treatment \end{pmatrix} + \begin{pmatrix} deviation \\ due\ to \\ jth\ block \end{pmatrix} + (residual)$$

The "residual" is

$$x_{gj} - \lfloor \overline{x} + (\overline{x}_{g.} - \overline{x}) + (\overline{x}_{.j} - \overline{x}) \rfloor,$$

which is the difference between the observation and

$$\overline{x} + (\overline{x}_{.j} - \overline{x}) + (\overline{x}_{.j} - \overline{x}),$$

obtained by taking into account the overall mean, the effect of the gth treatment, and the effect of the jth block. Algebra shows that the corresponding decomposition is true for sums of squares:

$$\sum_{g=1}^{k}\sum_{j=1}^{b}(x_{gj} - \overline{x})^2 = b\sum_{g=1}^{k}(\overline{x}_{g.} - \overline{x})^2 + k\sum_{j=1}^{b}(\overline{x}_{.j} - \overline{x})^2 + \sum_{g=1}^{k}\sum_{j=1}^{b}(x_{gj} - \overline{x}_{g.} - \overline{x}_{.j} + \overline{x})^2$$

that is,

SS(Total) = SS(Treatment) + SS(Blocks) + SS(Residuals).

The number of degrees of freedom of SS(Total) is kb - 1 = n - 1, the number of observations less 1 for the overall mean.

The number of degrees of freedom of SS(Treatments) is k - 1, the number of treatments less 1 for the overall mean.

Similarly, the number of degrees of freedom of SS(Blocks) is b - 1. There remain, as the number of degrees of freedom for SS(Residuals)

kb - 1 - (k - 1) - (b - 1) = (k - 1)(b - 1).

There is a hypothetical model behind the analysis. It is assumes that in repeated experiments the measurement for the gth treatment in the jth block would be the sum of a constant pertaining to the treatment, namely μ_g, a constant pertaining to the jth block, and a random "error" term with a variance of σ^2. The mean square for residuals,

MS(Residuals) = SS(Residuals) / (k - 1)(b - 1)

is an unbiased estimate of σ^2 regardless of whether the μ_g's differ (that is, whether there are true effects due to treatments). If there are no differences in the μ_g's,

MS(Treatments) = MS(Treatments) / (k - 1)

is an unbiased estimate of σ^2 (whether or not there are true effects due to blocks). If there are differences among the μ_g's, then MS(Treatments) will tend to be larger than σ^2. One tests H_0 by means of

F = MS(Treatments) / MS(Residuals)

When H_0 is true, F is distributed as an F-distribution based on (k - 1) numerator df and (k - 1) (b -1) df. One rejects H_0 if F is sufficiently large, that is, if F exceeds F_α. Table 10.14 is the analysis of variance table.

Table 10.14 *Analysis of variance table for randomized blocks design*

Sources of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	$b\sum_{g=1}^{k}(\bar{x}_{g.}-\bar{x})^2$	k - 1	MS(Treatments)	$\frac{MS(Treatments)}{MS(Residuals)}$
Blocks	$k\sum_{j=1}^{b}(\bar{x}_{.j}-\bar{x})^2$	b - 1	MS(Blocks)	$\frac{MS(Blocks)}{MS(Residuals)}$
Residuals	$\sum_{g=1}^{k}\sum_{j=1}^{b}(x_{gj}-\bar{x}_{g.}-\bar{x}_{.j}+\bar{x})^2$	(k -1)(b -1)	MS(Residuals)	
Total	$\sum_{g=1}^{k}\sum_{j=1}^{b}(x_{gj}-\bar{x})^2 = SS(Total)$	n - 1		

The computational formulas are

$$SS(Total)=\sum_{g=1}^{k}\sum_{j=1}^{b}x_{gj}{}^2-\frac{1}{kb}\left(\sum_{g=1}^{k}\sum_{j=1}^{b}x_{gj}\right)^2$$

$$SS(Treatments)=\frac{1}{b}\sum_{g=1}^{k}\left(\sum_{j=1}^{b}x_{gj}\right)^2-\frac{1}{kb}\left(\sum_{g=1}^{k}\sum_{j=1}^{b}x_{gj}\right)^2$$

$$SS(Blocks)=\frac{1}{k}\sum_{j=1}^{b}\left(\sum_{g=1}^{k}x_{gj}\right)^2-\frac{1}{kb}\left(\sum_{g=1}^{k}\sum_{j=1}^{b}x_{gj}\right)^2$$

$$SS(Residuals) = SS(Total) - SS(Treatments) - SS(Block)$$

For the data in Table 10.13 we have

$$\sum_{g=1}^{k}\sum_{j=1}^{b} x_{gj}{}^{2} = 6^2 + 5^2 + ... + 3^2 = 237$$

$$\frac{1}{kb}\left(\sum_{g=1}^{k}\sum_{j=1}^{b} x_{gj}\right)^2 = \frac{55^2}{15} = 201.67$$

$$SS(Total) \quad = 237 - 201.67 = 35.33$$

$$SS(Treatments) \quad = \frac{24^2 + 10^2 + 21^2}{5} - 201.67 = 223.40 - 201.67 = 21.73$$

$$SS(Blocks) \quad = \frac{14^2 + 12^2 + 12^2 + 10^2 + 7^2}{3} - 201.67 = 211 - 201.67 = 9.33$$

$$SS(Residuals) \quad = 35.33 - 21.73 = 4.27$$

The analysis of variance table is Table 10.15. From Table 4 in Appendix C, the tabulated value of $F_{.05}$ with 2 and 8 df is 4.46. Therefore, we will reject H_0 if the calculated value of F is $F > 4.46$. Since the computed value of test statistic, $F = 20.40$, exceeds 4.46, we have sufficient evidence to reject the null hypothesis of no difference in metallic taste of types of can at $\alpha = .05$.

Table 10.15 *Analysis variance table for "Metallic" scale*

Sources of variation	Sum of Squares	Degrees of freedom	Mean square	F
Cans	21.73	2	10.87	20.40
Persons	9.33	4	2.33	4.38
Residual	4.27	8	0.533	
Total	35.33	14		

The roles of cans and people can be interchanged. To test the hypothesis that there are no differences in scoring among persons (in the hypothetical population of repeated experiments), one uses the ration of MS(Blocks) to MS(Residuals) and rejects the null hypothesis if that ratio is greater than an F-value for $b - 1$ and $(k - 1)(b - 1)$ degrees of freedom. The value here of 4.38 is referred to Table 4 of Appendix C with 4 and 8 degrees of freedom, for which the 5% point is 3.84; it is barely significant.

10.9 Multiple comparisons of means and confidence regions

The F-test gives information about all means $\mu_1, \mu_2, \ldots, \mu_k$ simultaneously. In this section we consider inferences about differences of pairs of means. Instead of simply concluding that some of $\mu_1, \mu_2, \ldots, \mu_k$ are different, we may conclude that specific pairs μ_g, μ_h are different.

The variance of difference between two means, say $\bar{x}_1$ and $\bar{x}_2$, is $\sigma^2(1/n_1 + 1/n_2)$, which is estimated as $s^2(1/n_1 + 1/n_2)$. The corresponding estimated standard deviation is

$$s\sqrt{1/n_1 + 1/n_2}\ .$$

If one were interested simply in determining whether the first two population means differed, one would test the null hypothesis that $\mu_1 = \mu_2$ at significance level α by using a t-test, rejecting the null hypothesis if

$$\left|\bar{x}_1 - \bar{x}_2\right| / (s\sqrt{1/n_1 + 1/n_2}) > t_{\alpha/2}$$

where the number of degrees of freedom for the t-value is the number of degrees of freedom for s. However, now we want to consider each possible difference μ_g - μ_h; that is, we want to test all the null hypotheses

H_{gh}: $\mu_g = \mu_h$, with $g \neq h; g, h = 1, \ldots, k.$

There are $k(k$ - 1)/2 such hypotheses.

If, indeed, all the μ's were equal, so that there were no real differences, the probability that any particular one of the pair wise differences in absolute value would exceed the relevant t-value is α. Hence the probability that *at least one* of them would exceed the t-value, would be greater than α. When many differences are tested, the probability that some will appear to be "significant" is greater than the nominal significance level α when all the null hypotheses are true. How can one eliminate this false significance? It can be shown that, if m comparisons are to be made and the overall Type I error probability is to be at most α, it is sufficient to use α/m for the significance level of the individual tests. By overall Type I error we mean concluding $\mu_g \neq \mu_h$ for at least one pair g, h when actually $\mu_1 = \mu_2 = \ldots = \mu_k$.

Example 10.5 **We illustrate with Example 10.3 (Tables 10.6 and 10.9). Here s^2 = 1.867, based on 15 degrees of freedom (s = 1.366). Since all the sample sizes are 6, the value with which to compare each differences $\bar{x}_g - \bar{x}_h$ is**

$$t_{\alpha^*/2} \times (s\sqrt{1/n_1 + 1/n_2}) = t_{\alpha^*/2} \times 1.366 \times \sqrt{1/6 + 1/6} = t_{\alpha^*/2} \times 1.366 \times \sqrt{1/3} = t_{\alpha^*/2} \times .789$$

where α^* is to be the level of the individual tests.

The number of comparisons to be made for k = 3 is $k(k$ - 1)/2 = 3 = m. If we want the overall Type I error probability to be at most .03, then it suffices to choose the level α^* to be .03/3 = .01. The corresponding percentage point of Student's t-distribution with 15 degrees of freedom is $t_{.01/2}$ = $t_{.005}$ = 2.947. The value with which to compare $\bar{x}_g - \bar{x}_h$ is .789 x 2.947 = 2.33. In Table 10.6 the means are $\bar{x}_1 = 4, \bar{x}_2 = 9, \bar{x}_3 = 5.$

The difference $\bar{x}_2 - \bar{x}_1 = 9 - 4 = 5$ is significant; so is the $\bar{x}_2 - \bar{x}_3 = 9 - 5 = 4$. The difference $\bar{x}_3 - \bar{x}_1 = 5 - 4 = 1$ is not significant. The conclusion is that μ_2 is different from both μ_1 and μ_3, but μ_1 and μ_3 may be equal; Workbook 2 appears to be superior.

Confidence Regions

With confidence at least 1 - α, the following inequalities hold:

$$\bar{x}_g - \bar{x}_h - t_{\alpha^*/2} s\sqrt{1/n_g + 1/n_h} < (\mu_g - \mu_h) < \bar{x}_g - \bar{x}_h + t_{\alpha^*/2} s\sqrt{1/n_g + 1/n_h}$$

for $g \neq h; g, h = 1, \ldots, k$, if $\alpha^* = \alpha/m$ and the distribution of t is based on (n - k) degrees of freedom.

10.10 Summary

This chapter presented an extension of the methods for comparing two population means to allow for the comparison of more than two means. The completely randomized design uses independent random samples selected from each of k populations. The comparison of the population means is made by comparing the variance among the sample means, as measured by the mean square for treatments (MST), to the variation attributable to differences within the samples, as measured by the mean square for error (MSE). If the ratio of MST to MSE is large, we conclude that a difference exists between the means of at least two of the k populations.

We also presented an analysis of variance for a comparison of two or more population means using matched groups of experimental units in a randomized block design, an extension of the matched-pairs design. The design not only allows us to test for differences among the treatment means, but also enables us to test for differences among block means. By testing for differences among block means, we can determine whether blocking is effective in reducing the variation present when comparing the treatment means.

Remember that the proper application of these ANOVA techniques requires that certain assumptions are satisfied. In most applications, the assumptions will not be satisfied exactly. However, these analysis of variance procedures are flexible in the sense that slight departures from the assumptions will not significantly affect the analysis or the validity of the resulting inferences.

10.11 Exercises

10.1. A random sample of n = 500 observations were allocated to the k = 5 categories shown in the table. Suppose we want to test the null hypothesis that the category probabilities are p_1 =.1, p_2 =.1, p_3 =.5, p_4 =.1, and p_5 =.2.

		Category			Total
1	2	3	4	5	
27	62	241	69	101	500

a. Calculate the expected cell counts.

b. Find χ^2_α for α = .05.

c. State the alternative hypothesis for the test.

d. Do the data provide sufficient evidence to indicate that the null hypothesis is false?

10.2. Refer to the accompanying 2 × 3 contingency table.

		Columns			Totals
		1	2	3	
Rows	1	14	37	23	74
	2	21	32	38	91
Totals		35	69	61	165

a. Calculate the estimated expected cell counts for the contingency table.

b. Calculate the chi-square statistic for the table.

10.3. A partially completed ANOVA table for a completely randomized design is shown here.

Source	SS	df	MS	*F*
Between groups	24.7	4	—	—
Within groups	—	—	—	
Total	62.4	34		

a. Complete the ANOVA table.

b. How many treatments are involved in the experiment?

c. Do the data provide sufficient evidence to indicate a difference among the population means? Test using $\alpha = .10$.

10.4. A randomized block design was conducted to compare the mean responses for three treatments, A, B, and C, in four blocks. The data are shown in the accompanying table, followed by a partial summary ANOVA table.

	Block			
Treatment	1	2	3	4
A	3	6	1	2
B	5	7	4	6
C	2	3	2	2

Source	SS	df	MS	*F*
Treatments	23.167	—	—	—
Blocks	14.250	—	4.750	—
Residuals	—	—	.917	
Total	42.917	—		

a. Complete the ANOVA table.

b. Do the data provide sufficient evidence to indicate a difference among treatment means? Testing using $\alpha = .05$.

c. Do the data provide sufficient evidence to indicate that blocking was effective in reducing the experimental error? Testing using $\alpha = .10$.

d. What assumptions must the data satisfy to make the *F* test in parts b and c valid?

10.5. At the 5% level make the F-test of equality of population (treatment) means for the data in the table.

Treatment	Blocks		
	1	2	3
1	1	4	9
2	4	9	16
3	9	16	23

Chapter 11 Simple Linear regression and correlation

CONTENTS

11.1 Introduction: Bivariate relationships

Subject of this Chapter is to determine the relationship between variables.
In Chapter 10 we used chi-square tests of independence to determine whether a statistical relationship existed between two variables. The chi-square test tells us if there is such a relationship, but it does not tell us what the relationship is. Regression and correlation analyses will show how to determine both the nature and the strength of a relationship between two variables.
The term “regression “ was first used as a statistical concept by Sir Francis Galton. He designed the word regression as the name of the general process of predicting one variable (the height of the children) from another (the height of the parent). Later, statisticians coined the term multiple regression to describe the process by which several variables are used to predict another.
In regression analysis we shall develop an estimating equation – that is a mathematical formula that relates the known variables to the unknown variable. Then, after we have learned the pattern of this relationship we can apply correlation analysis to determine the degree to which the variables are related. *Correlation analysis tell us how well the* ***estimating equation*** *actually describes the relationship*.

Types of relationships

Regression and correlation analyses are based on the relationship or association between two or more variables.

Definition 11.1

The relationship between two random variables is known as a bivariate relationship. The known variable (or variables) is called the *independent variable*(s). The variable we are trying to predict is the *dependent variable*.

Example 11.1 A farmer may be interested in the relationship between the level of fertilizer x and the yield of potatoes y. Here the level of fertilizer x is independent variable and the yield of potatoes y is dependent variable.

Example 11.2 A medical researcher may be interested in the bivariate relationship between a patient's blood pressure x and heart rate y. Here x is independent variable and y is dependent variable.

Example 11.3 Economists might base their predictions of the annual gross national product (GDP) on the final consumption spending within the economy. Then, the final consumption spending is the independent variable, and the GDP would be the dependent variable.

In regression analysis we can have only one dependent variable in our estimating equation. However, we can use more than one independent variable. We often add independent variables in order to improve the accuracy of our prediction.

Definition 11.2

If when the independent variable x increases, the dependent variable y also increases then the relationship between x and y is *direct relationship*. In the case, the dependent variable y decreases as the independent variable x increases, we call the *relationship inverse*.

Scatter diagrams

The first step in determining whether there is a relationship between two variables is to examine the graph of the observed (or known) data, i.e. of the data points.

Definition 11.3

The graph of the data points is called a *scatter diagram* or *scatter gram*.

Example 11.4 In recent years, physicians have used the so-called diving reflex to reduce abnormally rapid heartbeats in humans by submerging the patient's face in old water. A research physician conducted an experiment to investigate the effects of various cold temperatures on the pulse rates of ten small children. The results are presented in Table 11.1.

Table 11.1
Temperature of water – Pulse rate data

Child	Temperature of Water, x^{o} F	Reduction in Pulse, y beats/minute
1	68	2
2	65	5
3	70	1
4	62	10
5	60	9
6	55	13
7	58	10
8	65	3
9	69	4
10	63	6

The scatter gram of the data set in Table 11.1 is depicted in Figure 11.1.

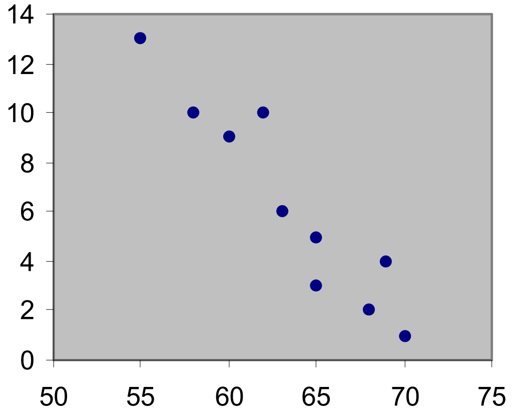


Figure 11.1 *Scatter gram for the data in Table 11.*

From the scatter gram we can visualize the relationship that exists between the two variables. As a result we can draw or “fit” a straight line through our scatter gram to represent the relationship. We have done this in Figure 11.2.

