

H.N.Aliyev

**Theory of Probability
and Statistics**

Solved problems and exercises

To the students

Nowadays, all companies use statistical methods in making decisions. Consequently, the study of statistical methods has taken on a prominent role in the education of student majoring in management and economics. Here is some advice that will help you to succeed in statistics (and in other subjects too).

Tip1: Understanding the process of solving a particular type of problem is emphasized on memorizing formulas. In most cases, if you understand the concepts, memorizing a formula becomes completely unnecessary, because you construct the necessary tools when needed.

Tip2: Classes are held for your benefit. If attending class was not important, all university courses would be by correspondence, and your tuition would be much lower. During class your instructor will go over examples, which are important, and most likely not in the book. Statistics courses are sequential, so the stuff you see in, for example lecture 6, will enable you to make sense of a much material you will see in lecture 7. As instructors, we note a definite correlation between grades and class attendance. **Go to class!!**

Tip3: Statistics books are not meant to be read like novels (even though they are often exciting). It is better generally to read the sections of the book to be covered in lecture through quickly to get some idea of what there is before going to lecture. After the lecture read it through carefully, with pencil and paper in hand, working through examples.

Tip4: Just as you must play a lot of football to be good at it, you must do a lot of statistics problems in order to be successful as well. At minimum, work on every problem your instructor suggests. If you are having trouble or want more practice, work on other problems in that section or get another book and work problems out of it. If you are

having trouble getting a correct answer to a problem, think about what is going wrong. By doing these you can learn something new and prevent yourself from making the same error in the future. Work on problems more than once. Work on problems until you can do them quickly. Remember, the process is usually more important than the result.

Tip5: The fastest way to get into trouble in statistics is to not do homework. Remember, similar problems will probably show up on quizzes and exams, where you will be expected to work them quickly and accurately, probably without the book in front of you.

Tip6: Contrary to many students' opinions, your instructor wants you to succeed. Extremely rare is the instructor who will intentionally put completely different material on an exam that was covered in class. For this reason, pay attention to your instructor and take notes. Then read your notes, and be sure you understand them, filling any missing details. Review your notes regularly.

The goal of this book is to present statistics in a clear and interesting way. Students, who will use this book, are not required to have a strong background in mathematics. The chapters are divided into sections, and each section contains necessary theoretical background and solved problems. A set of exercises appears at the end of each section. Answers are right after exercises.

In the end, any suggestion from readers would be greatly appreciated.

Dr. Humbet Aliyev

Chapter 1

Organization and description of data

1.1. Introduction

Statistics is a group of methods that are used to collect, analyse, present, interpret data and make decisions.

Statistics is sometimes divided into two main areas:

1. Descriptive statistics
2. Inferential statistics.

Descriptive statistics consists of the collection, organization, summation, and presentation of data.

A **population** is a complete set of units (usually people, objects, events) that we are interested in studying.

A subset of the population selected for study is called **a sample**.

Inferential statistics is an estimate or prediction about a population based on information contained in a sample.

1.2. The mean

The mean for ungrouped data, also known as the arithmetic average, is found by adding the values of the data and dividing by the total number of values. Thus,

$$\text{Mean for population data: } \mu = \frac{\sum_{i=1}^N x_i}{N}$$
$$\text{Mean for sample data: } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where N – is the population size, n – is sample size, μ – (Greek letter mu) is the population mean, and \bar{x} – (read as “ x -bar”) is the sample mean.

Example:

Calculate the mean of the following six sample observations:

$$5, 2, 6, 8, 7, 8$$

Solution:

Using the definition of sample mean, we find

$$\bar{x} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{5 + 2 + 6 + 8 + 7 + 8}{6} = \frac{36}{6} = 6.$$

Thus, the mean of this sample is 6.

Example:

The salaries of all 7 employees of a small company are:

$$\$ 320, 410, 310, 480, 530, 370, 240$$

Find the mean salary.

Solution:

Since the given data set includes all 7 employees of the company, it represents the population. Hence, $N = 7$. The population mean is

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{320 + 410 + 310 + 480 + 530 + 370 + 240}{7} = \frac{2660}{7} = \$380.$$

Thus, the mean salary of the employees of this company is \$380.

1.3. The median

The median is the middle term in a data set. Before one can find this point the data must be arranged in increasing (or decreasing) order. The calculation of the median for ungrouped data consists of the following two steps:

1. Rank the given data set in increasing (or decreasing) order.

2. Find $\left(\frac{n+1}{2}\right)^{th}$ term in a ranked data set.

The value of $\left(\frac{n+1}{2}\right)^{th}$ term is the median.

There are two possibilities

- 1) If n is odd, then the median is given by the value of the middle term in a ranked data.
- 2) If n is even, then the median is given by the average of the values of the two middle term.

Remark: If the given data set represents a population, replace n by N .

Example:

Consider again the seven salaries of employees of a small company

\$ 320, 410, 310, 480, 530, 370, 240

Calculate the median of this population.

Solution:

First of all, let us rank salaries in ascending order:

\$ 240, 310, 320, 370, 410, 480, 530

$$N=7 \quad \text{and} \quad \left[\frac{N+1}{2} \right]^{th} = \left[\frac{7+1}{2} \right]^{th} = 4^{th}$$

Therefore, the median is the value of the fourth term in the ranked data

\$ 240, 310, 320, 370, 410, 480, 530

↑
Median

Thus, the median value for this population is \$370.

Example:

The ages of a sample of 10 university students are

18, 22, 19, 20, 21, 18, 22, 19, 23, 17

Calculate the median of this sample.

Solution:

First we order the data in increasing order. The ordered values are

17, 18, 18, 19, 19, 20, 21, 22, 22, 23

There are 10 values in the data set. Hence,

$$n = 10 \quad \text{and} \quad \left[\frac{n+1}{2} \right]^{th} = \left[\frac{10+1}{2} \right]^{th} = 5.5^{th}$$

Therefore, the median is given by the mean of fifth and sixth values in the ranked data.