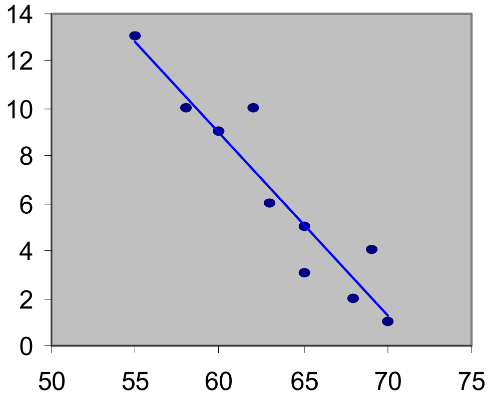


Figure 11.2 *Scatter gram with straight line representing the relationship between x and y "fitted" through it*

We see that the relationship described by the data points is well described by a straight line. Thus, we can say that it is a linear relationship. This relationship, as we see, is inverse because y decreases as x increases

Example 11.5 To model the relationship between the CO (Carbon Monoxide) ranking, y, and the nicotine content, x, of an American-made cigarette the Federal Trade commission tested a random sample of 5 cigarettes. The CO ranking and nicotine content values are given in Table 11.2

Table 11.2 *CO Ranking-Nicotine Content Data*

Cigarette	Nicotine Content, x, mgs	CO ranking, y, mgs
1	0.2	2
2	0.4	10
3	0.6	13
4	0.8	15
5	1	20

The scatter gram with straight line representing the relationship between Nicotine Content x and CO Ranking y "fitted" through it is depicted in Figure 11.3. From this we see that the relationship here is direct.

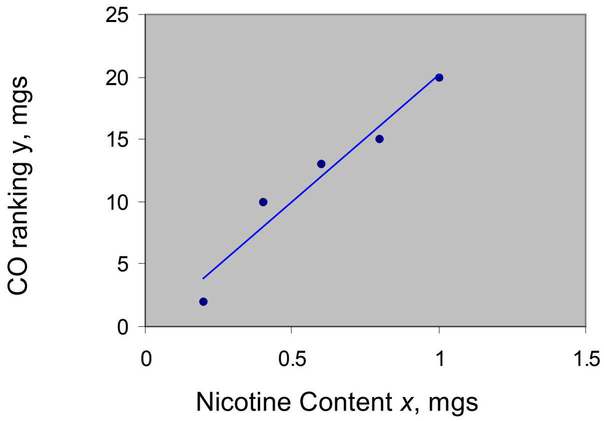


Figure 11.3 *Scatter gram with straight line representing the relationship between x and y "fitted" through it*

11.2 Simple Linear regression: Assumptions

Suppose we believe that the value of y tends to increase or decrease in a linear manner as x increases. Then we could select a model relating y to x by drawing a line which is well fitted to a given data set. Such a deterministic model – one that does not allow for errors of prediction – might be adequate if all of the data points fell on the fitted line. However, you can see that this idealistic situation will not occur for the data of Table 11.1 and 11.2. No matter how you draw a line through the points in Figure 11.2 and Figure 11.3, at least some of points will deviate substantially from the fitted line.

The solution to the proceeding problem is to construct a probabilistic model relating y to x- one that acknowledges the random variation of the data points about a line. One type of probabilistic model, a simple linear regression model, makes assumption that the mean value of y for a given value of x graphs as straight line and that points deviate about this line of means by a random amount equal to *e*, i.e.

$$y = A + B\,x + e,$$

where A and B are unknown parameters of the deterministic (nonrandom) portion of the model. If we suppose that the points deviate above or below the line of means and with expected value $E(e) = 0$ then the mean value of y is

$$y = A + B\,x.$$

Therefore, the mean value of y for a given value of x, represented by the symbol $E(y)$ graphs as straight line with y-intercept A and slope B.

A graph of the hypothetical line of means, $E(y) = A + B\,x$ is shown in Figure 11.4.

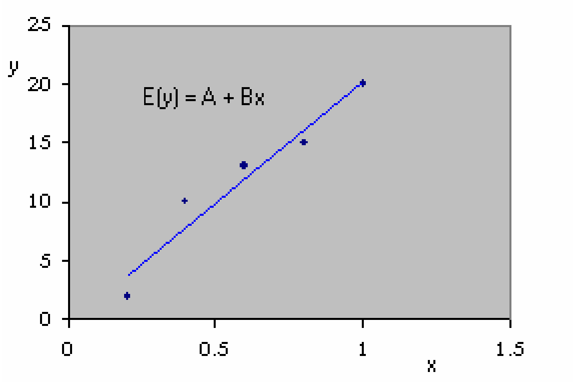


Figure 11.4 *The straight line of means*

A SIMPLE LINEAR REGRESSION MODEL

y = *A* + *B x* + e,

where

y = dependent variable (variable to be modeled – sometimes called the response variable)

x = independent variable (variable used as a predictor of *y*)

e = random error

A = *y*-intercept of the line

B = slope of the line

In order to fit a simple linear regression model to a set of data , we must find estimators for the unknown parameters *A* and *B* of the line of means *y* = *A* + *B x.* Since the sampling distributions of these estimators will depend on the probability distribution of the random error *e*, we must first make specific assumptions about its properties.

ASSUMPTIONS REQUIRED FOR A LINEAR REGRESSION MODEL

1. The mean of the probability distribution of the random error is 0, $E(e) = 0$. that is, the average of the errors over an infinitely long series of experiments is 0 for each setting of the independent variable x. this assumptionsimplies that the mean value of y, $E(y)$ for a given value of x is $y = A + B\,x.$
2. The variance of the random error is equal a constant, say σ^2, for all value of x.
3. The probability distribution of the random error is normal.
4. The errors associated with any two different observations are independent. That is, the error associated with one value of y has no effect on the errors associated with other values.

11.3 Estimating A and B: the method of least squares

The first problem of simple regression analysis is to find estimators of A and B of the regression model based on a sample data .

Suppose we have a sample of n data points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). The straight-line model for the response y in terms x is

$$y = A + B\,x + e.$$

The line of means is $E(y) = A + B\,x$ and the line fitted to the sample data is $\hat{y} = a + bx$. Thus, $\hat{y}$ is an estimator of the mean value of y and a predictor of some future value of y; and a, b are estimators of A and B, respectively.

For a given data point, say the point (x_i, y_i), the observed value of y is y_i and the predicted value of y would be

$$\hat{y}_i = a + bx_i$$

and the deviation of the ith value of y from its predicted value is

$$SSE = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2 \ .$$

The values of *a* and b that make the SSE minimum is called the *least squares estimators* of the population parameters A and B and the prediction equation $\hat{y} = a + bx$ is called the *least squares line.*

Definition 11.4

The *least squares line* is one that has a smaller than any other straight-line model.

FORMULAS FOR THE LEAST SQUARES ESTIMATORS

Slope: $b = \dfrac{SS_{xy}}{SS_{xx}}$, y-intercept: $a = \overline{y} - b\overline{x}$

where

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \quad SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 ,$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ,$$

n = sample size

Example 11.6 Refer to Example 11.5. Find the best-fitting straight line through the sample data points.

Solution By the least squares method we found the equation of the best-fitting straight line. It is $\hat{y} = -0.3 + 20.5x$. The graph of this line is shown in Figure 11.5

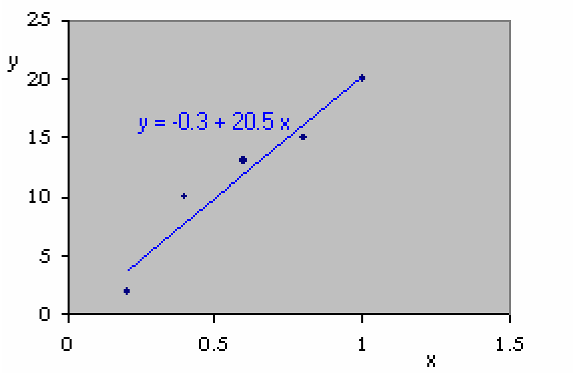


***Figure 11.5** Least squares line for Example 11.6*

11.4 Estimating σ^2

In most practical situations, the variance σ^2 of the random error e will be unknown and must be estimated from the sample data. Since σ^2 measures the variation of the y values about the regression line, it seems intuitively reasonable to estimate σ^2 by dividing the total error SSE by an appropriate number.

ESTIMATION OF σ^2

$$s^2 = \frac{SSE}{Degree\ of\ freedom\ for\ error} = \frac{SSE}{n-2}$$

where

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

From the following Theorem it is possible to prove that s^2 is an unbiased estimator of σ^2**,** that is $E(s^2) = \sigma^2$.

Theorem 11.1

Let $s^2 = \frac{SSE}{n-2}$. Then, when the assumptions of Section 11.2 are satisfied, the statistic

$\chi^2 = \frac{SSE}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2}$ has a chi-square distribution with $\nu = (n-2)$ degrees of freedom.

Usually, s is referred to as a *standard error of estimate*.

Example 11.7 Refer to Example 11.5. Estimate the value of the error variance σ^2 .

Data analysis or statistical softwares provide procedures or functions for computing the standard error of estimate *s*. For example, the function STEYX of MS-Excel gives, for the data of Example 11.5, the result s =1.816590.

Recall that the least squares line estimates the mean value of y for a given value of x. Since *s* measures the spread of distribution of y values about the least squares line, most observations will lie within 2s of the least squares line.

INTERPRETATION OF *s*, THE ESTIMATED STANDARD DEVIATION OF *e*

We expect most of the observed y values to lie within 2s of their respective least squares predicted value $\hat{y}$ **.**

11.5 Making inferences about the slope, B

In Section 11.2 we proposed the probabilistic model $y = A + B\,x + e$ for the relationship between two random variables x and y, where x is independent variable and y is dependent variable, A and B are unknown parameters, and e is a random error. Under the assumptions made on the random error e we have $E(y) = A + B\,x$. This is the *population regression line.* If we are given a sample of n data points (x_i, y_i), i =1,...,n, then by the least squares method in Section 11.3 we can find the straight line $\hat{y} = a + bx$ fitted to these sample data. This line is the *sample regression line*. It is an estimate for the population regression line. We should be able to use it to make inferences about the population regression line. In this section we shall make inferences about the slope B of the "true" regression equation that are based upon the slope b of the sample regression equation.

The theoretical background for making inferences about the slope B lies in the following properties of the least squares estimator *b:*

PROPERTIES OF THE LEAST SQUARES ESTIMATOR b

1. Under the assumptions in section 11.2, b will possess sampling distribution that is normally distributed.
2. The mean of the least squares estimator b is B, $E(b) = B$, that is, b is an unbiased estimator for B.
3. The standard deviation of the sampling distribution of b is

$$\sigma_b = \frac{\sigma}{\sqrt{SS_{xx}}} ,$$

where σ is the standard deviation of the random error *e*, $SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

We will use these results to test hypotheses about and to construct a confidence interval for the slope B of the population regression line.

Since σ is usually unknown, we use its estimator s and instead of $\sigma_b = \frac{\sigma}{\sqrt{SS_{xx}}}$ we use its estimate $s_b = \frac{s}{\sqrt{SS_{xx}}}$.

For testing hypotheses about B first we state null and alternative hypotheses:

$H_0 : B = B_0$

$H_a : B \neq B_0 \quad (or\ B < B_0\ or\ B > B_0)$

where B_0 is the hypothesized value of B.

Often, one tests the hypothesis if B = 0 or not, that is, if x does or does not contribute information for the prediction of y. The setup of our test of utility of the model is summarized in the box.

A TEST OF MODEL UTILITY

ONE-TAILED TEST

$H_0 : B = 0$

$H_a : B < 0$

(or $B > 0$)

Test statistic:

$$t = \frac{b}{s_b} = \frac{b}{s / \sqrt{SS_{xx}}}$$

Rejection region

$$t < -t_\alpha$$

(or $t > t_\alpha$),

where t_α is based on *(n - 2)* df.

TWO-TAILED TEST

$H_0 : B = 0$

$H_a : B \neq 0$

Test statistic:

$$t = \frac{b}{s_b} = \frac{b}{s / \sqrt{SS_{xx}}}$$

Rejection region

$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$,

where $t_{\alpha/2}$ is based on *(n-2)* df.

The values of t_α such that $P(t \geq t_\alpha) = \alpha$ are given in Table 7.4

Example 11.8 Refer to the nicotine-carbon monoxide ranking problem of Example 11.5. At significance level $\alpha = 0.05$, test the hypothesis that the nicotine content of a cigarette contributes useful information for the prediction of carbon monoxide ranking y, i.e. test the prediction ability of the least squares straight line model $\hat{y} = -0.3 + 20.5x$.

Solution Testing the usefulness of the model requires testing the hypothesis

$H_0 : B = 0$

$H_a : B \neq 0$

with n = 5 and $\alpha = 0.05$, the critical value based on (5 -2) = 3 df is obtained from Table 7.4

$t_{\alpha/2} = t_{0.025} = 3.182$.

Thus, we will reject H_0 if *t < -3.182* or *t > 3.182*.

In order to compute the test statistic we need the values of *b, s* and SS_{xx}. In Example 11.6 we computed b **=**20.5. In Example 11.7 we know s = 1.82 and we can compute SS_{xx} = 0.4. Hence, the test statistic is

$$t = \frac{b}{s/\sqrt{SS_{xx}}} = \frac{20.5}{1.82/\sqrt{0.4}} = 7.12$$

Since the calculated *t*-value is greater than the critical value $t_{0.025}$ = 3.182, we **reject** the null hypothesis and conclude that the slope $B \neq 0$. At the significance level α = 0.05, the sample data provide sufficient evidence to conclude that nicotine content does contribute useful information for prediction of carbon-monoxide ranking using the linear model.

Example 11.9 A consumer investigator obtained the following least squares straight line model (based on a sample on n = 100 families) relating the yearly food cost y for a family of 4 to annual income x:

$\hat{y} = 467 + 0.26x$.

In addition, the investigator computed the quantities s = 1.1, SS_{xx} = 26. Compute the observed p-value for a test to determine whether mean yearly food cost y increases as annual income x increases , i.e., whether the slope of the population regression line B is positive.

Solution The consumer investigator wants to test

$H_0 : B = 0$

$H_a : B > 0$

To compute the observed significance level (p-value) of the test we must first find the calculated value of the test statistic, t_c . Since b = 0.26, s =1.1, and SS_{xx} = 26 we have

$$t = \frac{b}{s/\sqrt{SS_{xx}}} = \frac{0.26}{1.1/\sqrt{26}} = 1.21$$

The observed significance level or p-value is given by

$P(t > t_c) = P(t > 1.21)$, where *t*-distribution is based on (n - 2) = (100 - 2) = 98 df. Since df >30 we can approximate the *t*-distribution with the *z*-distribution. Thus,

p-value = $P(t > 1.21) = P(z > 1.21)$ ≈ 0.5 – 0.3869 = 0.1131.

In order to conclude that the mean yearly food cost increases as annual income increases (B > *0)* we must tolerate $\alpha \geq 0.1131$. But it is a big risk and usually we take α = 0.05. Under this significance level we can not reject the hypothesis H_0. It means we consider *the sample result to be statistically insignificant*.

Another way to make inferences about the slope B is to estimate it using a confidence interval. This interval is formed as shown in the box.

A (1-α)100% CONFIDENCE INTERVAL FOR THE SLOPE B

$b \pm t_{\alpha/2} s_b$, where $s_b = \dfrac{s}{\sqrt{SS_{xx}}}$ and $t_{\alpha/2}$ is based on (n-2) df.

Example 11.10 Find the 95% confidence interval for B in Example 11.8.

Solution For a 95% confidence interval α = 0.05. Therefore, we need to find the value of $t_{\alpha/2} = t_{0.025}$ based on (5-2) = 3 df. In Example 11.8 we found that $t_{0.025}$ = 3.182. Also, we have b = 20.5, SS_{xx} = 0.4. Thus, a 95% confidence interval for the slope in the model relating carbon monoxide to nicotine content is

$$b \pm t_{\alpha/2}\left(\frac{s}{\sqrt{SS_{xx}}}\right) = 20.5 \pm 3.182\left(\frac{1.82}{\sqrt{0.4}}\right) = 20.5 \pm 9.16$$

Our interval estimate of the slope parameter B is then 11.34 to 29.66. Since all the values in this interval are positive, it appears that B is positive and that the mean of y, $E(y)$ increases as x increases.

Remark From the above we see the complete similarity between the t-statistic for testing hypotheses about the slope B and the t-statistic for testing hypotheses about the means of normal populations in Chapter 9 and the similarity of the corresponding confidence intervals. In each case, the general form of the test statistic is

$$t = \frac{\text{Parameter estimator} - \text{Its hypothesized mean}}{\text{Estimated standard error of the estimator}}$$

and the general form of the confidence interval is

Point estimator $\pm\, t_{\alpha/2}$ (Estimated standard error of the estimator)

11.6. Correlation analysis

Correlation analysis is the statistical tool that we can use to describe the degree to which one variable is linearly related to another. Frequently, correlation analysis is used in conjunction with regression analysis to measure how well the least squares line fits the data . Correlation analysis can also be used by itself, however, to measure the degree of association between two variables.

In this section we present two measures for describing the correlation between two variables: the coefficient of determination and the coefficient of correlation.