5^{th} value = 19, 6^{th} value = 20

$$\text{Median} = \frac{19+20}{2} = 19.5.$$

Hence, the median age is 19.5.

1.4. The Mode

The value that occurs most often in a data set is called the mode.

Example:

The following data give the GPA of 7 students

3.6; 3.2; 2.8; 3.6; 3.8; 3.6; 2.9

Find the mode.

Solution:

It is helpful to arrange the data in order, although it is not necessary

2.8; 2.9; 3.2; 3.6; 3.6; 3.6; 3.8

Since 3.6 occurs three times and 3.6 has a frequency larger than any other number- the mode for the data set is 3.6.

A data set can have more than one mode or no mode at all, whereas it will have only one mean and only one median.

A data set with each value occurring only once has no mode.

A data set with two (or more) values occurring with the same (highest) frequency has two (or more) modes.

Example:

Find the mode of the set data set:

2, 3, 3, 1, 2, 7, 7, 3, 1, 2

Solution: Since 2 and 3 both occur three times, the modes are 2 and 3.

This data set is said to be bimodal.

Example:

Last year's income of six randomly selected families were

210.000, 300.000, 325.000, 280.000, 315.000, 410.000

Find the mode.

Solution:

Since each value occurs only once, there is no mode.

Remark: One advantage of the mode is that it can be calculated for both kinds of data, quantitative and qualitative, whereas the mean and median can be calculated only for quantitative data.

Example:

6 students are selected at random. Their statuses are:

Senior, sophomore, senior, junior, senior, sophomore.
Find the mode.

Solution:

Since senior occurs more frequently than the other categories, it is the mode for this data set. However, we can not calculate the mean and median for this data set.

For a data set, the mean, median, and mode can be quite different.

Consider the following example.

Example:

The number of days the first six heart transplant patients survived after their operations were

$$15; 3; 46; 623; 126; 64$$

Find the mean, median, and mode.

Solution:

The sample mean is

$$\bar{x} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{15 + 3 + 46 + 623 + 126 + 64}{6} = \frac{877}{6} = 146.2 \text{ days}$$

To find median, first we rank the data. The ordered values are

$$3, 15, 46, 64, 126, 623 \quad n = 6$$

$$\text{Median} = \left[\frac{n+1}{2} \right]^{\text{th}} = \left[\frac{6+1}{2} \right]^{\text{th}} = 3.5^{\text{th}}$$

$$\text{Median} = \frac{\text{3}^{\text{rd}} \text{ value} + \text{4}^{\text{th}} \text{ value}}{2} = \frac{46 + 64}{2} = 55 \text{ days}.$$

Since each value occurs only once, there is no mode.

In this example, the mean is much higher than the median. Only 1 out of 6

patients survived longer than $\bar{x} = 146.2$ days. It is because of only large survival time that greatly inflates the mean. In this and similar situations, the median should be used as the measure of central tendency.

To sum up, we can not conclude which of the three measures of central tendency is a better measure overall. Each of them may be better under different situations. But the mean is the most used measure of central tendency. The advantage of the mean is that its calculation includes each value of the data set.

Exercises

1. Calculate the mean and median for each of the following data set.

- a) 6, 10, 7, 14, 8
- b) 2, 1, 4, 2, 1

2. The following data set belongs to a population

5, -7, 2, 0, -9, 12, 10, 7

Calculate the mean, median and mode.

3. Find the mode of each of the following samples

- a) 5, 8, 11, 9, 8, 6, 8
- b) 7, 12, 8, 7, 10, 11, 8, 6, 10, 13, 7, 8

4. Twelve secretaries were given a typing test, and the time (in minutes) to complete it were as follows:

8, 12, 15, 9, 6, 8, 10, 9, 8, 6, 7, 8

Find the mean, median, and mode.

5. The grade point average (GPA) of 10 students who applied for financial aid are shown below

3.15, 3.62, 2.54, 2.81, 3.97, 1.85, 1.93, 2.63, 2.50, 2.80

Find the mean, median, and mode.

6. During a year, the major earthquakes had Richter magnitudes as shown below

7.0, 6.2, 7.7, 8.0, 6.4, 6.2, 7.2, 5.4, 6.4, 6.5, 7.2, 5.4

Find the mean, median, and mode.

7. Eight participants in a bike race had following finishing times in minutes
28, 22, 26, 33, 21, 23, 37, 24

Find the mean and median for the finishing times.

8. The monthly income in dollars for seven families are

950, 775, 925, 2500, 1150, 850, 975

a) Calculate the mean and median income

b) Which of the two is preferable as a measure of center,
and why?

9. The numbers of defective parts observed on 16 different days are shown below:

11, 14, 18, 14, 21, 17, 13, 21, 25, 19, 17, 13, 28, 13, 17, 18

What statistical measures of central tendency would help summarize this data set?

What information would you report to the production manager?

10. Consider the following two data sets:

Data set I: 10, 12, 17, 8, 15

Data set II: 13, 15, 20, 11, 18

Notice that each value of the second data set is obtained by adding 3 to the corresponding value of the first data set. Calculate the mean for each of these data sets. Comment on the relationship between the two means.

11. Consider the following two data sets:

Data set I: 2, 5, 7, 9, 8

Data set II: 6, 15, 21, 27, 24

Notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 3. Calculate the mean for each of these data sets. Comment on the relationship between the two means.

Answers

1. a) $\bar{x} = 9$; Median = 8; b) $\bar{x} = 2$; Median = 2 ; **2.** $\bar{x} = 2.5$;

Median = 3.5; no mode ; **3.** a) 8; b) 7, 8; **4.** $\bar{x} = 8.83$; Median = 8;

mode = 8; **5.** $\bar{x} = 2.78$; Median = 2.715; no mode ; **6.** $\bar{x} = 6.6$;

Median = 6.45; no mode ; **7.** $\bar{x} = 26.75$; Median = 25;

8. a) $\bar{x} = 1160.7$; Median = 950 .