11.6.1 The coefficient of correlation

Definition 11.5

The Pearson product moment *coefficient of correlation* (or simply, the coefficient of correlation) *r* is a measure of the strength of the linear relationship between two variables x and y. It is computed (for a sample of *n* measurements on x and y) as follows

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} ,$$

where

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i ,$$

Some properties of the coefficient of correlation:

i) $-1 \le r \le 1$ (this follows from the Cauchy-Bunhiacopskij inequality)

ii) *r* and b (the slope of the least squares line) have the same sign

iii) A value of *r* near or equal to 0 implies little or no linear relationship between x and y. The closer *r* is to 1 or to –1, the stronger the linear relationship between x and y.

Keep in mind that the correlation coefficient *r* measures the correlation between x values and y values in the sample, and that a similar linear coefficient of correlation exists for the population from which the data points were selected. The population correlation coefficient is denoted by ρ (rho). As you might expect, ρ is estimated by the corresponding sample statistic *r*. Or, rather than estimating ρ, we might want to test the hypothesis H_0: $\rho = 0$ against H_a: $\rho \neq 0$, i.e., test the hypothesis that x contributes no information for the predicting y using the straight line model against the alternative that the two variables are at least linearly related. But it can be shown *that the null hypothesis H_0: $\rho = 0$ is equivalent to the hypothesis H_0: $B = 0$.* Therefore, we omit the test of hypothesis for linear correlation.

11.6.1 The coefficient of determination

Another way to measure the contribution of x in predicting y is to consider how much the errors of prediction of y can be reduced by using the information provided by x.

The sample coefficient of determination is develped from the relationship between two kinds of variation: the variation of the y values in a data set around:

1. The fitted regression line
2. Their own mean

The term variation in both cases is used in its usual statistical sense to mean " the sum of a group of squared deviations".

The first variation is the variation of y values around the regression line, i.e., around their predicted values. This variation is the **sum of squares for error** (**SSE)** of the regression model

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The second variation is the variation of y values around their own mean

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Definition 11.6

The coefficient of determination is

$$\boxed{\frac{SS_{yy}-SSE}{SS_{yy}}}$$

It is easy to verify that

$$r^2=\frac{SS_{yy}-SSE}{SS_{yy}}=1-\frac{SSE}{SS_{yy}},$$

where *r* is the coefficient of correlation, defined in Subsection 11.6.1.

Therefore, usually **we call *r*^{2} the coefficient of determination**.

Statisticians interpet the coefficient of determination by looking at the amount of the variation in y that is explained by the regression line. To understand this meaning of *r*^{2} consider Figure 11.6.

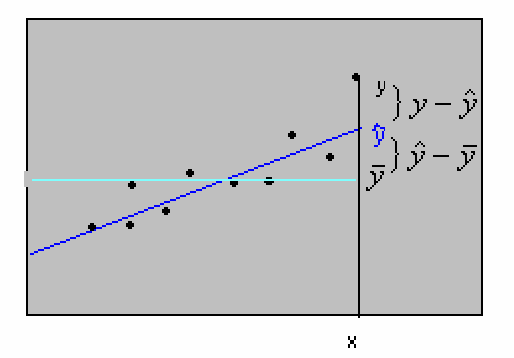


***Figure 11.6** The explained and unexplained deviations*

Here we singled out one observed value of y and showed the **total variation** of this y from its mean $\bar{y}$, $y-\bar{y}$, the **unexplained deviation** $y-\hat{y}$ and the remaining **explained deviation** $\hat{y}-\bar{y}$. Now consider a whole set of observed y values instead of only one value. **The total variation**, i.e., the sum of squared deviations of these points from their mean would be

$$SS_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

The **unexplained portion of the total variation** of these points from the regression line is

$$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

The **explained portion of the total variation** is

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

It is true that

Total variation = Explained variation + Unexplained variation.

Therefore,

$$r^2 = \frac{Explained \quad \text{variation}}{Total \quad \text{variation}}$$

PRACTICAL INTERPRETATION OF THE COEFFICIENT OF DETERMINATION, *r*²

About 100(r^2) % of the total sum of squares of deviations of the sample y-values about their mean $\bar{y}$ can be explained by (or attributed to) using x to predict y in the straight-line model.

Example 11.11 Refer to Example 11.5. Calculate the coefficient of determination for the nicotine content-carbon monoxide ranking and interpret its value.

Solution By the formulas given in this section we found r^2 = 0.9444. We interpret this value as follows: The use of nicotine content, x, to predict carbon monoxide ranking, y, with the least squares line

$$\hat{y} = -0.3 + 20.5x$$

accounts for approximately 94% of the total sum of squares of deviations of the five sample CO rankings about their mean. That is, we can reduce the total sum of squares of our prediction errors by more than 94% by using the least squares equation instead of $\bar{y}$.

11.7 Using the model for estimation and prediction

The most common uses of a probabilistic model can be divided into two categories:

1) The use of the model for estimating the mean value of y, $E(y)$, for a specific value of x
2) The second use of the model entails predicting a particular y value for a given x value.

In case 1) we are attempting to estimate the mean result of a very large number of experiments at the given x value. In case 2) we are trying to predict the outcome of a single experiment at the given x value.

The difference in these two model uses lies in the relative accuracy of the estimate and the prediction. These accuracies are best measured by the repeated sampling errors of the least squares line when it is used as estimator and as a predictor, respectively. These errors are given in the next box.

<table>
<tr><th colspan="2">SAMPLING ERRORS FOR THE ESTIMATOR OF THE MEAN AND THE PREDICTOR OF AN INDIVIDUAL y</th></tr>
<tr><td>The standard deviation of the sampling distribution of the estimator $\hat{y}$ of the mean value of y at a fixed x is
$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$</td><td>The standard deviation of the prediction error for the predictor $\hat{y}$ of an individual y-value at a fixed x is
$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}$</td></tr>
<tr><td colspan="2">where σ is the square root of σ^2, the variance of the random error (see Section 11.2)</td></tr>
</table>

The true value of σ will rarely be known. Thus, we estimate σ by s and calculate the estimation and prediction intervals as follows

A (1-α)100% CONFIDENCE INTERVAL FOR THE MEAN VALUE OF y FOR $x = x_p$	**A (1-α)100% CONFIDENCE INTERVAL FOR AN INDIVIDUAL y FOR $x = x_p$**
$\hat{y} \pm t_{\alpha/2}(Estimate\ std\ of\ \hat{y})$ or $\hat{y} \pm t_{\alpha/2}.s.\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$ where $t_{\alpha/2}$ is based on (n-2) df	$\hat{y} \pm t_{\alpha/2}[Estimate\ std\ of\ (y-\hat{y})]$ or $\hat{y} \pm t_{\alpha/2}.s.\sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$ where $t_{\alpha/2}$ is based on (n-2) df

Example 11.12 Find a 95% confidence interval for the mean carbon monoxide ranking of all cigarettes that have a nicotine content of 0.4 milligram. Also, find a 95% prediction interval for a particular cigarette if its nicotine content is 0.4 mg.

Solution For a nicotine content of 0.4 mg, x_p = 0.4 and the confidence interval for the mean of y is calculated by the formula in left of the above box with s = 1.82, n = 5, df = n - 2 = 5 - 2 = 3, $t_{0.025}$ = 3.182 $\hat{y} = -0.3 + 20.5x_p = -0.3 + 20.5 * 0.4 = 7.9$, SS_{xx} = 0.4. Hence, we obtain the confidence interval (7.9 ± 3.17).

Also, by the formula in the right cell we obtain the 95% prediction interval for a particular cigarette with nicotine content of 0.4 mg as (7.9 ± 6.60).

From the Example 11.12 it is important note that the prediction interval for the carbon monoxide ranking of an individual cigarette is wider than corresponding confidence interval for the mean carbon monoxide ranking. By examining the formulas for the two intervals, we can see that this will always be true.

Additionally, over the range of sample data, the width of both intervals increase as the value of x gets further from $\bar{x}$ (see Figure 11.7).

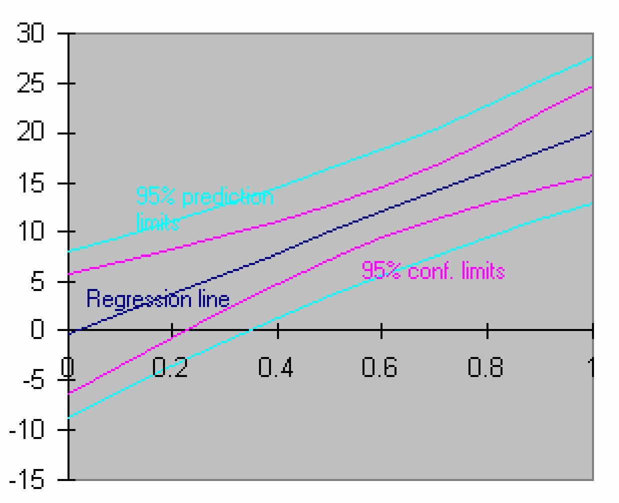


Figure 11.7 *Comparison of 95% confidence interval and prediction interval*

11.8. Simple Linear Regression: An Example

In the previous sections we have presented the basic elements necessary to fit and use a straight-line regression model. In this section we will assemble these elements by applying them to an example.

Example 11.13 The international rice research institute in the Philippines wants to relate the grain yield of rice varieties, y, to the tiller number, x . They conducted experiments for some rice varieties and tillers. Below there are the results obtained for the rice variety Milfor 6

Table 11.3 *The grain yield of rice, y, for the tiller number, x*

Grain Yield, kg/ha	**Tillers, no./m^2**
4,862	160
5,244	175
5,128	192
5,052	195
5,298	238
5,410	240
5,234	252
5,608	282

Step 1 Suppose that the assumptions listed in Section 11.2 are satisfied, we hypothesize a straight line probabilistic model for the relationship between the grain yield, y, and the tillers, x

$$y = A + B\,x + e.$$

Step 2 Use the sample data to find the least squares line. For the purpose we make calculations:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \ ,$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$b = \frac{SS_{xy}}{SS_{xx}}\ , \ a = \overline{y} - b\overline{x}$$

for the data. As a result, we obtain the least squares line

$$\hat{y} = 4242 + 4.56x$$

The scattergram for the data and the least squares line fitted to the data are depicted in Figure 11.8.

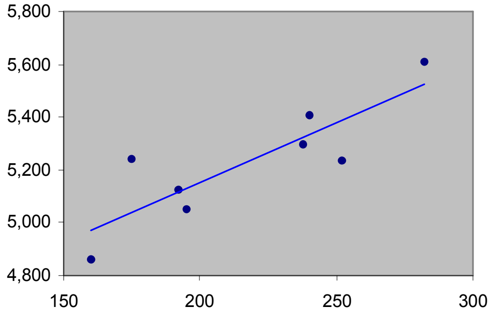


Figure 11.8 *Simple linear model relating Grain Yield to Tiller Number*

Step 3 Compute an estimator, s^2, for the variance σ^2 of the random error e :

$$s^2 = \frac{SSE}{n-2}$$

where

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

The result of computations gives s^2 = 16,229.66, *s* = 127.39. The value of s implies that most of the observed 8 values will fall within 2*s* = 254.78 of their respective predicted values.

Step 4 Check the utility of the hypothesized model, that is, whether x really contributes information for the prediction of y using the straight-line model. First test the hypothesis that the slope B is 0, i.e., there is no linear relationship between the grain yield, y, and the tillers, x. We test:

$$H_0 : B = 0$$

$$H_a : B \neq 0$$

Test statistic:

$$t = \frac{b}{s_b} = \frac{b}{s / \sqrt{SS_{xx}}}$$

For the significance level α = 0.05, we will reject H_0 if $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$,

where $t_{\alpha/2}$ is based on (n-2) = (8 – 2) = 6 df. On this df we find $t_{0.025}$ = 2.447,

$$t = \frac{4.56}{127.39 / \sqrt{125415}} = 4.004 .$$

This *t*-value is greater than $t_{0.025}$. Thus, we reject the hypothesis B = 0.

Next, we obtain additional information about the relationship by forming a confidence interval for the slope B. A 95% confidence interval is

$$b \pm t_{\alpha/2} \left(\frac{s}{\sqrt{SS_{xx}}} \right) = 4.56 \pm 2.447 \left(\frac{127.39}{\sqrt{12541.5}} \right) = 4.56 \pm 2.78 .$$

It is the interval (1.78, 7.34).

Another measure of the utility of the model is the coefficient of correlation

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}\text{ , where } SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

Computations give r = 0.853.

The high correlation confirms our conclusion that B differs from 0. It appears that the grain yield and tillers are rather highly correlated.

The coefficient of determination is r^2 = 0.7277, which implies that 72.77% of the total variation is explained by the tillers.

Step 5 Use the least squares model:

Suppose the researchers want to predict the grain yield if the tillers are 210 per m^2, i.e., x_p =210. The predicted value is

$$\hat{y} = 4242 + 4.56x_p = 4242 + 4.56 * 210 = 5199.6\text{.}$$

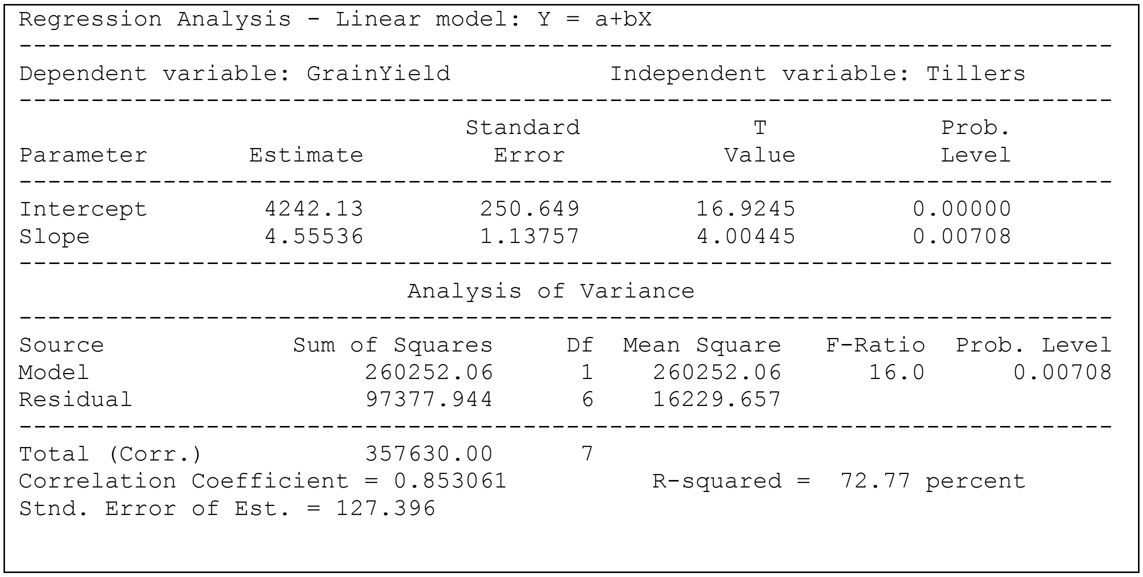
If we want a 95% prediction interval, we calculate

$$\hat{y} \pm t_{\alpha/2}.s.\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} = 5199.6 \pm 2.447 * 127.39\sqrt{1 + \frac{1}{8} + \frac{(210 - 26.75)^2}{12541.5}}$$

$$= 5199 \pm 331.18 = (4867.82, 5530.18)$$

Thus, the model yields a 95% prediction interval for the grain yield for the given value 210 of tillers from 4867.82 kg/ha to 5530.18 kg/ha.

Below we include the STATGRAPHICS printout for this example.



```
Regression Analysis - Linear model: Y = a+bX
-----------------------------------------------------------------------------
Dependent variable: GrainYield          Independent variable: Tillers
-----------------------------------------------------------------------------
                               Standard             T             Prob.
Parameter       Estimate         Error            Value           Level
-----------------------------------------------------------------------------
Intercept        4242.13        250.649          16.9245         0.00000
Slope            4.55536        1.13757          4.00445         0.00708
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square   F-Ratio  Prob. Level
Model                    260252.06      1    260252.06      16.0      0.00708
Residual                 97377.944      6    16229.657
-----------------------------------------------------------------------------
Total (Corr.)            357630.00      7
Correlation Coefficient = 0.853061          R-squared =  72.77 percent
Stnd. Error of Est. = 127.396
```

Figure 11.9 *STATGRAPHICS printout for Example 11.13*

11.9 Summary

In this chapter we have introduced bivariate relationships and showed how to compute the coefficient of correlation, r , a measure of the strength of the linear relationship between two variables. We have also presented the method of least squares for fitting a prediction equation to a data set. This procedure, along with associated statistical tests and estimations, is called a regression analysis. The steps that we follow in the simple linear regression analysis are:
To hypothesize a probabilistic straight-line model $y=A + Bx + e$.
To make assumptions on the random error component *e.*
To use the method of least squares to estimate the unknown parameters in the deterministic component, $y=A + Bx$.
To assess the utility of the hypothesized model. Included here are making inferences about the slope B, calculating the coefficient of correlation r and the coefficient of determination r^2.
If we are satisfied with the model we used it to estimate the mean y value, $E(y)$, for a given x value and to predict an individual y value for a specific x value

11.10 Exercises

1. Consider the seven data points in the table

x	-5	-3	-1	0	1	3	5
y	0.8	1.1	2.5	3.1	5.0	4.7	6.2

   a) Construct a scatter diagram for the data. After examining the scattergram, do you think that x and y are correlated? If correlation is present, is it positive or negative?

   b) Find the correlation coefficient r and interpret its value.

   c) Find the least squares prediction equation.

   d) Calculate SSE for the data and calculate s^2 and s.

   e) Test the null hypothesis that the slope B = 0 against the alternative hypothesis that $B \neq 0$. Use α = 0.05.

   f) Find a 90% confidence interval for the slope B.