1.5. Measures of dispersion for ungrouped data

In statistics, in order to describe the data set accurately statisticians must know more than measures of central tendency. Two data sets with the same mean may have completely different spreads. The variation among values of observations for one data set may be much larger or smaller than for the other data set.

Remark:

The words dispersion, spread, and variation have the same meaning.

Example:

Consider the following two samples:

Sample1: 66, 66, 66, 67, 67, 67, 68, 69

Sample2: 43, 44, 50, 54, 67, 90, 91, 97

The mean of sample1 is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{536}{8} = 67$$

The mean of sample2 is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{536}{8} = 67.$$

Each of these samples has a mean equal to 67. However, the dispersion of the observations in the two samples differs greatly. In the first sample all observations are grouped within 2 units of the mean. Only one observation (67) is closer than 13 units to the mean of the second sample, and some are as far away as 30 units. Thus, the mean, median, or mode is usually not by itself a sufficient measure to reveal the shape of the distribution of a data set. We also need a measure that can provide some information about the variation among data values. The measures that help us to know about the spread of data set are called the measures of dispersion. The measures of central tendency and dispersion taken together give a better picture of a data set than measure of central tendency alone. Several quantities that are used as measures of dispersion are the **range**, the **mean absolute deviation**, the **variance**, and the **standard deviation**.

1.5.1. Range

The simplest measure of variability for a set of data is the range.

Definition:

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

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

Example:

Find the range for the sample observations

$$13, 23, 11, 17, 25, 18, 14, 24$$

Solution:

We see that the largest observation is 25 and the smallest observation is 11. The range is $25 - 11 = 14$.

Example:

A sample is composed of the observations

$$67, 79, 87, 97, 93, 57, 44, 80, 47, 78, 81, 90, 88, 91$$

Find the range.

Solution:

The largest observation is 97; the smallest observation is 44.

The range is $97 - 44 = 53$.

1.5.2. The mean absolute deviation

The mean absolute deviation is defined exactly as the words indicate. The word “deviation” refers to the deviation of each member from the mean of the population. The term “absolute deviation” means the numerical (i.e. positive) value of the deviation, and the “mean absolute deviation” is simply the arithmetic mean of the absolute deviations.

Let $x_1, x_2, x_3, \dots, x_N$ denote the N members of a population, whose mean is μ . Their mean absolute deviation, denoted by $M.A.D.$ is

$$M.A.D. = \frac{\sum_{i=1}^N |x_i - \mu|}{N}$$

For the sample of n observations, with mean \bar{x} , mean absolute deviation is defined analogously

$$M.A.D. = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

To calculate mean absolute deviation it is necessary to take following steps:

1. Find \bar{x} (or μ)
2. Find and record the signed differences
3. Find and record the absolute differences
4. Find $\sum_{i=1}^n |x_i - \bar{x}|$ (or $\sum_{i=1}^N |x_i - \mu|$)
5. Find the mean absolute deviation.

Example:

Suppose that sample consists of the observations

21, 17, 13, 25, 9, 19, 6, and 10

Find the mean absolute deviation.

Solution:

Perhaps the best manner to display the computations in steps 1, 2, 3, and 4 is to make use of a table 1.1 composed of three columns

Table 1.1

x_i	$x_i - \bar{x}$	$ x_i - \bar{x} $
21	21-15=6	6
17	17-15=2	2
13	13-15=-2	2
25	25-15=10	10
9	9-15=-6	6
19	19-15=4	4
6	6-15=-9	9
10	10-15=-5	5

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{120}{8} = 15$$

$$M.A.D. = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = \frac{44}{8} = 5.5$$

On the average, each observation is 5.5 units from the sample.

1.5.3. The variance and the standard deviation

A key step in developing a measure of variability that includes all the data items involves the computations of the differences between the data values and the mean for the data set. The difference between x_i and the mean

(\bar{x} for a sample, μ for a population) is called a deviation about the mean. Since we are seeking a descriptive statistical measure that summarizes the variability or dispersion in the entire data set, we want to consider the deviation of each data value about the mean. Thus for a sample size n and data values x_1, x_2, \dots, x_n , we will need to compute the deviations

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x}).$$

We might think of summarizing the dispersion in a data set by computing the average deviation about the mean. The only trouble with such an attempted definition is that it would not give us much information about the

$$\sum_{i=1}^n (x_i - \bar{x})$$

variation present in the data; the mean $\frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$ would be zero for every

sample, because the sum $\sum_{i=1}^n (x_i - \bar{x})$ equals zero for every sample. The

positive and negative deviations cancel each other out. Hence if we are to use the deviations from the mean as a measure of dispersion we must find another approach. As we already know, one way is computing the average absolute deviation as a measure of variability. While this measure is sometimes used, the one most often used is based on squaring the deviations to eliminate the negative values. The average of the squared deviations for a data set representing a population or sample is given a special name in statistics. It is called the **variance**.

The **population variance** is denoted by the Greek symbol σ^2 (pronounced “sigma squared”). The formula for population variance is

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

where x_i – population data

μ – population mean

N – population size.

It is frequently desirable to have a measure of dispersion whose units are the same as those of the observations. Since the variance is given in squared units, the square root of the variance would be given in units that we need. Thus, if we take the square root of the variance, we have the measure of dispersion that is known as the population standard deviation and denoted by σ . By definition we have

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

In many statistical applications, the data set we are working with is a sample. When we compute a measure of variability for the sample, we often are interested in using the sample statistic obtained as an estimate of the population parameter, σ^2 . At this point it might seem that the average of the squared deviations in the sample would provide a good estimate of the population variance. However, statisticians have found that the average squared deviation for the sample has the undesirable feature that it tends to underestimate the population variance σ^2 . Because of this tendency toward underestimation we say it provides a biased estimate.

Fortunately, it can be shown that if the sum of the squared deviations in the sample is divided by $(n - 1)$, and not n , then the resulting sample statistic will provide an unbiased estimate of the population variance. For this reason the **sample variance** is not defined to be the average squared deviation in the sample. Sample variance is denoted by s^2 and is defined as follows:

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

To find the sample standard deviation (denoted by s), one must take the square root of the sample variance:

$$\text{Sample standard deviation } s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Example:

Find the variance and the standard deviation for the sample data
21, 17, 13, 25, 9, 19, 6, and 10

Solution:

When we compute s^2 by applying formula (2), the computations can most conveniently be shown in a table. The table will be composed of three columns: a column for the observations x_i , a column for the deviations of the observations from the sample mean $(x_i - \bar{x})$, and a column for the squared deviations $(x_i - \bar{x})^2$. (Table 1.2)

(Table 1.2)

$$\sum_{i=1}^8 x_i = 120; \bar{x} = \frac{\sum_{i=1}^8 x_i}{n} = \frac{120}{8} = 15$$

$$s^2 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^2}{n-1} = \frac{302}{7} = 43.14;$$

$$\text{and } s = \sqrt{s^2} = \sqrt{43.14} = 6.57.$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
21	21-15=6	36
17	17-15=2	4
13	13-15=-2	4
25	25-15=10	100
9	9-15=-6	36
19	19-15=4	16
6	6-15=-9	81
10	10-15=-5	25
120		302

From the computational point of view, it is easier and more efficient to use **short-cut** formulas to calculate the variance. By using the short-cut formula, we reduce the computation time and round off errors.

The short-cut formulas for calculating variance are as follows:

$$\sigma^2 = \frac{1}{N} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{N} \right] = \frac{1}{N} \left[\sum x_i^2 - N \cdot \mu^2 \right]$$

and

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{n-1} \left[\sum x_i^2 - n \cdot \bar{x}^2 \right]$$

Example:

Find the variance and the standard deviation for the sample of
16, 19, 15, 15, and 14

Solution:

Let us apply

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right]$$

Step1: Find the sum of values,

$$\sum x = 16 + 19 + 15 + 15 + 14 = 79$$

Step2: Square each value and find the sum

$$\sum x^2 = 16^2 + 19^2 + 15^2 + 15^2 + 14^2 = 1263$$

Step3: Substitute in the formula and calculate

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] = \frac{1}{4} \left[1263 - \frac{79^2}{5} \right] = 3.7$$

$$s = \sqrt{3.7} = 1.9$$

Hence the sample variance is 3.7 and sample standard deviation is 1.9.

1.5.4. Interpretation of the population standard deviation

Often in statistical studies we are interested in specifying the percentage of items in a data set that lie within some specified interval when only the mean and standard deviation for the data set are known. Two rules are commonly used for forming such estimates.

The first is true for any data set.

Chebyshev's theorem:

For any set of data and any $k \geq 1$, at least $100 \cdot \left(1 - \frac{1}{k^2}\right)\%$ of the values in

the data set must be within plus or minus k standard deviations of the mean.

Remark:

In applying Chebyshev's theorem we treat every data set as if it were a population, and the formula for a population standard deviation is used.

k	1.5	2	2.5
$100 \cdot \left(1 - \frac{1}{k^2}\right)\%$	55.6%	75%	84%

According to Chebyshev's rule, at least 55.6% of the population data lie within 1.5 standard deviations around the mean, at least 75% of the population data lie within 2 standard deviations around the mean and so on.

Example:

Let $\mu = 70$, $\sigma = 1.5$

If we let $k = 3$ from $100 \cdot \left(1 - \frac{1}{k^2}\right)\%$ we obtain that $100 \cdot \left(1 - \frac{1}{k^2}\right)\% = 100 \left(1 - \frac{1}{9}\right)\% = \frac{8}{9} \cdot 100\% = 88.89\%$.

The theorem states that at least 88.89% of data values will fall within 3 standard deviations of the mean. 88.89% of data falls within $(\mu \pm 3\sigma)$ or

$$70 + 3 \cdot 1.5 = 74.5 \text{ and}$$

$$70 - 3 \cdot 1.5 = 65.5$$

For $\mu = 70$, $\sigma = 1.5$, at 88.89% of the data values fall between 74.5, 65.5.

Rule of Thumb.

When a distribution is bell-shaped the following statements, which are called Thumb rule, are true:

Approximately 68% of the population members lie within one standard deviation of the mean.

Approximately 95% of the population members lie within two standard deviations of the mean.

Approximately 99.7% of the population members lie within three standard deviations of the mean.

For example, suppose that scores on entrance exam have a mean of 480 and standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 ;

$$(480 - 1 \cdot 90 = 390 \text{ and } 480 + 1 \cdot 90 = 570)$$

Approximately 95% of the scores will fall between 300 and 660

$$(480 - 2 \cdot 90 = 300 \text{ and } 480 + 2 \cdot 90 = 660).$$

Approximately 99.7 % of the scores will fall between 210 and 750

$$(480 - 3 \cdot 90 = 210 \text{ and } 480 + 3 \cdot 90 = 750).$$

1.5.5. The interquartile range

Quartiles are the summary measures that divide a ranked data set into four equal parts. Three measures will divide any data set into four equal parts.

These three measures are the first quartile (denoted by Q_1), the second quartile (denoted by Q_2), and the third quartile (denoted by Q_3). The data should be ranked in increasing order before the quartiles are determined. The quartiles are defined as follows:

$$Q_1 = \left[\frac{(n+1)}{4} \right]^{th} \quad \text{- ordered observation}$$

$$Q_3 = \left[\frac{3 \cdot (n+1)}{4} \right]^{th} \quad \text{- ordered observation.}$$

The difference between the third and the first quartiles gives the interquartile range. That is

$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1.$$

Example:

A teacher gives a 20-point test to 10 students. The scores are shown below

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Find the interquartile range.