2. In fitting a least squares line to n = 22 data points, suppose you computed the following quantities:

   SS_{xx} = 25 SS_{yy} = 17 SS_{xy} = 20

   $\bar{x} = 2$ $\bar{y} = 3$

   a) Find the least squares line.
   b) Calculate SSE.
   b) Calculate s^2 .
   d) Find a 95% confidence interval for the mean value of y when x = 1.
   e) Find a 95% prediction interval for y when x = 1.
   f)) Find a 95% confidence interval for the mean value of y when x = 0.

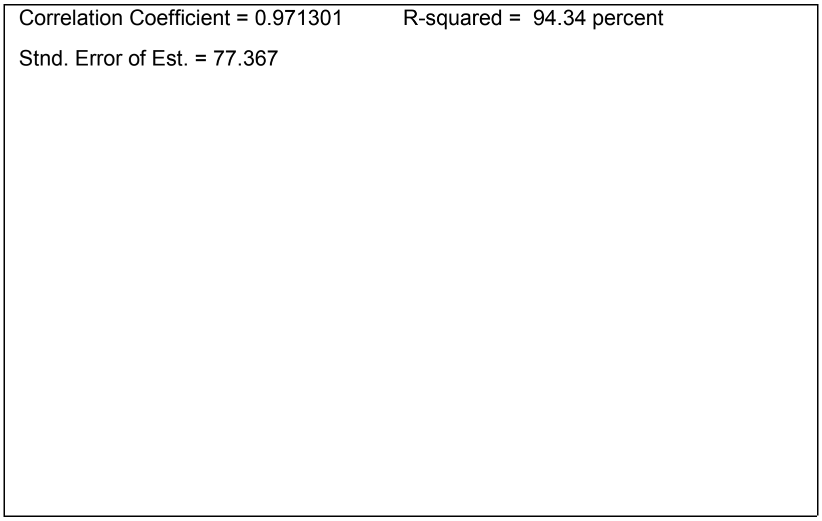
3. A study was conducted to examine the inhibiting properties of the sodium salts of phosphoric acid on the corrosion of iron. The data shown in the table provide a measure of corrosion of Armco iron in tap water containing various concentrations of $NaPO_4$ inhibitor:

Concentration of $NaPO_4$, x, parts per million	**Measure of corrosion rate, y**	**Concentration of $NaPO_4$,** x, parts per million	**Measure of corrosion rate, y**
2.50	7.68	26.20	0.93
5.03	6.95	33.00	0.72
7.60	6.30	40.00	0.68
11.60	5.75	50.00	0.65
13.00	5.01	55.00	0.56
19.60	1.43		

a) Construct a scatter diagram for the data .
b) Fit the linear model $y = A + B x + e$ to the data.
c) Does the model of part b) provide an adequate fit? Test using $\alpha = 0.05$.
d) Construct a 95% confidence interval for the mean corrosion rate of iron in tape water in which the concentration of $NaPO_4$ is 20 parts per milllion.

4. For the relationship between the variables x and y one uses a linear model and for some data collected STATGRAPHICS gives the following printout

```
Regression Analysis - Linear model: Y = a+bX
--------------------------------------------------------------------------------
Dependent variable: ELECTRIC.Y              Independent variable: ELECTRIC.X
--------------------------------------------------------------------------------
                          Standard          T          Prob.
Parameter     Estimate      Error         Value        Level
--------------------------------------------------------------------------------
Intercept      279.763     116.445       2.40252      0.04301
Slope         0.720119    0.0623473      11.5501      0.00000
--------------------------------------------------------------------------------
                     Analysis of Variance
-------------------------------------------------------------------------------
Source          Sum of Squares   Df   Mean Square    F-Ratio    Prob. Level
Model              798516.89      1    798516.89      133.4       0.00000
Residual           47885.214      8      5985.652
--------------------------------------------------------------------------------
Total (Corr.)      846402.10      9
```



```
Correlation Coefficient = 0.971301          R-squared =  94.34 percent
Stnd. Error of Est. = 77.367
```

Figure 11.10 *STATGRAPHICS printout for Exercise 11.4*

.

a) Identify the least squares model fitted to the data.

b) What are the values of SSE and s^2 for the data?

c) Perform a test of model adequacy. Use α = 0.05.

Chapter 12 Multiple regression

CONTENTS

12.1. Introduction: the general linear model

The models for a multiple regression analysis are similar to simple regression model except that they contain more terms.

Example 12.1 The researchers in the international rice research institute suppose that Grain Yield , y, relates to Plant Height, x_1, and Tiller Number, $x_{2,}$ by the linear model

$E(y) = B_0 + B_1x_1 + B_2 x_2$.

Example 12.2 Suppose we think that the mean time $E(y)$ required to perform a data-processing job increases as the computer utilization increases and that relationship is curvilinear. Instead of using the straight line model $E(y) = A + Bx_1$ to model the relationship, we might use the quadratic model $E(y) = A + B_1x_1 + B_2x_1^2$, where x_1 is a variable measures computer utilization.

A quadratic model often referred to as a second-order linear model in contrast to a straight line or first-order model.

If, in addition, we think that the mean time required to process a job is also related to the size x_2 of the job, we could include x_2 in the model. For example, the first-order model in this case is

$$E(y) = B_0 + B_1x_1 + B_2x_2$$

and the second-order model is

$$E(y) = B_0 + B_1x_1 + B_2x_2 + B_3 x_1x_2 + B_4 x_1^2 + B_5 x_2^2.$$

All the models that we have written so far are called linear models, because $E(y)$ is a linear function of the unknown parameters B_0, B_1, B_2, ...

The model

$$E(y) = A\ e^{-Bx}$$

is not a linear model because $E(y)$ is not a linear function of the unknown model parameters A and B.

Note that by introducing new variables, second-order models may be written in the form of first-order models. For example, putting $x_2 = x_1^2$, the second-order model

$$E(y) = B_0 + B_1x_1 + B_2x_1^2$$

becomes the first-order model

$$E(y) = B_0 + B_1x_1 + B_2x_2.$$

Therefore, in the future we consider only multiple first-order regression model.

THE GENERAL MULTIPLE LINEAR MODEL

$y = B_0 + B_1x_1 + ... + B_kx_k + e$,

where

y = dependent variable (variable to be modeled – sometimes called the response variable)

$x_1, x_2, ..., x_k$ = independent variable (variable used as a predictor of y)

e = random error

B_i determines the contribution of the independent variable x_i

12.2 Model assumptions

ASSUMPTIONS REQUIRED FOR A MULTIPLE LINEAR REGRESSION MODEL

1. $y = B_0 + B_1x_1 + ... + B_kx_k + e$,
   where e is random error.
2. For any given set of values $x_1, x_2, ..., x_k$, *the* random *error e* has a normal probability distribution with the mean equal 0 and variance equal σ^2.
3. The random errors are independent.

12.3 Fitting the model: the method of least squares

The method of fitting a multiple regression model is identical to that of fitting the straight-line model.

Suppose we are given the sample data that are presented in Table 12.1.

Table 12.1

DATA POINT	*Y* VALUE	x_1	x_2	...	x_k
1	y_1	x_{11}	x_{21}	...	x_{k1}
2	y_2	x_{12}	x_{22}	...	x_{k2}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	y_n	x_{1n}	x_{2n}		x_{kn}

We will use the method of least squares and choose estimates of B_0, B_1, B_2,..., B_k that minimize

$$SSE = \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 = \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} ... + b_k x_{ki})]^2$$

In order to briefly write the solution of the least squares problem we introduce the matrix notations

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \mathrm{K} & x_{k1} \\ 1 & x_{12} & x_{22} & \mathrm{K} & x_{k2} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 1 & x_{1n} & x_{2n} & \mathrm{K} & x_{kn} \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \mathrm{M} \\ b_k \end{bmatrix},$$

Then we can write the least squares equations in matrix form as

THE LEAST SQUARES MATRIX EQUATION

$(X'X\,)b = X'Y,$

where X' is the transpose of X

. The solution of the least squares equations therefore is

LEAST SQUARES SOLUTION

$b = (X'X)^{-1}XY$.

Example 12.3 Refer to Example 12.1 relating Grain Yield , y, to Plant Height, x_1, and Tiller Number, $x_{2,}$ by the linear model

$E(y) = B_0 + B_1x_1 + B_2\, x_2$.

Find the least squares estimates of B_0, B_1, B_2. The data are shown in Table 12.2

Table 12.2 *Data for Grain Yield Study*

VARIETY NUMBER	GRAIN YIELD, kg/ha (y)	PLANT HEIGHT, cm (x_1)	TILLER, no./hill (x_2)
1	5755	110.5	14.5
2	5939	105.4	16.0
3	6010	118.1	14.6
4	6545	104.5	18.2
5	6730	93.6	15.4
6	6750	84.1	17.6
7	6899	77.8	17.9
8	7862	75.6	19.4

Solution The Y, X and b are shown below

$$Y = \begin{bmatrix} 5755 \\ 5939 \\ 6010 \\ 6545 \\ 6730 \\ 6750 \\ 6899 \\ 7862 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 110.5 & 14.5 \\ 1 & 105.4 & 16.0 \\ 1 & 118.1 & 14.6 \\ 1 & 104.5 & 18.2 \\ 1 & 93.6 & 15.4 \\ 1 & 84.1 & 17.6 \\ 1 & 77.8 & 17.9 \\ 1 & 75.6 & 19.4 \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix},$$

After calculations, finally, we obtain b = (6335.59, -23.75, 150.31)'.

Thus, the prediction equation is

y = 6335.59 -23.75 x_1 + 150.31 $x_{2.}$

Below we include the STATGRAPHICS printout for this example.

```
Model fitting results for: GRAIN.Y
--------------------------------------------------------------------------------------------------------
Independent variable            coefficient     std. error          t-value         sig. level
--------------------------------------------------------------------------------------------------------
CONSTANT                        6335.596495    2942.930958         2.1528          0.0839
GRAIN.X1                         -23.748104      12.895492        -1.8416           0.1249
GRAIN.X2                         150.312641     112.069368         1.3412          0.2375
--------------------------------------------------------------------------------------------------------
```

```
R-SQ. (ADJ.) = 0.7474    SE=     340.427774     MAE=     248.149078      DurbWat=  2.337
Previously:   0.0000                0.000000                0.000000               0.000
8 observations fitted, forecast(s) computed for 0 missing val. of dep. var.
```

12.4 Estimating σ^2

We recall that the variances of the estimators of all the B parameters and of $\hat{y}$ will depend on the value of σ^2, the variance of the random error e that appears in the linear model. Since σ^2 will rarely be known in advance, we must use the sample data to estimate its value.

ESTIMATOR OF σ^2, THE VARIANCE OF *e* IN A MULTIPLE REGRESSION MODEL

$$s^2 = \frac{SSE}{Degree\ of\ freedom\ for\ error} = \frac{SSE}{n - \text{Number of } B \text{ parameters in model}}$$

where

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

It can be proved that s^2 is an unbiased estimator of σ^2**,** that is $E(s^2) = \sigma^2$.
Notice that in softwares SSE often is referred to as Sum of Squares for Error and s^2 is refereed to as Mean Squares for Error. For example, for the data for Grain Yield Study in **Table 12.2** the STATGRAPHICS printout is following

```
Analysis of Variance for the Full Regression
-----------------------------------------------------------------------------
Source                    Sum of Squares     DF      Mean Square      F-Ratio    P-value
-----------------------------------------------------------------------------
Model                           2632048.      2         1316024.      11.3557     0.0138
Error                            579455.      5          115891.
-----------------------------------------------------------------------------
Total (Corr.)                   3211504.      7

R-squared = 0.819569                                 Stnd. error of est. = 340.428
R-squared (Adj. for d.f.) = 0.747396             Durbin-Watson statistic = 2.33739
```

We see on this printout that SSE = 579455 and s^2 = 115891.

12.5 Estimating and testing hypotheses about the B parameters

12.5.1 Properties of the sampling distributions of $b_0, b_1, ..., b_k$

Before making inferences about the B parameters of the multiple linear model we provide some properties of the least squares estimators b , which serve the theoretical background for estimating and testing hypotheses about B.

From Section 12.3 we know that the least squares estimators b are computed by the formula $b = (X'X)^{-1}XY$. Now, we can rewrite b in the form

$b = [(X'X)^{-1}X]Y$.

From this form we see that the components of b: $b_0, b_1, ..., b_k$ are linear functions of n normally distributed random variables $y_1, y_2,..., y_n$. Therefore, b_i (i =0,1, ..., k) has a normal sampling distribution.

One showed that the least squares estimators provide unbiased estimators of $B_0, B_{1, ...,} B_k$, that is, $E(b_i) = B_i$ (i = 0,1, ..., k).

The standard errors and covariances of the estimators are defined by the elements of the matrix $(X'X)^{-1}$. Thus, if we denote

$$(X'X)^{-1} = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{ok} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ c_{20} & c_{21} & \cdots & c_{2k} \\ \vdots & \vdots & & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{bmatrix},$$

then the standard deviation of the sampling distributions of $b_0, b_1, ..., b_k$ are

$$\sigma_{b_i} = \sigma\sqrt{c_{ii}} \quad (i = 0, 1, ..., k)$$

where σ is the standard deviation of the random error e.

The properties of the sampling distributions of the least squares estimators are summarized in the box.

THEOREM 12.1 (properties of the sampling distributions of $b_0, b_1, ..., b_k$)

The sampling distribution of b_i (i = 0, 1,..., k) is normal with:

mean $E(b_i) = B_i$, variance $V(b_i) = c_{ii}$,

standard deviation: $\sigma_{b_i} = \sigma\sqrt{c_{ii}} \quad (i = 0, 1, ..., k)$

The covariance of two parameter estimators is equal to

$$Cov(b_i, b_j) = c_{ij}\sigma^2 \ (i \neq j) \ .$$

12.5.2 Estimating and testing hypotheses about the *B* parameters

A (1-α)100% confidence interval for a model parameter B_i (i = 0, 1,..., k) can be constructed using the t statistic

$$t = \frac{b_i - B_i}{s_{b_i}} = \frac{b_i - B_i}{s\sqrt{c_{ii}}}$$

where s is an estimate of σ.

A (1-α)100% CONFIDENCE INTERVAL FOR B_i

$b_i \pm t_{\alpha/2}$ (Estimated standard error of b_i) or

$$b_i \pm t_{\alpha/2} s\sqrt{c_{ii}}$$

where $t_{\alpha/2}$ is based on [n – (k+1)] df.

Similarly, the test statistic for testing the null hypothesis H_0: B_i = 0 is

$$t = \frac{b_i}{\text{Estimated standard error of } b_i}$$

The test is summarized in the box:

TEST OF AN INDIVIDUAL PARAMETER COEFFICIENT IN THE MULTIPLE REGRESSION MODEL $y = B_0 + B_1x_1 + \ldots + B_kx_k + e$,

ONE-TAILED TEST

$H_0 : B_i = 0$

$H_a : B_i < 0$

(or $B_i > 0$)

Test statistic:

$$t = \frac{b_i}{s_{b_i}} = \frac{b_i}{s\sqrt{c_{ii}}}$$

Rejection region

$$t < -t_\alpha$$

(or $t > t_\alpha$)

where $t_{\alpha/2}$ is based on [n- (k+1)] df,
n = number of observations,
k= number of independent variables in the model

TWO-TAILED TEST

$H_0 : B_i = 0$

$H_a : B_i \neq 0$

Test statistic:

$$t = \frac{b_i}{s_{b_i}} = \frac{b_i}{s\sqrt{c_{ii}}}$$

Rejection region

$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$,

where $t_{\alpha/2}$ is based [n- (k+1)] df,
n = number of observations,
k= number of independent variables in the model

The values of t_α such that $P(t \geq t_\alpha) = \alpha$ are given in Table 7.4

Example 12.4 An electrical utility company wants to predict the monthly power usage of a home as a function of the size of the home based on the model

$y = B_0 + B_1x + B_2x^2 + e$.

Data are shown in Table 12.3.

***Table 12.3** Data for Power Usage Study*

SIZE OF HOME *x*, square feet	MONTHY USAGE *y*, kilowatt-hours
1290	1182
1350	1172
1470	1264
1600	1493
1710	1571
1840	1711
1980	1804
2230	1840
2400	1956
2390	1954

a. Find the least squares estimators of B_0, B_1, B_2.
b. Compute the estimated standard error for b_1.
c. Compute the value of the test statistic for testing H_0: $B_2 = 0$.
d. Test H_0: $B_2 = 0$ against H_a: $B_2 \neq 0$. State your conclusions.