Solution:

First, we rank the given data in increasing order:

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

$$n = 10$$

$$Q_1 = \left[\frac{(n+1)}{4} \right]^{th} \quad \text{- ordered observation.}$$

$$Q_1 = \left[\frac{10+1}{4} \right]^{th} = \left[2\frac{3}{4} \right]^{th}.$$

Hence, the first quartile is three-quarter way from the 2nd data (3) to the 3rd third (5). Therefore,

$$\text{First quartile} = Q_1 = 3 + \frac{3}{4}(5 - 3) = 3 + \frac{3}{2} = \frac{9}{2}$$

$$\text{Similarly, since } Q_3 = \left[\frac{3(n+1)}{4} \right]^{th} = \left[\frac{3 \cdot 11}{4} \right]^{th} = \left[\frac{33}{4} \right]^{th} = \left[8\frac{1}{4} \right]^{th}$$

The third quartile is one-quarter of the way from the 8th observation (15) to the 9th observation (18). Thus we have

$$\text{Third quartile} = Q_3 = 8^{th} + \frac{1}{4}(9^{th} - 8^{th}) = 15 + \frac{1}{4}(18 - 15) = 15\frac{3}{4}.$$

Finally, the interquartile range is the difference between the third and first quartiles:

$$\begin{aligned}\text{Interquartile range} &= IQR = Q_3 - Q_1 = \\ &= 15 \frac{3}{4} - 3 \frac{3}{2} = \frac{63}{4} - \frac{9}{2} = \frac{45}{4} = 11.25\end{aligned}$$

Example:

The following are the ages of nine employees of an insurance company
47, 28, 39, 51, 33, 37, 59, 24, 33

Find the interquartile range.

Solution:

Let us arrange the data in order from smallest to largest

$$24, 28, 33, 33, 37, 39, 47, 51, 59$$

$$\begin{aligned}Q_1 &= \left[\frac{(n+1)}{4} \right]^{th} = \left[\frac{9+1}{4} \right]^{th} = (2.5)^{th} = 2^{nd} + \frac{1}{2}(3^{rd} - 2^{nd}) = \\ &= 28 + \frac{1}{2}(33 - 28) = 28 + 2.5 = 30.5\end{aligned}$$

$$\begin{aligned}Q_3 &= \left[\frac{3(n+1)}{4} \right]^{th} = \left[\frac{3 \cdot 10}{4} \right]^{th} = \left[7 \frac{1}{2} \right]^{th} = 7^{th} + \frac{1}{2}(8^{th} - 7^{th}) = \\ &= 47 + \frac{1}{2}(51 - 47) = 47 + 2 = 49\end{aligned}$$

The interquartile range is

$$IQR = Q_3 - Q_1 = 49 - 30.5 = 18.5.$$

Exercises

1. Fifteen students were selected randomly and asked how many hours each studied for the final exam in statistics. Their answers are recorded here

$$8, 6, 3, 0, 0, 5, 9, 2, 1, 3, 7, 10, 0, 3, 6$$

- Find the range
- Find the mean absolute deviation
- Find the sample variance and sample standard deviation
- Find the interquartile range.

2. The following data give the hourly wage rate of all 12 employees of a small company

21, 22, 27, 36, 22, 29, 22, 23, 22, 28, 36, 33

- a) Find the population variance and standard deviation
- b) Find the mean absolute deviation
- c) Find the range
- d) Find the interquartile range.

3. The number of words printed in each of 12 randomly selected storybooks for children is listed below

502, 213, 335, 197, 414, 469, 497, 367, 409, 297, 309, 414

- a) Find the sample variance and sample standard deviation
- b) Find the range
- c) Find the mean absolute deviation
- d) Find the interquartile range.

4. The weights of sample of nine football players are recorded as follows:

78, 72, 68, 73, 75, 69, 74, 73, 72

- a) Find the range
- b) Find the variance
- c) Find the standard deviation

5. The following data give the number of cars that stopped at a service station during each of the 10 hours observed

29, 35, 42, 31, 24, 18, 16, 27, 39, 34

Find the range, variance, and standard deviation.

6. The following data give the number of new cars sold at a dealership during a 12-day period

13, 5, 9, 6, 8, 11, 9, 15, 4, 11, 7, 5

Find the range, variance, standard deviation, and interquartile range.

7. Consider the following two data sets:

Data set I : 12, 25, 37, 8, 41

Data set II: 19, 32, 44, 15, 48

Notice that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Calculate the standard deviation for each of these two data sets using the formula for sample data. Comment on the relationship between the standard deviations.

8. Consider the following two data sets:

Data set I : 4, 8, 15, 9, 11

Data set II: 8, 16, 30, 18, 22

Notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 2. Calculate the standard deviation for each of these data sets using the formula for the sample data. Comment on the relationship between the standard deviations.

9. The number of patients treated at the hospital per day are shown below. Data are from a random sample of 12 days:

45, 50, 36, 59, 28, 42, 55, 67, 33, 35, 40, 50

Compute the mean, median, mode, range, variance, and standard deviation for these data.

10. Light bulbs manufactured by a well-known electrical equipment firm are known to have a mean life of 800 hours with a standard deviation of 100 hours.

- Find a range in which it can be guaranteed that 84% of lifetimes of light bulbs lie.
- Using the rule of thumb, find a range in which it can be estimated that approximately 68% of these light bulbs lie.

11. Tires of a particular brand have lifetimes with mean of 29.000 km and standard deviation of 3.000 km.

- Find a range in which it can be guaranteed that 75% of the lifetimes of tire of this brand lie.
- Using the rule of thumb, find a range in which it can be estimated that approximately 95% of the lifetimes of tires of this brand lie.

12. The mean of a distribution is 20 and the standard deviation is 2.

Use Chebyshev's theorem to answer:

- At least what percentage of the values will fall between 10 and 30 ?
- At least what percentage of the values will fall between 12 and 28 ?

13. A sample of hourly wages of employees who work in restaurants in a large city has a mean of \$5.02 and a standard deviation of \$0.09. Using Chebyshev's theorem, find the range in which at least 75% of the data will lie.

14. The average score on a special test of knowledge has a mean of 95 and a standard deviation of 2. Using Chebyshev's theorem, find the range in which at least 88.89 % of the data will fall.

15. During a recent football season, it was reported that the average attendance for games was 45.000. The standard deviation in the attendance figure was $\sigma = 4.000$. Use Chebyshev's theorem to answer the following:

- a) Develop an interval that contains the attendance figure for at least 75% of the games.
- b) The commissioner claims that at least 90% of the games had attendances between 29.000 and 61.000. Is this statement warranted given information we have?

Answers

1. a) 10; b) 2.8; c) 11.3; 3.4; d) 6; 2. a) 29.52; 5.43; b) 4.75; c) 15; d) 10; 3. a) $s^2 = 10325.9$; $s = 101.62$; b) 305; c) 82.25; d) 156.5; 4. a) 10; b) 9; c) 3 ; 5. range = 26 ; b) $s^2 = 72.25$; c) $s = 8.50$; 6. range = 11; $s^2 = 11.720$; $s = 3.42$; 7. $s = 14.64$ for both data sets; 8. $s_1 = 4.04$ and $s_2 = 8.08$; 9. 45; 43.5; 50; range = 39 ; $s^2 = 134.36$; $s = 11.59$; 10. a) 550-1050; b) 700-900; 11. a) 23.000-35.000; b) 23.000-35.000; 12. a) $1 - \frac{1}{5^2} = 0.96$ or 96%; b) 0.9375 or 93.75% ; 13. \$4.84-\$5.20; 14. 89-101; 15. a) 37.000-53.000.

1.6. Numerical summary of grouped data

1.6.1. Mean for data with multiple-observation values

Suppose that a data set contains values m_1, m_2, \dots, m_k occurring with frequencies, f_1, f_2, \dots, f_k respectively.

1. For a population of N observations, so that

$$N = \sum_{i=1}^k f_i$$

The mean is

$$\mu = \frac{\sum_{i=1}^k f_i \cdot m_i}{N}$$

2. For a sample of n observations, so that

$$n = \sum_{i=1}^k f_i$$

The mean is

$$\bar{x} = \frac{\sum_{i=1}^k f_i \cdot m_i}{n}$$

The arithmetic is most conveniently set out in tabular form.

Example:

The score for the sample of 25 students on a 5-point quiz are shown below.
Find the mean.

Solution:

We must find $\bar{x} = \frac{\sum_{i=1}^6 f_i \cdot m_i}{n}$. We need a

column to display the computation of the quantity $f_i \cdot x_i$ (Table 1.3):

Score (m_i)	Frequency (f_i)
0	1
1	2
2	6
3	12
4	3
5	1
	$n = 25$

Table 1.3

In the end,

$$\bar{x} = \frac{\sum_{i=1}^6 f_i \cdot m_i}{n} = \frac{67}{25} = 2.68 \approx 2.7.$$

Hence the mean of the scores
Is approximately 2.7.

Score (m_i)	Frequency (f_i)	$f_i \cdot m_i$
0	1	$0 \cdot 1 = 0$
1	2	$1 \cdot 2 = 2$
2	6	$2 \cdot 6 = 12$
3	12	$3 \cdot 12 = 36$
4	3	$4 \cdot 3 = 12$
5	1	$5 \cdot 1 = 5$
	$n = 25$	$\sum_{i=1}^6 f_i \cdot m_i = 67$

1.6.2. Median for data with multiple-observation values

For an ungrouped frequency distribution, find the median by examining the cumulative frequency to locate the middle value, as shown in the next example.

Example:

The number of videocassette recorders sold per month over a two-year period is recorded below. Find the median.

Solution:

As we know the median is $\left[\frac{n+1}{2} \right]^{th}$ observation.

Since $n = 24$ then median = $\left[\frac{24+1}{2} \right]^{th} = \frac{12^{th} + 13^{th}}{2}$.

To find 12^{th} and 13^{th} observations we write corresponding cumulative frequency distribution (Table 1.4).

Table 1.4

Class	Number of sets sold	Frequency (month)	Cumulative frequency
1	1	3	3
2	2	8	11
3	3	5	16
4	4	4	20
5	5	2	22
6	6	1	23
7	7	1	24
		$n = 24$	

The 12^{th} and 13^{th} values fall in class 3.

12^{th} value=3 ; 13^{th} value=3.

Therefore, Median = $\frac{3+3}{2} = 3$.

1.6.3. Mode for data with multiple-observation values

As we already know, the mode is the most frequently occurring value. A similar concept can be used when the data are available in multiple-observation form.

Example:

The following data were collected on the number of blood tests a hospital conducted for a random sample of 50 days. Find the mode.

Number of tests per day	Frequency (days)
26	5
27	9
28	12
29	18
30	5
31	0
32	1

Solution:

Since 29 days were given on 18 days (the number of tests that occurs most often), the mode is 29.

1.6.4. Variance for data with multiple-observation values

Suppose that a data set contains values m_1, m_2, \dots, m_k occurring with frequencies, f_1, f_2, \dots, f_k respectively.

1. For a population of N observations, so that

$$N = \sum_{i=1}^k f_i$$

The variance is

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (m_i - \mu)^2}{N} = \frac{\sum_{i=1}^k f_i \cdot m_i^2}{N} - \mu^2$$

The standard deviation is $\sigma = \sqrt{\sigma^2}$.