Solution We use computer with the software STATGRAPHICS to do this example. Below is a part of the printout of the procedure " Multiple regression ".

```
                    Model fitting results for: ELECTRIC.Y
--------------------------------------------------------------------------------
Independent variable                coefficient  std. error     t-value   sig.level
--------------------------------------------------------------------------------
CONSTANT                           -1303.382558  415.209833     -3.1391      0.0164
ELECTRIC.X                             2.497984     0.46109      5.4176      0.0010
ELECTRIC.X * ELECTRIC.X               -0.000477    0.000123     -3.8687      0.0061
--------------------------------------------------------------------------------
R-SQ. (ADJ.) = 0.9768  SE=      46.689335  MAE=      32.230298  DurbWat=  2.094
Previously:    0.9768           46.689335            32.230298            2.094
10 observations fitted, forecast(s) computed for 0 missing val. of dep. var.
```

Figure 12.1 *STATGRAPHICS printout for Example 12.4*

From the printout we see that

a. The least squares model are y = -1303.382558 + 2.497884 x – 0.000477 x^2.

b. The estimated standard error for b_1 is 0.461069 (in std.error column)

c. The value of the test statistic for testing H_0: $B_2 = 0$ is t = –3.8687.

d. At significance level $\alpha = 0.05$, for df = [10 – (2+1)] =7 we have $t_{\alpha/2} = 2.365$. Therefore, we will reject H_0: $B_2 = 0$ if t < -2.365 or t >2.365. Since the observed value of t = –3.8687 is less than -2.365, we reject H_0, that is, x^2 contributes information for the prediction of y.

Below we include also a printout from SPSS for the **Example 12.4.**

Coefficients							
Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for *B*	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	-1303.383	415.210	-3.139	.016	-2285.196	-321.570
	X	2.498	.461	5.418	.001	1.408	3.588
	X2	-4.768E-04	.000	-3.869	.006	-.001	.000

Figure 12.2 *A part of SPSS printout for Example 12.4*

12.6. Checking the utility of a model

Conducting t-tests on each B parameter in a model is not a good way to determine whether a model is contributing information for the prediction of y. If we were to conduct a series of t-tests to determine whether the individual variables are contributing to the predictive relationship . it is very likely that we would make one or more errors in deciding which terms to retain in the model and which to exclude.

To test the utility of a multiple regression model, we will need a global test (one that encompasses all the B parameters). We would like to find some statistical quantity that measures how well the model fits the data.

We begin with the easier problem – finding a measure of how well a linear model fits a set of data. For this we use the multiple regression equivalent of r^2, the coefficient of determination for the straight line model (Chapter 11).

Definition 12.1

The multiple coefficient of determination R^2 is defined as

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$

where

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and $\hat{y}_i$ is the predicted value of y_i for the multiple regression model.

From the definition we see that R^2 = 0 implies a complete lack of fit of the model to the data, , R^2 = 1 implies a perfect fit with the model passing through every data point. In general, the larger the value of R^2, the better the model fits the data.

R^2 is a sample statistic that tells how well the model fits the data , and thereby represents a measure of the utility of the entire model . It can be used to make inferences about the utility of the model for predicting y values for specific settings of the independent variables.

TESTING THE OVERALL UTILITY OF THE MODEL

$$E(y) = B_0 + B_1x_1 + \ldots + B_kx_k$$

H_0 : $B_1 = B_2 = \ldots = B_k = 0$ (Null hypothesis: y doesn't depend on any x_i)
H_a : At least one $B_i \neq 0$ (Alternative hypothesis: y depends an at least one of the x_i's.

Test statistic:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{\text{Mean Square for Model}}{\text{Mean Square for Error}} = \frac{SS(\text{Model})/k}{SSE/[n-(k+1)]}$$

Rejection region: $F > F_\alpha$, where F_α is value that locate area α in the upper tail of the F-distribution with $\nu_1 = k$ and $\nu_2 = n - (k+1)$,
n = Number of observations, k = Number of parameters in the model (excluding B_0)
R^2 = Multiple coefficient of determination.

Example 12.5 Refer to Example 12.4. Test to determine whether the model contributes information for the prediction of the monthly power usage.

Solution For the electrical usage example, n = 10, k = 2 and n – (k+1) = 7. At the significance level α = 0.05 we will reject H_0 : $B_1 = B_2 = 0$ if $F > F_{0.05}$. where ν_1 = 2 and ν_2 = 7, or $F >$ 4.74. From the computer printout (see Figure 12.3) we find that the computed F is 190.638. Since this value greatly exceeds 4.74 we reject H_0 and conclude that at least one of the model coefficients B_1 and B_2 is nonzero. Therefore, this F test indicates that the second order model $y = B_0 + B_1x + B_2x^2 + e$, is useful for predicting electrical usage.

```
Analysis of Variance for the Full Regression
-----------------------------------------------------------------------------------------------------------------
Source              Sum of Squares    DF    Mean Square      F-Ratio   P-value
-----------------------------------------------------------------------------------------------------------------
Model                    831143.        2      415571.         190.638    0.0000
Error                    15259.3        7       2179.89
-----------------------------------------------------------------------------------------------------------------
Total (Corr.)            846402.        9

R-squared = 0.981972                              Stnd. error of est. = 46.6893
R-squared (Adj. for d.f.) = 0.976821           Durbin-Watson statistic = 2.09356
```

Figure 12.3 *STATGRAPHICS Printout for Electrical Usage Example*

Example 12.6 Refer to Example 12.3. test the utility of the model $E(y) = A + B_1x_1 + B_2x_2$.

Solution From the SPSS Printout (Figure 12.4) we see that the F value is 11.356 and the corresponding observed significance level is 0.014. Thus, at the significance level greater than 0.014 we reject the null hypothesis, and conclude that the linear model $E(y) = A + B_1x_1 + B_2x_2$ is useful for prediction of the grain yield.

ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	2632048.153	2	1316024.076	11.356	.014
Residual	579455.347	5	115891.069		
Total	3211503.500	7			

Figure 12.4 *SPSS Printout for Grain Yield Example*

12.7. Using the model for estimating and prediction

After checking the utility of the linear model and finding it to be useful for prediction and estimation, we may decide use it for those purposes. Our methods for prediction and estimation using any general model are identical to those discussed in Section 11.7 for the simple straight-line model. We will use the model to form a confidence interval for the mean $E(y)$ for a given value x^* of x, or a prediction interval for a future value of y for a specific x^*.

The procedure for forming a confidence interval for $E(y)$ is shown in following box.

A (1-α)100% CONFIDENCE INTERVAL FOR E(y)

$$\hat{y} \pm t_{\alpha/2} s\sqrt{(x^*)'(X'X)^{-1}x^*}$$

where

$$\hat{y} = b_0 + b_1 x_1^* + b_2 x_2^* + \Lambda\ + b_k x_k^*$$

$x^* = \begin{pmatrix}1 & x_1^* & x_2^* & \Lambda & x_k^*\end{pmatrix}'$ is the given value of x,

s and $(X'X)^{-1}$ are obtained from the least squares analysis,

$t_{\alpha/2}$ is based on the number of degrees of freedom associated with s, namely, [n-(k+1)]

The procedure for forming a prediction interval for y for a given x^* is shown in following box.

A (1-α)100% PREDICTION INTERVAL FOR y

$$\hat{y} \pm t_{\alpha/2} s\sqrt{1 + (x^*)'(X'X)^{-1}x^*}$$

where

$$\hat{y} = b_0 + b_1 x_1^* + b_2 x_2^* + \Lambda\ + b_k x_k^*$$

$x^* = \begin{pmatrix}1 & x_1^* & x_2^* & \Lambda & x_k^*\end{pmatrix}'$ is the given value of x,

s and $(X'X)^{-1}$ are obtained from the least squares analysis,

$t_{\alpha/2}$ is based on the number of degrees of freedom associated with s, namely, [n-(k+1)]

12.8 Multiple linear regression: An overview example

In the previous sections we have presented the basic elements necessary to fit and use a multiple linear regression model . In this section we will assemble these elements by applying them to an example.

Example 12.7 Suppose a property appraiser wants to model the relationship between the sale price of a residential property in a mid-sized city and the following three independent variables:

(1) appraised land value of the property,
(2) appraised value of improvements (i.e., home value)
(3) area of living space on the property (i.e., home size)

Consider the linear model

$y = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + e$

where

y = Sale price (dollars)
x_1 = Appraised land value (dollars)
x_2 = Appraised improvements (dollars)
x_3 = Area (square feet)
In order to fit the model, the appraiser selected a random sample of $n = 20$ properties from the thousands of properties that were sold in a particular year. The resulting data are given in Table 12.4.

Table 12.4 *Real Estate Appraisal Data*

Property # (Obs.)	**Sale price, y**	**Land value, x_1**	**Improvements value , x_2**	**Area, x_3**
1	68900	5960	44967	1873
2	48500	9000	27860	928
3	55500	9500	31439	1126
4	62000	10000	39592	1265
5	116500	18000	72827	2214
6	45000	8500	27317	912
7	38000	8000	29856	899
8	83000	23000	47752	1803
9	59000	8100	39117	1204
10	47500	9000	29349	1725
11	40500	7300	40166	1080
12	40000	8000	31679	1529
13	97000	20000	58510	2455
14	45500	8000	23454	1151
15	40900	8000	20897	1173
16	80000	10500	56248	1960
17	56000	4000	20859	1344
18	37000	4500	22610	988
19	50000	3400	35948	1076
20	22400	1500	5779	962

Step 1 Hypothesize the form of the linear model

$y = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + e$

Step 2 Use the sample data to find least squares prediction equation. Using the formulas given in Section 12.3 we found

$$\hat{y} = 1470.28 + 0.8145x_1 + 0.824x_2 + 13.53x_3 .$$

This is the same result obtained by computer using STATGRAPHICS (see Figure 12.5)

Step 3 Compute an estimator, s^2, for the variance σ^2 of the random error e :

$$s^2 = \frac{SSE}{n-(k+1)}$$

where

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

STATGRAPHICS gives s = 7919.48 (see Stnd. error of est. in Figure 12.6)

Step 4 Check the utility of the model

a) Does the model fits the data well?

For this purpose calculate the coefficient of determination

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$

You can see in the printout in Figure 12.6 that SSE = 1003491259 (in column “Sum of Squares” and row “Error”) and SS_{yy} = 9783168000 (in column “Sum of Squares” and row “Total”), and R^2 is R-squared =0.897427. This large value of R^2 indicates that the model provides a good fit to the n = 20 sample data points.

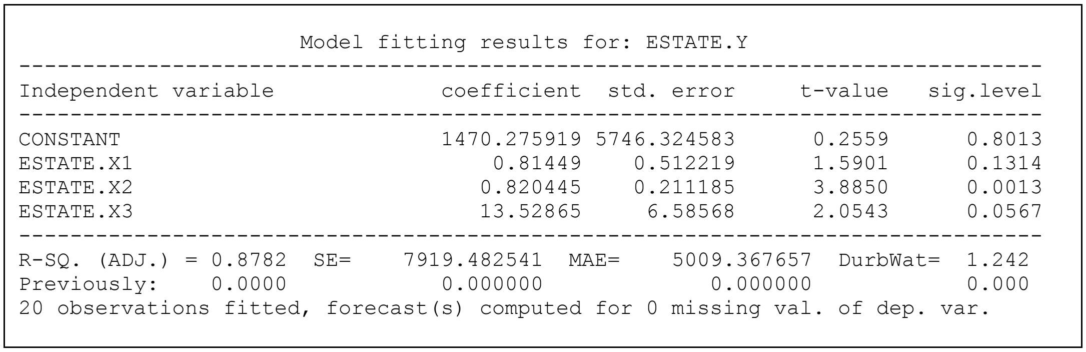
b) Usefulness of the model

Test $H_0 : B_1 = B_2 = ...= B_k = 0$ (Null hypothesis) against H_a : At least one $B_i \neq 0$ (Alternative hypothesis).

Test statistic:

$$F = \frac{R^2 / k}{(1 - R^2)/[n - (k+1)]} = \frac{\text{Mean Square for Model}}{\text{Mean Square for Error}} = \frac{SS(\text{Model})/k}{SSE/[n-(k+1)]}$$

In the printout F= 46.6620, the observed significance level for this test is 0.0000 (under the column *P*-value). This implies that we would reject the null hypothesis for any level, for example 0.01. Thus, we have strong evidence to reject H_0 and conclude that the model is useful for predicting the sale price of residential properties.



```
                      Model fitting results for: ESTATE.Y
--------------------------------------------------------------------------------
Independent variable             coefficient  std. error     t-value   sig.level
--------------------------------------------------------------------------------
CONSTANT                         1470.275919 5746.324583      0.2559      0.8013
ESTATE.X1                            0.81449    0.512219      1.5901      0.1314
ESTATE.X2                           0.820445    0.211185      3.8850      0.0013
ESTATE.X3                           13.52865     6.58568      2.0543      0.0567
--------------------------------------------------------------------------------
R-SQ. (ADJ.) = 0.8782  SE=    7919.482541  MAE=    5009.367657  DurbWat=  1.242
Previously:    0.0000           0.000000             0.000000            0.000
20 observations fitted, forecast(s) computed for 0 missing val. of dep. var.
```

Figure 12.5 *STATGRAPHICS Printout for Estate Appraisal Example*

```
Analysis of Variance for the Full Regression
-------------------------------------------------------------------------------
Source               Sum of Squares     DF    Mean Square      F-Ratio   P-value
--------------------------------------------------------------------------------
Model                   8779676741.      3    2926558914.      46.6620    0.0000
Error                   1003491259.     16      62718204.
--------------------------------------------------------------------------------
(Total (Corr.)           9783168000.     19

R-squared = 0.897427                             Stnd. error of est. = 7919.48
R-squared (Adj. for d.f.) = 0.878194          Durbin-Watson statistic = 1.24161
```

Figure 12.6 *STATGRAPHICS Printout for Estate Appraisal Example*

Step 5 Use the model for estimation and prediction

(1) **Construct a confidence interval for E(y) for particular values of the independent variables**.

Estimate the mean sale price, $E(y)$, for a property with x_1 = 15000, x_2 = 50000 and x_3 = 1800, using 95% confidence interval. Substituting these particular values of the independent variables into the least squares prediction equation yields the predicted value equal 79061.4. In the printout reproduced in Figure 12.7 the 95% confidence interval for the sale price corresponding to the given (x_1, x_2, x_3) is (733379.3, 84743.6).

Regression results for ESTATE.Y

Observation Number	Observed Values	Fitted Values	Lower 95% CL for means	Upper 95% CL for means
1	68900	68556.7		
2	48500	44212.9		
3	55500	50235.2		
4	62000	59212		
5	116500	105834		
6	45000	43143.7		
7	38000	44643.6		
8	83000	83773.6		
9	59000	56449.5		
10	47500	56216.8		
11	40500	54981		
12	40000	54662.4		

13	97000	98977.1		
14	45500	42800.4		
15	40900	41000.1		
16	80000	82686.9		
17	56000	40024.4		
18	37000	37052		
19	50000	48289.7		
20	22400	20447.9		
21		79061.4	73379.3	84743.6

Figure 12.7 *STATGRAPHICS Printout for estimated mean and corresponding confidence interval for x_1 = 15000, x_2 = 50000 and x_3 = 1800*

(2) **Construct a confidence interval for prediction y for particular values of the independent variables**.

For example, construct a 95% prediction interval for y with x_1 = 15000, x_2 = 50000 and x_3 = 1800.

The printout reproduced in Figure 12.8 shows that the prediction interval for y with the given x is (61333.4, 96789.4).

We see that the prediction interval for a particular value of y is wider than the confidence interval for the mean value.

Regression results for ESTATE.Y

Observation Number	Observed Values	Fitted Values	Lower 95% CL for forecasts	Upper 95% CL for forecasts
1	68900	68556.7		
2	48500	44212.9		
3	55500	50235.2		
4	62000	59212		
5	116500	105834		
6	45000	43143.7		
7	38000	44643.6		
8	83000	83773.6		
9	59000	56449.5		
10	47500	56216.8		
11	40500	54981		
12	40000	54662.4		

13	97000	98977.1		
14	45500	42800.4		
15	40900	41000.1		
16	80000	82686.9		
17	56000	40024.4		
18	37000	37052		
19	50000	48289.7		
20	22400	20447.9		
21		79061.4	61333.4	96789.4

Figure 12.8 *STATGRAPHICS Printout for estimated mean and corresponding prediction interval for x_1 = 15000, x_2 = 50000 and x_3 = 1800*

12.8. Model building: interaction models

Suppose the relationship between the dependent variable y and the independent x_1 and x_2 is described by first-order linear model $E(y) = B_0 + B_1x_1 + B_2 x_2$. When the values of one variable, say x_2, are fixed then $E(y)$ is a linear function of the other variable (x_1):

$E(y) = (B_0 + B_2 x_2) + B_1x_1$.

Therefore, the graph of $E(y)$ against x_1 is a set of parallel straight lines.
For example, if

$E(y)=1 + 2x_1 – x_2$,

the graphs of $E(y)$ for $x_2 = 0$, $x_2 = 2$ and $x_2 = -3$ are depicted in Figure 12.9.

When this situation occurs (as it always does for a first-order model), we say that the relationship between $E(y)$ and any one independent variable does not depend on the value of the other independent variable(s) in the model – that is, we say that ***the independent variables do not interact.***

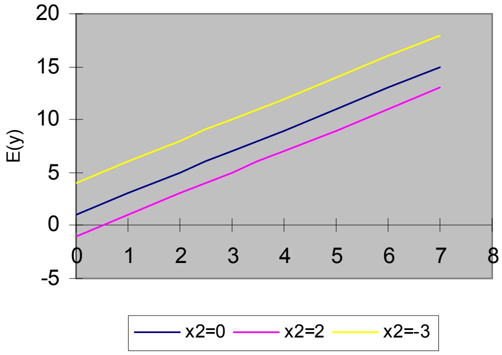


Figure 12.9 *Graphs of E(y) = 1 + 2x_1 – x_2 versus x_1 for fixed values of x_2*

However, if the relationship between $E(y)$ and x_1 does, in fact, depend on the value of x_2 held fixed, then the first-order model is not appropriate for predicting y. In this case we need another model that will take into account this dependence. This model is illustrated in the next example

Example 12.8 Suppose that the mean value $E(y)$ of a response y is related to two quantitative variables x_1 and x_2 by the model
$E(y) = 1 + 2x_1 – x_2 + x_1x_2$.
Graph the relationship between $E(y)$ and x_1 for $x_2 = 0$, 2 and –3. Interpret the graph.