2. For a sample of n observations, so that

$$n = \sum_{i=1}^k f_i$$

The variance is

$$s^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \bar{x}^2 \right]$$

The standard deviation is $s = \sqrt{s^2}$.

The arithmetic is most conveniently set out in tabular form.

Example:

The score for the sample of 25 students on a 5-point quiz are shown below. Find a sample variance and standard deviation.

Score (m_i)	Frequency (f_i)
0	1
1	2
2	6
3	12
4	3
5	1

Solution:

$$\sum_{i=1}^6 f_i \cdot m_i$$

Remark: The denominator in the formula $\bar{x} = \frac{\sum_{i=1}^6 f_i \cdot m_i}{n}$ is obtained by

summing the frequencies $(\sum_i f_i = n)$. It is not number of classes.

To calculate variance we need three columns to display the computation of the quantities $(m_i - \bar{x})^2$ a column for the m_i , a column for the $(m_i - \bar{x})$ and a column for the $(m_i - \bar{x})^2$. We also need a column for f_i and a final column for the products $f_i \cdot (m_i - \bar{x})^2$. (Table 1.5)

The necessary computations for finding $\sum_{i=1}^k f_i (m_i - \bar{x})^2$ are shown below.

Table 1.5

Score (m_i)	Frequency (f_i)	$(m_i - \bar{x})$	$(m_i - \bar{x})^2$	$f_i \cdot (m_i - \bar{x})^2$
0	1	0-2.7=-2.7	7.29	0· 7.29=0
1	2	1-2.7=-1.7	2.89	1· 2.89=2.89
2	6	2-2.7=-0.7	0.49	2· 0.49=0.98
3	12	3-2.7=0.3	0.09	3· 0.09=0.27
4	3	4-2.7=1.3	1.69	4· 1.69=6.76
5	1	5-2.7=2.3	5.29	5· 5.29=26.45
	25			$\sum_{i=1}^6 f_i (m_i - \bar{x})^2 = 37.35$

Thus we have

$$s^2 = \frac{\sum_{i=1}^6 f_i (m_i - \bar{x})^2}{n-1} = \frac{37.35}{24} = 1.56$$

$$s = \sqrt{1.56} = 1.25 .$$

Example:

The number of television sets sold per month over a two year period is reported below. Find the variance and standard deviation for the data.

Number of sets sold (m_i)	Frequency (month) (f_i)
5	2
6	3
7	8
8	1
9	6
10	4

Solution:

Let us apply

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \bar{x}^2 \right].$$

Make a table as shown below

Sets (m_i)	Frequency (f_i)	$m_i \cdot f_i$	m_i^2	$f_i \cdot m_i^2$
5	2	$5 \cdot 2 = 10$	25	50
6	3	$6 \cdot 3 = 18$	36	108
7	8	$7 \cdot 8 = 56$	49	392
8	1	$8 \cdot 1 = 8$	64	64
9	6	$9 \cdot 6 = 54$	81	486
10	4	$10 \cdot 4 = 40$	100	400
	$n = 24$	$\sum_{i=1}^6 m_i \cdot f_i = 186$		$\sum_{i=1}^6 f_i \cdot m_i^2 = 1500$

$$\bar{x} = \frac{\sum_{i=1}^6 f_i \cdot m_i}{n} = \frac{186}{24} = 7.75$$

$$s^2 = \frac{1}{23} [1500 - 24 \cdot (7.75)^2] = 2.5$$

To find standard deviation we take the square root of variance

$$s = \sqrt{s^2} = \sqrt{2.5} = 1.6.$$

Exercises

1. The following numbers of books were read by each of the 28 students in a literature class.

- a) Find the mean
- b) Find the median
- c) Find the mode
- d) Find the variance and standard deviation.

Number of books	Frequency (students)
0	2
1	6
2	12
3	5
4	3

2. The all forty students in a class found the following figures for number of hours spent studying in the week before final exam

- a) Find the mean time for study
- b) Find the median
- c) Find the mode
- d) Find the variance and standard deviation for this population.

Time (hours)	Number of students
1	1
2	7
3	15
4	10
5	7

3. A sample of fifty personal property insurance policies found the following numbers of claims over the past 2 years

Number of claims	0	1	2	3	4	5	6
Number of policies	21	13	5	4	2	3	2

- a) Find the mean number of claims per day policy
- b) Find the sample median of claims
- c) Find the modal number of claims for this sample
- d) Find the sample variance and standard deviation.

4. For sample of 50 antique car owners, the following numbers of cars' ages was obtained

- a) Find the mean age of cars
- b) Find the median
- c) Find the modal number
- d) Find the sample variance and standard deviation.

Ages (in years)	Frequency (cars)
17	20
18	18
19	8
20	4

5. The following data represents the net worth (in million of dollars) of 45 national corporations

- a) Find the sample mean net worth
- b) Find the median
- c) Find the mode
- d) Find the sample variance and standard deviation

Net worth (in million of dollars)	Frequency
15	2
20	8
25	15
30	7
35	10
40	3

Answers.

1. a) 2.04; b) 2; c) 2; d) 1.09; 1.04; 2. a) 3.375; b) 3; c) 3; d) 1.08; 1.04;
3. a) 1.4; b) 1; c) 0; d) 3.061; 1.75; 4. a) 17.92; b) 18; c) 17; d) 0.89; 0.94; 5. a) 27.7; b) 25; c) 25; d) 41.98; 6.48.