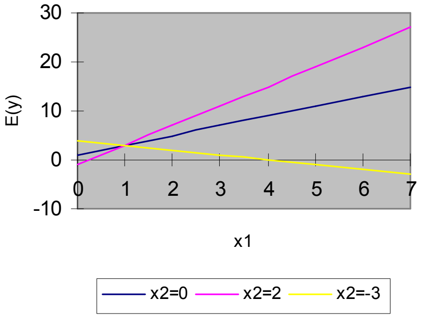


Figure 12.10 *Graphs of E(y) = 1 + 2x_1 – x_2 + x_1x_2 versus x_1 for fixed values of x_2*

Solution For fixed values of x_2, $E(y)$ is linear functions of x_1. Graphs of the straight lines of $E(y)$ for
$x_2 = 0$, 2 and –3 are depicted in Figure 12.10. Note that the slope of each line is represented by $2+ x_2$. The effect of adding a term involving the product x_1x_2 can be seen in the figure. In contrast to Figure 12.9, the lines relating $E(y)$ to x_1 are no longer parallel. The effect on $E(y)$ of a change in x_1 (i.e. the slope) now depends on the value of x_2 .
When this situation occurs, we say that **x_1 and x_2 interact.**. The cross-product term, x_1x_2, *is called an interaction term* and the model

$$E(y) = B_0 + B_1x_1 + B_2x_2 + B_3x_1x_2$$

is called an ***interaction model with two independent variables.***
Below we suggest a practical procedure for building a interaction model.

Procedure to build a interaction model for the relationship between $E(y)$ and two independent variables x_1 and x_2

1. If from observations it is known that the rate of change of $E(y)$ in x_1 depends on x_2 and vice versa, then the interaction model
$$E(y) = B_0 + B_1x_1 + B_2x_2 + B_3x_1x_2$$
is hypothesized.
2. Fit the model to the data.
3. Check if the model fits the data well.
4. Test whether the model is useful for predicting y i.e., test hypothesis H_0 : $B_1 = B_2 = B_3 = 0$ (Null hypothesis) against H_a : At least one $B_i \neq 0$ (Alternative hypothesis).
5. If model is useful for predicting y (i.e. reject H_0), test whether the interaction term contributes significantly to the model:
   H_0 : $B_3 = 0$ (no interaction between x_1 and x_2)
   H_a : $B_3 \neq 0$ (x_1 and x_2 interact)

12.9. Model building: quadratic models

A quadratic (second-order) model in a single quantitative independent variable

$$E(y) = B_0 + B_1x + B_2x^2$$

where B_0 = y-intercept of the curve
B_1 = shift parameter
B_2 = rate of curvature

Procedure to build a quadratic model for the relationship between E(y) and independent variables x

1. State the hypothesized model
   $E(y) = B_0 + B_1x + B_2x^2$
2. Fit the model to the data.
3. Check if the model fits the data well.
4. Test whether the model is useful for predicting y i.e., test hypothesis H_0 : $B_1 = B_2 = = 0$ (Null hypothesis) against H_a : At least one $B_i \neq 0$ (Alternative hypothesis).
5. If model is useful for predicting y (i.e. reject H_0), test whether the second-order term contributes significantly to the model:
   H_0 : $B_2 = 0$
   H_a : $B \neq 0$.

12.11 Summary

In this chapter we have discussed some of the methodology of multiple regression analysis, a technique for modeling a dependent variable y as a function of several independent variables $x_1, x_2, ..., x_k$. The steps employed in a multiple regression analysis are much the same as those employed in a simple regression analysis:

1. The form of the probabilistic model is hypothesized.
2. The appropriate model assumptions are made.
3. The model coefficients are estimated using the method of least squares.
4. The utility of the model is checked using the overall F-test and t-tests on individual B-parameters.
5. If the model is deemed useful and the assumptions are satisfied, it may be used to make estimates and to predict values of y to be observed in the future.

12.12 Exercises

1. Suppose you fit the first-order multiple regression model
$y = B_0 + B_1x_1 + B_2x_2 + e$
to n = 20 data points and obtain the prediction equation
$\hat{y} = 6.4 + 3.1x_1 + 0.92x_2$
The estimated standard deviations of the sampling distributions of b_1, b_2 (least squares estimators of B_0, B_1) are 2.3 and 0.27, respectively.
a) Test H_0: $B_1 = 0$ against H_a: $B_1 > 0$. Use $\alpha = 0.05$.
b) Test H_0: $B_2 = 0$ against H_a: $B_2 > 0$. Use $\alpha = 0.05$.
c) Find a 95% confidence interval for B_1. Interpret the interval.
d) Find a 99% confidence interval for B_2. Interpret the interval.

Suppose you fit the first-order multiple regression model
$y = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + e$
to n = 20 data points and obtain $R^2 = 0.2632$. Test the null hypothesis H_0: $B_1 = B_2 = B_3 = 0$ against the alternative hypothesis that at least one of the B parameters in nonzero. Use α = 0.05.

Plastics made under different environmental conditions are known to have differing strengths. A scientist would like to know which combination of temperature and pressure yields a plastic with a high breaking strength. A small preliminary experiment was run at two pressure levels and two temperature levels. The following model is proposed:
$E(y) = B_0 + B_1x_1 + B_2x_2$
where
y = Breaking strength (pounds)
x_1 = Temperature (^{0}F)
x_2 = Pressure (pounds per square inch).
A sample of n = 16 observations yield
$\hat{y} = 226.8 + 4.9x_1 + 1.2x_2$
with $s_{b1} = 1.11$, $s_{b2} = 0.27$.
Do the data indicate that the pressure is important predictor of breaking strength?
Test using $\alpha = 0.05$.

Suppose you fit the interaction model
$\boldsymbol{E(y) = B_0 + B_1x_1 + B_2x_2 + B_3x_1x_2}$
in n = 32 data points and obtain the following results:
$SS_{yy} = 479$ $SSE = 21$ $b_3 = 10,\ s_{b3} = 4.$
a) Find R^2 and interpret its value.
b) Is the model adequate for predicting y? Test at $\alpha = 0.05$.
c) Use a graph to explain the contribution for the x_1x_2 term to the model.
d) Is there evidence that x_1 and x_2 interact? Test at $\alpha = 0.05$.

The researchers in the international rice research institute in the Philippines conducted a study on the Yield Response of Rice Variety IR661-1-170 to Nitrogen Fertilizer. They obtained the following data

Pair Number	Grain Yield, kg/ha, *y*	Nitrogen Rate, kg/ha, *x*
1	4878	0
2	5506	30
3	6083	60
4	6291	90
5	6361	120

and suggested the quadratic model

$\boldsymbol{E(y) = B_0 + B_1x + B_2x^2}$

The portions of STATGRAPHICS printouts are shown below.

a) Identify the least squares model fitted to the data.

b) What are the values of SSE and s^2 for the data?

c) Perform a test of overall model adequacy. Use $\alpha = 0.05$.

d) Test whether the second-order term contributes significantly to the model. Use $\alpha = 0.05$.

```
                         Model fitting results for: NITROGEN. y
Independent variable          coefficient          std. error          t-value        sig.level
CONSTANT                             4861.457143          47.349987       102.6707       0.0001
x                                       26.64619           1.869659        14.2519       0.0049
x *x                                   -0.117857           0.014941        -7.8884       0.0157
R-SQ. (ADJ.) = 0.9935      SE=     50.312168          MAE=      25.440000         DurbWat=  3.426
5 observations fitted, forecast(s) computed for 0 missing val. of dep. var.
```

```
                         Analysis of Variance for the Full Regression
Source             Sum of Squares        DF          Mean Square      F-Ratio       P-value
Model                      1564516              2           782258      309.032        0.0032
Error                       5062.63              2          2531.31
Total (Corr.)              1569579              4
R-squared = 0.996775                              Stnd. error of est. = 50.3122
```

Chapter 13 Nonparametric statistics

CONTENTS

13.1. Introduction

The majority of hypothesis tests (t- and F-tests) discussed so far have made inferences about population parameters, such as the mean and the proportion. These parametric tests have used the parametric statistics of samples that came from the population being tested. To formulate these tests, we made restrictive assumptions about the populations from which we drew our samples. In each case of Chapter 9, for example, we assumed that our samples either were large or came from normally distributed populations. But populations are not always normal. And even if a goodness-of-fit test indicates that a population is approximately normal, we can not always be certain we're right, because the testis not 100 percent reliable. Clearly, there are certain situations in which the use of the normal curve is not appropriate.

An another case in which the t- and F-tests are inappropriate is when the data are not measurements but can be ranked in order of magnitude. For example, suppose we want to compare the ease of operation of two types of computer software based on subjective evaluations by trained observers. Although we can not give an exact value to the variable Ease of operation of the software package, we may be able to decide that package A is better than package B. If packages A and B are evaluated by each of ten observers, we have the standard problem of comparing the probability distributions for two populations of ratings – one for package A and one for package B. But the t-test of Chapter 9 would be inappropriate, because the only data that can be recorded are preferences; that is, each observer decides either that A is better than B or vice versa.

For the two types of the situations statisticians have developed useful techniques called **nonparametric methods** or nonparametric statistics. The nonparametric counterparts of the t- and F-tests compare the relative locations of the probability distributions of the sampled populations, rather than specific parameters of these populations (such as the means or variances). Many nonparametric methods use the **relative ranks** of the sample observations rather than their actual numerical values.
A large number of nonparametric tests exist, but this chapter will examine only a few of the better known and more widely used ones.

13.2. The sign test for a single population

Recall from Chapter 9 that small-sample procedures for testing a hypothesis about a population mean, require that the population have an approximately normal distribution. For situations in which we collect a small sample (n < 30) from a non-normal distribution, the *t*-testis not valid and we must resort to a nonparametric procedure. The simplest nonparametric technique to apply in this situation is the **sign test**. The sign test is specifically designed for testing hypotheses about the median of any continuous population. Like the mean, the median is a measure of the center, or location, of the distribution; therefore, the sign test is sometimes referred to as a **test for location**.

The theoretical background of the sign test follows.
Let x_1, x_2, ..., x_n be a random sample form a population with unknown median M. Suppose we want to test the null hypothesis H_0: M = M_0 against the one-side alternative H_a: M > M_0. From **Definition 3.2** we know that the median is a number such that half the area under the probability distribution lies to the left of M and half lies to the right. Therefore, the probability that a x-value selected from the population is larger than M is 0.5, i.e., $P(x_i > M) = 0.5$. If, in fact, the null hypothesis is true, then we should expect to observe approximately half the sample x-value greater than M= M_0.
The sign test utilizes the test statistic S, where

$$S = \{ \text{ number of values } x_i \text{ that exceed } M_0\}.$$

Notice that S depends only on the sign (positive or negative) of the difference x_i - M_0. That is, we simply count the number of positive (+) signs among the differences x_i - M_0. If S is "too large" the we will reject H_0 in favor of H_a: M > M_0.
The rejection region for the sign test is derived as follows. Let each sample difference x_i - M_0 denote the outcome of a single trial in an experiment consisting of n identical trials. If we call a positive difference a "Success" and a negative difference a "Failure", then S is the number of successes in n trials. Under H_0 the probability of observing a success on any one trial is

$p = P(\text{Success}) = P(x_i - M_0 > 0) = P(x_i > M_0) = 0.5$

Since the trials are independent, the properties of a binomial distribution, listed in Section 5.3, are satisfied. Therefore, *S* has a binomial distribution with parameters n and p = 0.5. We can use this fact to calculate the observed significance level (*p*-value) of the sign test.

The procedure for the sign test is presented in the following box.

SIGN TEST FOR A POPULATION MEDIAN

ONE-TAILED TEST	**TWO-TAILED TEST**
$H_0 : M = M_0$	$H_0 : M = M_0$
$H_a : M > M_0$ (or $M < M_0$)	$H_a : M \neq M_0$
Test statistic: S = Number of sample observations greater than M_0 (or S = Number of sample observations less than M_0)	**Test statistic:** $S = max\ (S_1, S_2)$, where S_1 = Number of sample observations greater than M_0, S_2 = Number of sample observations less than M_0

Observed significance level: p-value = $P(S \geq S_c)$	[Note: By definition $S_2 = n - S_1$] **Observed significance level:** p-value =2 $P(S \geq S_c)$

where S_c is the computed value of the test statistic and S has a binomial distribution with parameters n and p = 0.5.

Rejection region: Reject H_0 if α > p-value.

Example 13.1 Suppose from a population the following sample is randomly selected:
41 33 43 52 46 37 44 49 53 30.
Do the data provide sufficient evidence to indicate that the median percentage of the population is greater than 40? Test using α = 0.05.

Solution We want to test

H_0: M = 40
H_a: M > 40
using the sign test. The test statistic is
S = {Number of sample observations greater than 40}
ha s binomial distribution with n =10 and p = 0.5.
The computed test statistic Sc = 7 and p-value = $P(S \geq 7)$ = 1 – $P(S \leq 6)$ = 1 – 0.828 = 0.172.
Since p-value > α = 0.05, we can not reject the null hypothesis. That is, there is insufficient evidence to indicate the median percentage of the population exceeds 40.

Recall from Section 5.8 that a normal distribution with the mean $\mu = np$ and the variance $\sigma^2 = np(1-p)$ can be used to approximate the binomial distribution for large n. When p = 0.5, the normal approximation performs reasonably well even for n as small as 10 (see Figure 5.6 or Table 5.4).
Thus, for $n \geq 10$ we can conduct the sign test using the familiar standard normal z-statistic.

SIGN TEST BASED ON A LARGE SAMPLE $(n \geq 10)$

ONE-TAILED TEST	**TWO-TAILED TEST**
$H_0 : M = M_0$	$H_0 : M = M_0$
$H_a : M > M_0$ (or $M < M_0$)	$H_a : M \neq M_0$

Test statistic:

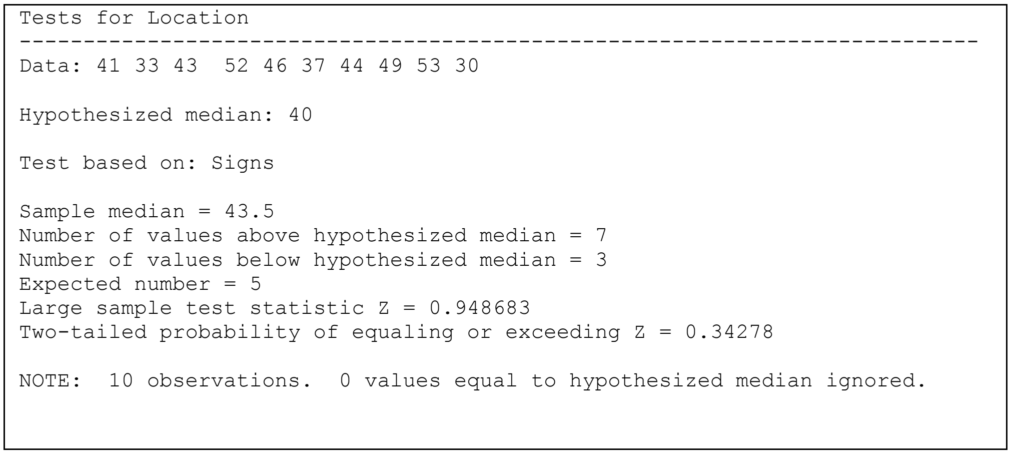
$$z = \frac{S - E(S)}{\sqrt{\sigma(S)}} = \frac{S - 0.5n}{\sqrt{(0.5)(0.5)n}} = \frac{S - 0.5n}{0.5\sqrt{n}}$$

S = Number of sample observations greater than M_0 (or S = Number of sample observations less than M_0)	$S = max\ (S1,\ S2)$, where S_1 = Number of sample observations greater than M_0, S_2 = Number of sample observations less than M_0 [Note: By definition $S_2 = n - S_1$]
Rejection region:	**Rejection region:**
$z > z_\alpha$ (or $z < -z_\alpha$)	$z < -z_{\alpha/2}$ (or $z > z_{\alpha/2}$)

where z_α and $z_{\alpha/2}$ are tabulated values given in any table of normal curve areas.

Example 13.2 Refer to Example 13.1 using the sign test based on z-statistic.

Solution For this example the software STATGRAPHICS provides the following printout.



```
Tests for Location
---------------------------------------------------------------------------
Data: 41 33 43  52 46 37 44 49 53 30

Hypothesized median: 40

Test based on: Signs

Sample median = 43.5
Number of values above hypothesized median = 7
Number of values below hypothesized median = 3
Expected number = 5
Large sample test statistic Z = 0.948683
Two-tailed probability of equaling or exceeding Z = 0.34278

NOTE:  10 observations.  0 values equal to hypothesized median ignored.
```

Figure 13.1 *STATGRAPHICS printout for Example 13.2.*

From the printout we see that the computed statistic z_c = 0.948683 and $P(|z| \geq z_c) = 0.34278$.

Therefore $P(z \geq z_c) = 0.17139$, that is, p-value = 0.17139.

Since p-value > α = 0.05, we can not reject the null hypothesis. That is, there is insufficient evidence to indicate the median percentage of the population exceeds 40.

13.3 Comparing two populations based on independent random samples: Wilcoxon rank sum test

In Chapter 9 we presented parametric tests (tests about population parameters) based on the z- and the t-statistics, to test for a difference between two population means. Recall that the mean of a population measures the location of the population distribution. Another measure of the location of the population distribution is the median M. Thus, if the data provides sufficient evidence to indicate that $M_1 > M_2$, we imagine the distribution for the population 1 shifted to right of population 2.

The equivalent nonparametric test is not a test about the difference between population means. Rather, it is a test to detect whether distribution 1 is shifted to the right of distribution 2 or vice versa. The test based on independent random samples of n_1 and n_2 observations from the respective populations, is known as the **Wilcoxon rank sum test**.

To use the Wilcoxon rank sum test, we first rank all ($n_1 + n_2$) observations, assigning a rank of 1 to the smallest, 2 to the second smallest, and so on. Tied observations (if they occur) are assigned ranks equal to the average of the ranks of the tied observations. For example, if the second and the third ranked observations were tied, each would be assigned the rank 2.5. The sum of the ranks, called a rank sum, is then calculated for each sample. If the two distributions are identical, we would expect the same rank sums, designated as T_1 and T_2, to be nearly equal. In contrast, if one rank sum – say, T_1 – is much larger than the other, T_2, then the data suggest that the distribution for population 1 is shifted to the right of the distribution for population 2. The procedure for conducting a Wilcoxon rank sum test is summarized in the following box.

WILCOXON RANK SUM TEST FOR A SHIFT IN POPULATION LOCATIONS:

INDEPENDENT RANDOM SAMPLES

ONE-TAILED TEST

H_0: The sampled populations have identical probability distributions

H_a: The probability distribution for population 1 is shifted to the right of that for population 2

TWO-TAILED TEST

H_0: The sampled populations have identical probability distributions

H_a: The probability distribution for population 1 is shifted either to the left or to the right of that for population 2

Rank the $n_1 + n_2$ observations in the two samples from the smallest (rank 1) to the largest (rank $n_1 + n_2$). Calculate T_1 and T_2, the rank sums associated with sample 1 and sample 2, respectively. Then calculate the test statistic.

Test statistic:

T_1 if $n_1 < n_2$ or T_2 if $n_2 \le n_1$

Test statistic:

T_1 if $n_1 \le n_2$; T_2 if $n_2 \le n_1$. We will denote this rank sum as T.

Rejection region:	**Rejection region:**
$T_1 \geq T_U$ if T_1 is test statistic; or $T_2 \leq T_L$ if T_2 is test statistic	$T \geq T_U$, or $T \leq T_L$
where T_U and T_L are obtained from Table 1 of Appendix D	

Example 13.3 Independent random samples were selected from two populations. The data are shown in Table 13.1. Is there sufficient evidence to indicate that population 1 is shifted to the right of population 2. Test using α = 0.05.

Table 13.1 *Data for Example 13.3*

Sample from Population 1	Sample from Population 2
17	10
14	15
12	7
16	6
23	13
18	11
10	12
8	9
19	17
22	14

Solution The ranks of the 20 observations from lowest to highest, are shown in Table 13.2. We test

H_0: The sampled populations have identical probability distributions

H_a: The probability distribution for population 1 is shifted to the right of that for population 2

The test statistic T_2 =78. Examining Table 13.3 we find that the critical values, corresponding to $n_1 = n_2$ =10 are T_L = 79 and T_U = 131. Therefore, for one-tailed test at α = 0.025, we will reject H_0 if $T_2 \leq T_L$, i.e., reject H_0 if $T_2 \leq 79$. Since the observed value of the test statistic, T_2 =78 <79 we reject H_0 and conclude (at α = 0.025) that the probability distribution for population 1 is shifted to the right of that for population 2.

Table 13.2 *Calculations of rank sums for Example 13.3*

Sample from Population 1		**Sample from Population 2**	
Raw data	Rank	Raw data	Rank

17	15.5	10	5.5
14	11.5	15	13
12	8.5	7	2
16	14	6	1
23	20	13	10
18	17	11	7
10	5.5	12	8.5
8	3	9	4
19	18	17	15.5
22	19	14	11.5
	$T_1 = 132$		$T_2 = 78$

Table 13.3 *A Partial Reproduction of Table 1 of Appendix D*

Critical values of T_L and T_U for the Wilcoxon Rank Sum Test: Independent samples

a. Alpha = 0.025 one-tailed; alpha = 0.05 two-tailed

n_1	3		4		5		6		7		8		9		10	
n_2																
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

Many nonparametric test statistics have sampling distributions that are approximately normal when n_1 and n_2 are large. For these situations we can test hypotheses using the large-sample z-test.

WILCOXON RANK SUM TEST FOR LARGE SAMPLES $(n_1 \geq 10 \text{ and} n_2 \geq 10)$

ONE-TAILED TEST

H_0: The sampled populations have identical probability distributions
H_a: The probability distribution for population 1 is shifted to the right of that for population 2
(or population 1 is shifted to the left of population 2)

TWO-TAILED TEST

H_0: The sampled populations have identical probability distributions
H_a: The probability distribution for population 1 is shifted either to the left or to the right of that for population 2

Test statistic:

$$z = \frac{T_1 - \left[\dfrac{n_1 n_2 + n_1(n_1+1)}{2}\right]}{\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Rejection region:

$z > z_\alpha$ (or $z < -z_\alpha$)

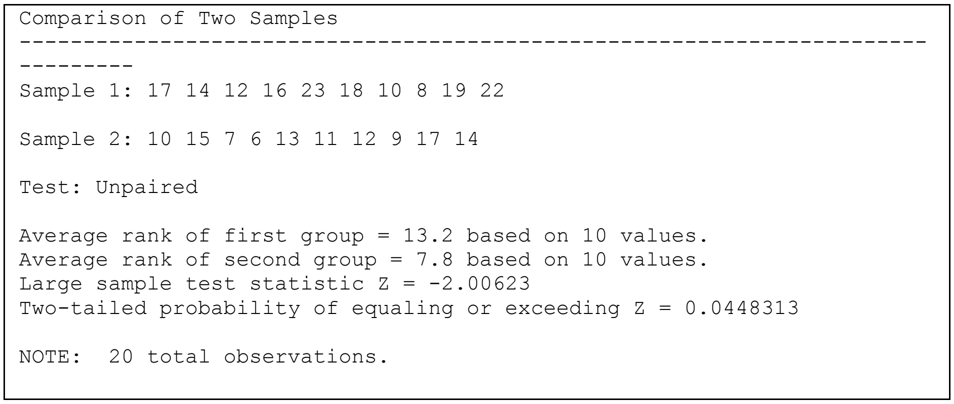
Rejection region:

$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

where z_α and $z_{\alpha/2}$ are tabulated values given in Table 1 of Appendix C

Example 13.4 Refer to Example 13.3. Using the above large-sample z-test check whether there is sufficient evidence to indicate that population 1 is shifted to the right of population 2. Test using α = 0.05.

Solution We do this example with the help of computer using STATGRAPHICS. The printout is given in Figure 13.2.



```
Comparison of Two Samples
-----------------------------------------------------------------------
---------
Sample 1: 17 14 12 16 23 18 10 8 19 22

Sample 2: 10 15 7 6 13 11 12 9 17 14

Test: Unpaired

Average rank of first group = 13.2 based on 10 values.
Average rank of second group = 7.8 based on 10 values.
Large sample test statistic Z = -2.00623
Two-tailed probability of equaling or exceeding Z = 0.0448313

NOTE:  20 total observations.
```

Figure 13.2 *STATGRAPHICS printout for Example 13.4*

From the printout we see that the computed test statistic z_c = -2.00623 and the two-tailed probability $P(|z| \geq z_c) = 0.0448313$. Therefore, $P(z \leq z_c) = 0.022415$. Hence, at significance level α < 0.023 we reject the null hypothesis and conclude that the probability distribution for population 1 is shifted to the right of that for population 2 at this significance level.

13.4. Comparing two populations based on matched pairs: the Wilcoxon signed ranks test

Recall from Chapter 9 that the analysis of matched-pairs data is based on the differences within the matched pairs of observations. The Wilcoxon signed ranks test is a nonparametric test to detect shifts in locations for population probability distributions. The test is summarized in the box.

WILCOXON SIGNED RANKS TEST: MATCHED PAIRS

ONE-TAILED TEST	**TWO-TAILED TEST**
H_0: The sampled populations have identical probability distributions H_a: The probability distribution for population 1 is shifted to the right of that for population 2	H_0: The sampled populations have identical probability distributions H_a: The probability distribution for population 1 is shifted either to the left or to the right of that for population 2

Calculate the differences within each of the n matched pairs of observations. Then rank the absolute values of the n differences from smallest (rank 1) to the highest (rank n) and calculate the rank sum T^- of the negative differences and the rank sum T^+ of the positive differences.

Test statistic:	**Test statistic:**
T^-, the rank sum of the negative differences	$T = \min(T^-, T^+)$
Rejection region:	**Rejection region:**
$T^- \leq T_0$	$T \leq T_0$

where T_0 is given in Table 2 of Appendix D

[Note: differences equal to 0 are eliminated and the number n of differences is reduced accordingly. Tied absolute differences receive ranks equal to the average of the ranks they would have received had they not been tied.]

Example 13. 5 Suppose that a company wants to know the opinion of customers about the quality of its product before and after introducing a new technology. The company selects randomly 10 customers and each of them is given a sample of the product before (B) and after

(A) introducing the new technology. Each customer rates the quality of each product on a scale from 1 to 10. The results of the experiment are shown in Table 13. . Is there sufficient evidence to indicate that the product after introducing the new technology is rated higher than the one before new technology.
Test using α = 0.05.

Table 13.4 *Product quality ratings*

Customer	Product A	B	Difference $(A - B)$	Absolute value $\lvert A - B \rvert$	Rank of $\lvert A - B \rvert$
1	6	4	2	2	5
2	8	5	3	3	7.5
3	4	5	-1	1	2
4	9	8	1	1	2
5	4	1	3	3	7.5
6	7	9	-2	2	5
7	6	2	4	4	9
8	5	3	2	2	5
9	6	7	-1	1	2
10	8	2	6	6	10
			T^+ = Sum of positive ranks = 46		
			T^- = Sum of negative ranks = 9		

Solution We must test the hypotheses:

H_0: The sampled populations have identical probability distributions
H_a: The probability distribution for population 1 (A) is shifted to the right of that for population 2 (B).
We will use T^- as the test statistic and reject H_0 if $T^- \leq T_0$.
For our example, the computed value T^- = 9. Examining Table 13.5 in the column corresponding to a one-tailed test, the row corresponding to α = 0.05, and the column for n = 10, we read T_0 = 11. Since T^- = 9 < 11 we reject H_0 and conclude that there is sufficient evidence to indicate that the probability distribution of population A is shifted to the right of the probability distribution of population B, that is, after introducing the new technology the product is rated higher than before.

Table 13.5 *A Partial Reproduction of Table 2 of Appendix D*

Critical values of T_0 in the Wilcoxon Matched Pairs Signed Ranks Test

α ONE-TAILED	α TWO-TAILED	n=5	n=6	n=7	n=8	n=9	n=10
0.05	0.1	1	2	4	6	8	11
0.025	0.05		1	2	4	6	8
0.01	0.02			0	2	3	5
0.005	0.01				0	2	3
		n=11	n=12	n=13	n=14	n=15	n=16
	0.1	14	17	21	26	30	36
0.025	0.05	11	14	17	21	25	30
0.01	0.02	7	10	13	16	20	24
0.005	0.01	5	7	10	13	16	19
		n=17	n=18	n=19	n=20	n=21	n=22
	0.1	41	47	54	60	68	75
0.025	0.05	35	40	46	52	59	66
0.01	0.02	28	33	38	43	49	56
0.005	0.01	23	28	32	37	43	49
		n=23	n=24	n=25	n=26	n=27	n=28
	0.1	83	92	101	110	120	130
0.025	0.05	73	81	90	98	107	117
0.01	0.02	62	69	77	85	93	102
0.005	0.01	55	61	68	76	84	92

The Wilcoxon signed ranks test for large samples

The Wilcoxon signed ranks test statistic has a sampling distribution that is approximately normal when the number n of pairs is large – say, $n \geq 25$. This large sample nonparametric matched-pairs test is summarized in the following box.

WILCOXON SIGNED RANK SUM TEST FOR LARGE SAMPLES (n ≥ 25)

ONE-TAILED TEST

H_0: The sampled populations have identical probability distributions
H_a: The probability distribution for population 1 is shifted to the right of that for population 2
(or population 1 is shifted to the left of population 2)

TWO-TAILED TEST

H_0: The sampled populations have identical probability distributions
H_a: The probability distribution for population 1 is shifted either to the left or to the right of that for population 2

Test statistic:

$$z = \frac{T^{+} - [n(n+1)/4]}{\sqrt{[n(n+10(2n+1)]/24}}$$

Rejection region:

$z > z_\alpha$ (or $z < -z_\alpha$)

Rejection region:

$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

where z_α and $z_{\alpha/2}$ are tabulated values given in any table of normal curve areas.

(See Table 1 of Appendix C)

Example 13.6 Suppose from each of two populations we select a sample. They are 30 matched pairs

Sample 1	4 5 6 4 7 8 6 9 7 4 10 7 6 8 5 4 6 7 9 7 4 6 7 9 6 10 9 7 8 5
Sample 2	5 6 7 8 5 9 6 8 3 7 5 7 5 8 9 4 6 8 4 6 7 9 10 6 8 5 7 8 9 6

Use the Wilcoxon signed ranks test to check whether the probability distributions of the populations are identical.

Solution For this example using STATGRAPHICS we obtain the following printout.

```
Comparison of Two Samples
--------------------------------------------------------------------------
Sample 1: 4 5 6 4 7 8 6 9 7 4 10 7 6 8 5 4 6 7 9 7 4 6 7 9 6 10 9 7 8 5

Sample 2: 5 6 7 8 5 9 6 8 3 7 5 7 5 8 9 4 6 8 4 6 7 9 10 6 8 5 7 8 9 6

Test: Ranks

Number of positive differences = 10 with average rank = 15.4
Number of negative differences = 15 with average rank = 11.4
Large sample test statistic Z = 0.242162
Two-tailed probability of equaling or exceeding Z = 0.80865

NOTE:  30 total pairs.  5 tied pairs ignored.
```

Figure 13.3 *STATGRAPHICS printout for Example 13.6*

From the printout we see that p-value for two-tailed test is 0.80865. This is not small. Therefore, we can not reject the hypothesis that the probability distributions of the populations are identical.

13.5. Comparing population using a completely randomized design: The Kruskal-Wallis H test

In Chapter 10 we compare the means of k populations based on data collected according to a completely randomized design. The analysis of variance F-test, used to test the null hypothesis of equality of means, is based on the assumption that the populations are normally distributed with common variance σ^2.

The Kruskal-Wallis H-test is the nonparametric equivalent of the analysis of variance F-test. It tests the null hypothesis that all k populations possess the same probability distribution against the alternative hypothesis that the distributions differ in location – that is, one or more of the distributions are shifted to the right or left of each other. the advantage of the Kruskal-Wallis H-test is that we need make no assumptions about the nature of the sampled populations.

A completely randomized design specifies that we select independent random samples of n_1, n_2, ..., n_k observations form the k populations. To conduct the test, we first rank all $n = n_1 + n_2 + ... + n_k$ observations and compute the rank sums, $R_1, R_2, ..., R_k$ for the k samples. The ranks of tied observations are averaged in the same manner as for the Wilcoxon rank sum test. Then, if H_0 is true, and if the sample sizes, n_1, n_2, ..., n_k, each equal 5 or more, then the test statistic

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{{R_i}^2}{n} - 3(n+1)$$

will have a sampling distribution that can be approximated by a chi-square distribution with (k – 1) degrees of freedom. Large values of H imply rejection of H_0. Therefore, the rejection region for the test is $H > \chi^2_\alpha$ where χ^2_α is the value that locates α in the upper tail of the chi-square distribution.

The test is summarized in the following box.

KRUSKAL-WALLIS H- TEST FOR COMPARING k POPULATION PROBABILITY

DISTRIBUTIONS

H_0: The k population probability distributions are identical
H_a: At least two of the k population probability distributions differ in location
Test statistic:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n} - 3(n+1)$$

where
n_i = Number of observations in sample i
R_i = Rank sum of sample i, where the rank of each observation is computed according to its relative magnitude in the totality of data for the k samples
$n = n_1 + n_2 + ... + n_k$
Rejection region:
$H > \chi^2_\alpha$ with df = k –1
Assumptions:

1. The k samples are random and independent
2. $n_i \geq 5$ for each i
3. The observations can be ranked.

No assumptions have to be made about the shape of the population probability distribution.

Example 13.7 Independent random samples of three different brands of magnetron tubes were subjected to stress testing, and the number of hours each operated without repair was recorded. Although these times do not represent typical lifetimes, they do indicate how well the tubes can withstand extreme stress.. The data are shown in the table. Experience has shown that the distributions of lifetimes for manufactured products are usually non-normal.

A	B	C
36	49	71
48	33	31
5	60	140
67	2	59
53	55	42

Use the Kruskal-Wallis H-test to determine whether evidence exists to conclude that the brands of magnetron tubes tend to differ in length of life under stress. Test using α = 0.05.
Solution The first step in performing the Kruskal-Wallis H-test is to rank the n = 15 observations in the complete data set. The ranks and rank sums for three samples are shown in Table 13.6

Table 13.6 *Ranks and Rank Sums for Example 13.7*

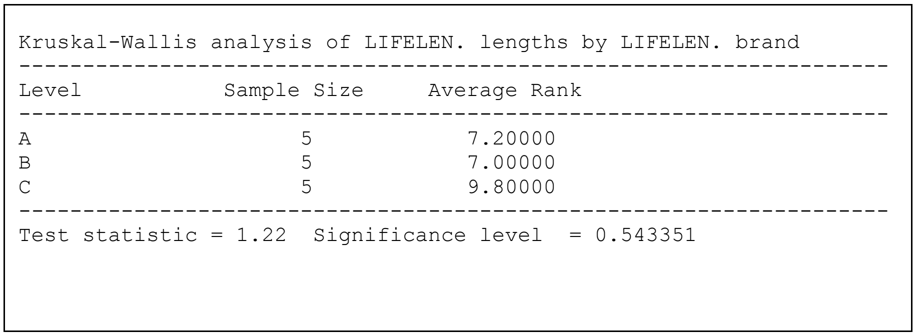
A	RANK	B	RANK	C	RANK
36	5	49	8	71	14
48	7	33	4	31	3
5	2	60	12	140	15
67	13	2	1	59	11
53	9	55	10	42	6
	R_1 =36		R_2 =35		R_3 =49

We want to test the null hypothesis
H_0: The population probability distributions lifetimes under stress are identical for three brands of magnetron tubes
against the alternative hypothesis
H_a : At least two of the population probability distributions differ in location using the test statistic

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n} - 3(n+1) = \frac{12}{(15)(16)}\left[\frac{(36)^2}{5}+\frac{(35)^2}{5}+\frac{(49)^2}{5}\right] - 3(16) = 1.22$$

The rejection region for the H-test is $H > \chi^2_\alpha$ with df = k –1 = 3 –1 = 2. For α = 0.05 and df = 2, $\chi^2_\alpha = 5.99147$. Since the computed value of H =1.22 is less than 5.99147 we can not reject H_0. There is insufficient evidence to indicate a difference in location among the distributions of lifetimes for the three brands of magnetron tubes.

For this example the STATGRAPHICS printout is given in Figure 13.3. In the printout we see that Test statistic = 1.22, Significance level = 0.543351. Therefore, at significance level α = 0.05 we can not reject the hypothesis H_0.



```
Kruskal-Wallis analysis of LIFELEN. lengths by LIFELEN. brand
--------------------------------------------------------------------
Level           Sample Size     Average Rank
--------------------------------------------------------------------
A                      5           7.20000
B                      5           7.00000
C                      5           9.80000
--------------------------------------------------------------------
Test statistic = 1.22  Significance level  = 0.543351
```

Figure 13.4 *STATGRAPHICS printout for Example 13.7*

13.6. Rank Correlation: Spearman's r_s statistic

Several different nonparametric statistics have been developed to measure and to test for correlation between two random variables. One of these statistics is the Spearman's rank correlation coefficient r_s.

The first step in finding r_s is to rank the values of each of the variables separately; ties are treated by averaging the tied ranks. Then r_s is computed in exactly the same way as the simple correlation coefficient r. The only difference is that the values of x and y that appear in the formula for r_s denote the ranks of the raw data rather than the raw data themselves.

Formulas for computing Spearman's rank correlation coefficient

Rank the values for each of the variables and let x and y denote the ranks of a pair of observations. Then

$$r_s = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum (x-\bar{x})^2, \quad SS_{yy} = \sum (y-\bar{y})^2, \quad SS_{xy} = \sum (x-\bar{x})(y-\bar{y})$$

When there are no ties, the formula for r_s, reduces to

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where d is the difference between the values of x and y corresponding to a pair of observations. This simple formula will provide a good approximation to r_s when the number of ties in the ranks is small.

The nonparametric test of hypothesis for rank correlation is shown in the box.

Spearman's Nonparametric Test for Rank Correlation

ONE-TAILED TEST	**TWO-TAILED TEST**
H_0: There is no correlation between the ranked pairs H_a: Ranked pairs are positively correlated (or Ranked pairs are negatively correlated)	H_0: There is no correlation between the ranked pairs H_a: Ranked pairs are correlated
Test statistic: r_s	**Test statistic: r_s**
Rejection region:	**Rejection region:**
$r_s \geq r_0$ (or $r_s \leq -r_0$)	$r_s \geq r_0$ or $r_s \leq -r_0$

where the value of r_0 is given in Table 3 of Appendix D

Example 13.8 A large manufacturing firm wants to determine whether a relationship exists between the number of works-hours an employee misses per year and the employee's annual wages (in thousands of dollars). A sample of 15 employees produced the data shown in Table 13.7.

Table 13.7 *Data for Example 13.8*

EMPLOYEE	HOURS	WAGES
1	49	15.8
2	36	17.5
3	127	11.3
4	91	13.2
5	72	13.0
6	34	14.5
7	155	11.8
8	11	20.2
9	191	10.8
10	6	18.8
11	63	13.8
12	79	12.7
13	43	15.1
14	57	24.2
15	82	13.9

a) Calculate Spearman's rank correlation coefficient as a measure of the strength of the relationship between work-hours missed and annual wages.

b) Is there sufficient evidence to indicate that work-hours missed decrease as annual wages increases , i.e., that work-hours missed and annual wages are negatively correlated? Test using $\alpha = 0.01$.

Solution

a) First we rank the values of work-hours missed and rank the values of the annual salaries. Let these rankings are x_i and y_i, respectively, and they are shown in Table 13.8. The next step is

Table 13.8 *Calculations for Example 13.8*

EMPLOYEE	HOURS	RANK	WAGES	RANK	d_i	d_i^2
1	49	6	15.8	11	-5	25
2	36	4	17.5	12	-8	64

3	127	13	11.3	2	11	121
4	91	12	13.2	6	6	36
5	72	9	13.0	5	4	16
6	34	3	14.5	9	-6	36
7	155	14	11.8	3	11	121
8	11	2	20.2	14	-12	144
9	191	15	10.8	1	14	196
10	6	1	18.8	13	-12	144
11	63	8	13.8	7	1	1
12	79	10	12.7	4	6	36
13	43	5	15.1	10	-5	25
14	57	7	24.2	15	-8	64
15	82	11	13.9	8	3	9
						$\sum d_i^2=1038$

b) To calculate the differences $d_i = x_i - y_i$ (i = 1, 2, ..., 15). These differences d_i and their squares are shown in the table. Since there are no ties, we calculate r_s by the formula

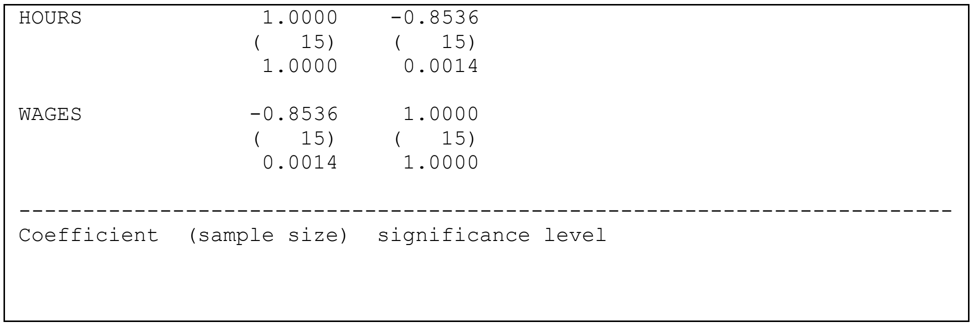
$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(1038)}{15(224)} = -0.854$$

This large negative value of r_s implies that a negative correlation exists between work-hours missed and annual wages in the sample of 15 employees.

c) To test $H_{0:}$ No correlation exists between work-hours missed and annual wages in the population against H_1: Work-hours missed and annual wages are negatively correlated, we use r_s as the test statistic and obtain the critical value r_0 from Table
This table gives the critical values of r_0 for an upper-tailed test, i.e., a test to detect a positive rank correlation. For our example, α = 0.01, n =15, the critical value is r_0 = 0.623. Therefore, we reject the null hypothesis in favor of the alternative hypothesis if the computed r_s statistic is less or equal –0.623. Since our computed r_s = -0.854 < -0.623, we reject H_0 and conclude that there is ample evidence to indicate that work-hours missed decrease as annual wages increases.

Below we reproduce the STATGRAPHICS printout for our example. In this printout we see that the rank correlation coefficient between the variable HOURS (work-hours missed) and the variable WAGES (annual wages) is –0.8536 and the significance level is 0.0014. Since the observed significance level is very small, it is naturally to reject the null hypothesis.

```
Spearman Rank Correlations
---------------------------------------------------------------------
                    HOURS      WAGES
```



```
HOURS              1.0000    -0.8536
                  (   15)    (   15)
                   1.0000     0.0014

WAGES             -0.8536     1.0000
                  (   15)    (   15)
                   0.0014     1.0000

--------------------------------------------------------------------
Coefficient  (sample size)  significance level
```

Figure 13.4 *STATGRAPHICS printout for Example 13.8*

13.7 Summary

We have presented several useful nonparametric techniques for testing the location of a single population, or for comparing two or more populations. Nonparametric techniques are useful when the underlying assumptions for their parametric counterparts are not justified or when it is impossible to assign specific values to the observations. Nonparametric methods provide more general comparisons of populations than parametric methods, because they compare the probability distributions of the populations rather than specific parameters.

Rank sums are the primary tools of nonparametric statistics. The Wincoxon rank sum test can be used to compare two populations based on independent random samples, and Wincoxon signed ranks test can be used for a matched-pairs experiment. The Kruskal-Wallis H-test is applied when comparing k populations using a completely randomized design.

13.8 Exercises

1. Suppose you want to use the sign test to test the null hypothesis that the population median equals 75, i.e., H_0: M = 75. Use the table of binomial probabilities to find the observed significance level (p-value) of the test for each of the following situations:
   a) H_a: M > 75, n = 5, S = 2
   b) H_a: $M \neq 75$, n = 15, S = 9
   c) H_a: M < 75, n = 10, S = 7

2. A random sample of 8 observations from a continuous population resulted in the following:

   17   16.5   20 18.2   19.6   14.9   21.1   19.4

   Is there sufficient evidence to indicate that the population median differs from 20? test using α = 0.05.

3. Independent random variables were selected from two populations. The data are shown in the table

Sample from population 1	15	16	13	14	12	17		
Sample from population 2	6	13	8	9	7	5	4	10

a) Use the Wilcoxon rank sum test to determine whether the data provide sufficient evidence to indicate a shift in the locations of the probability distributions of the sampled populations. Test using $\alpha = 0.05$.

b) Do the data provide sufficient evidence to indicate that the probability distribution for population 1 is shifted to the right of the probability distribution for population 2? Use the Wilcoxon rank sum test with $\alpha = 0.05$.

4. The following data show employee' rates of defective work before and after a change in wage incentive plan. Compare the two sets of data to see if the change lowered the defective units produced (Use the Wilcoxon signed rank test for a matched pairs design with $\alpha = 0.01$)

Before	8	7	6	9	7	10	8	6	5	8	10	8
After	6	5	8	6	9	8	10	7	5	6	9	5

5. The following table shows sample retail prices for three brands of shoes. Use the Kruskal-Wallis test to determine if there is any difference among the retail prices of the brands throughout the country. Use 0.05 level of significance.

Brand A	$89	90	92	81	76	88	85	95	97	86	100
Brand B	$78	93	81	87	89	71	90	96	82	85	
Brand C	$80	88	86	85	79	80	84	85	90	92	

6. A random sample of seven pairs of observations are recorded on two variables, X and Y. The data are shown in the table. use Spearman's nonparametric test for rank correlation to answer the following :
a) Do the data provide sufficient evidence to conclude that the rank correlation between X and Y is greater than 0? Test using $\alpha = 0.05$.
b) Do the data provide sufficient evidence to conclude that the rank correlation between X and Y is not 0? Test using $\alpha = 0.05$.

X	65	57	55	38	29	43	49
Y	58	61	58	23	34	38	37

7. Below are ratings of aggressiveness (X) and amount of sales in the last year (Y) for eight salespeople. Is there a significant rank correlation between the two measures? Use the 0.05 significance level.

X	30	17	35	28	42	25	19	29
Y	35	31	43	46	50	32	33	42

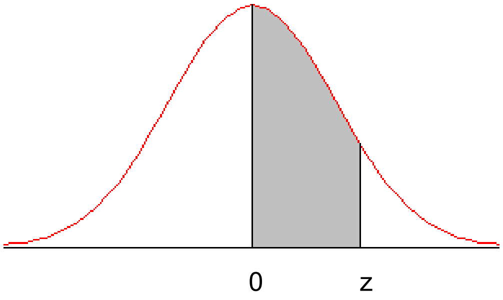
References

1. Berenson, M.L. and D.M. Levine, Basic Business Statistics: Concepts and Applications, 4th ed. Englewood Cliffs, NJ, Prentice Hall, 1989.
2. McClave, J.T. & Dietrich, F.H. Statistics, 4th ed., San Francisco: Dellen,1988.
3. Fahrmeir L. and Tutz G., Multivariate statistical modeling based on generalized linear models, New York: Springer-Verlag, 1994.
4. Gnedenko B.V., The theory of probability, Chelsea Publ. Comp., New York, 1962.
5. Goldstein H. (ed.) Multilevel statistical models, London: Edward Arnold, 1995.
6. Iman, R. L., and W. J. Conover, Modern Business Statistics, 2nd ed. New York, NY, John Wiley & Sons, 1989.
7. Kwanchai A. Gomez and Arturo A. Gomez, Statistical procedures for agricultural research, John Wiley & Sons, 1982.
8. Levin, R.I. and D. S. Rubin, Statistics for management, 5th ed. Englewood Cliffs, NJ, Prentice Hall, 1991.
9. Moore D. S. and G.P. McCabe, Introduction to the Practice of Statistics, W.H. Freeman and Company, 1989.
10. Mendehall, W. and Sincich T., Statistics for the engineering and computer sciences, 2nd edition, Dellen Publ. Comp., 1989.
11. Rosenbaum P.R., Observational Studies, New York: Springer-Verlag, 1995.
12. STATGRAPHICS Plus, Reference manual, Manugistics Inc., 1992.

Table 1

Appendix C

Normal Curve
Areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Table 2

Appendix C Critical Values for Student's *t*

1	3.078	6.314	12.706	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.102	3.852	4.221
14	1.345	1.760	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.528	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table 1 Critical values of T_L and T_U for the Wincoxon Rank Sum Test: Independent samples

a. Alpha = 0.025 one-tailed; alpha = 0.05 two-tailed

n_1 / n_2	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U		T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

a. Alpha = 0.05 one-tailed; alpha = 0.10 two-tailed

n_1 / n_2	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Table 2 Critical values of T0 in the Wincoxon Matched Pairs Signed Ranks Test

Alpha	Alpha						
ONE-TAILED	TWO-TAILED	n=5	n=6	n=7	n=8	n=9	n=10
0.05	0.1	1	2	4	6	8	11
0.025	0.05		1	2	4	6	8
0.01	0.02			0	2	3	5
0.005	0.01				0	2	3
		n=11	n=12	n=13	n=14	n=15	n=16
	0.1	14	17	21	26	30	36
0.025	0.05	11	14	17	21	25	30
0.01	0.02	7	10	13	16	20	24
0.005	0.01	5	7	10	13	16	19
		n=17	n=18	n=19	n=20	n=21	n=22
	0.1	41	47	54	60	68	75
0.025	0.05	35	40	46	52	59	66
0.01	0.02	28	33	38	43	49	56
0.005	0.01	23	28	32	37	43	49
		n=23	n=24	n=25	n=26	n=27	n=28
	0.1	83	92	101	110	120	130
0.025	0.05	73	81	90	98	107	117
0.01	0.02	62	69	77	85	93	102
0.005	0.01	55	61	68	76	84	92
		n=29	n=30	n=31	n=32	n=33	n=34
	0.1	141	152	163	175	188	201
0.025	0.05	127	137	148	159	171	183
0.01	0.02	111	120	130	141	151	162
0.005	0.01	100	109	118	128	138	149
		n=35	n=36	n=37	n=38	n=39	
	0.1	214	228	242	256	271	
0.025	0.05	195	208	222	235	250	
0.01	0.02	174	186	198	211	224	
0.005	0.01	160	171	183	195	208	
		n=40	n=41	n=42	n=43	n=44	n=45
	0.1	287	303	319	336	353	371
0.025	0.05	264	279	295	311	327	344
0.01	0.02	238	252	267	281	297	313
0.005	0.01	221	234	248	262	277	292
		n=46	n=47	n=48	n=49	n=50	
	0.1	389	408	427	446	466	
0.025	0.05	361	379	397	415	434	
0.01	0.02	329	345	362	380	398	
0.005	0.01	307	323	339	365	373	

Table 3 Critical values of Spearman's Rank Correlation Coefficient

The alpha-values correspond to a one-tailed test of Null hypothesis. The value should be doubled for two-tailed tests

n	alpha=0.05	alpha=0.025	alpha=0.01	alpha=0.005
5	0.900			
6	0.829	0.886	0.943	
7	0.714	0.786	0.893	
8	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.523	0.623	0.736	0.818
12	0.497	0.591	0.703	0.780
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.490	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.450	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.400	0.475	0.526
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.448	0.496
29	0.311	0.370	0.440	0.487
30	0.305	0.364	0.432	0.478

Index