1.7. Frequency distribution. Grouped data and histograms

Suppose a researcher wished to do study on the monthly earnings of sample of 50 employees of a large company. The researcher would first have to collect the data by asking each of 50 employees. When data are collected in original form, they are called **raw data**. In this case, the data are as follows:

405	510	520	880	820	780	810	580	555
790	505	610	620	650	680	350	530	495
480	695	610	710	810	525	530	680	705

370	760	590	705	300	590	390	460	590
450	540	690	480	420	410	595	750	620
850	585	690	570	560				

Many persons do not like to examine a mass of numbers, and many others do not have the time to do so. Therefore, it would be advantageous if the information could somehow be "compressed" so that the distribution of the observations could be seen at a glance. We find, after some searching, that the smallest observation is 300 and the largest observation is 880. Let us group the observations. We could subdivide the range of data and count the number of values in each subinterval. If the lowest and highest values in a data set are known, the following expression often is helpful in determining both the width of the class interval and the number of classes desired:

$$(1) \text{ Approximate number of classes} = \frac{\text{Highest value} - \text{Lowest value}}{\text{Width of class}}$$

Using this formula with a trial class width of 100 shows that

$$\frac{880 - 300}{100} = 5.8$$

Rounding up, we find that 6 classes would be required for the data.

Table 1.6

Monthly earnings (in dollars)	Number of employees
301- 400	4
401- 500	8
501 - 600	16
601 - 700	10
701 - 800	7
801 - 900	5

The numbers 301, 400, 401, 500 are known as class limits. To find the midpoint of the upper limit of the first class and the lower limit of the second class in table 1.6 we divide the sum of these two limits by 2. Thus, midpoint is

$$\frac{400+401}{2} = 400.5$$

The value 400.5 is called the upper boundary of the first class and the lower boundary of the second class. By using this technique, we can convert the class limits of table 1.7 to class boundaries, which are also called real class limits.

Table 1.7

Class limits	Class boundaries	Class width	Class midpoint
301 to 400	300.5-400.5	100	350.5
401 to 500	400.5-500.5	100	450.5
501 to 600	500.5-600.5	100	550.5
601 to 700	600.5-700.5	100	650.5
701 to 800	700.5-800.5	100	750.5
801 to 900	800.5-900.5	100	850.5

Definition:

The class boundary is given by the midpoint of the upper limit of one class and the lower limit of the next class.

Definition:

The difference between the two boundaries of a class is called the class width.

$$\text{Class width} = \text{Upper boundary} - \text{Lower boundary}$$

Definition:

The class midpoint (or mark) is the average of the two limits (or two boundaries)

$$\text{Class midpoint(or mark)} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Remark:

Other class widths may be considered in (1); the decision on the class width and the number of classes is up to the user.

Definition:

A frequency distribution is a table used to organize data. The left column (called classes or groups) included numerical intervals on a variable being studied. The right column is a list of the frequencies, or number of observations, for each class. Data presented in the form of a frequency distributions are called **grouped data**.

The subintervals into which the data are broken down are called classes. In this distribution the values 300 and 400 of the first class are called class limits. For any particular class, the **cumulative frequency** is the total number of observations in that and previous classes. (Table 1.8)

Table 1.8

Monthly earnings (in dollars)	Number of employees	Cumulative frequencies
301 to 400	4	4
401 to 500	8	12
501 to 600	16	28
601 to 700	10	38
701 to 800	7	45
801 to 900	5	50

Definition:

Relative frequency is the proportion of observations in each class. It is defined as:

$$\text{Relative frequency of a class} = \frac{\text{frequency of that class}}{\text{sum of all frequencies}} = \frac{f_i}{\sum_{i=1}^n f_i}$$

In addition, we often want to consider the proportion of observations that are either in that or one of the earlier classes. These proportions are called **cumulative relative frequencies**.

Example in the table 1.9 illustrates how to construct relative frequency and cumulative relative frequency distributions.

Table 1.9

Monthly earnings (in dollars)	Number of employees	Cumulative frequencies	Relative frequencies	Cumulative relative frequencies
301 but less than 400	4	4	4/50	4/50=0.08
401 but less than 500	8	12	8/50	12/50=0.24
501 but less than 600	16	28	16/50	28/50=0.56
601 but less than 700	10	38	10/50	38/50=0.76
701 but less than 800	7	45	7/50	45/50=0.9
800 but less than 900	5	50	5/50	50/50=1
	50		50/50=1	50/50=1

Definition:

A histogram is a graph in which classes are marked on a horizontal axis and either the frequencies, relative frequencies, or cumulative relative frequencies are marked on the vertical axis. The frequencies, relative frequencies, or cumulative relative frequencies are represented by the heights of the bars. In a histogram, the bars are drawn adjacent to each other.

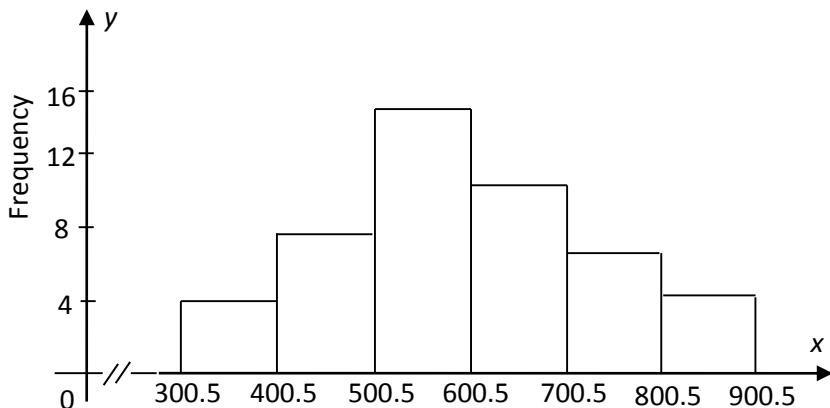


Fig. 1.1 Monthly earnings salaries

Remark: frequency histogram

The symbol “-//–“ used in the horizontal axis represents a break, called the truncation, in the horizontal axis. It indicates that entire horizontal axis is not shown in this figure. As can be noticed, the zero to 300.5 portion of the horizontal axis has been omitted in the figure 1.1.

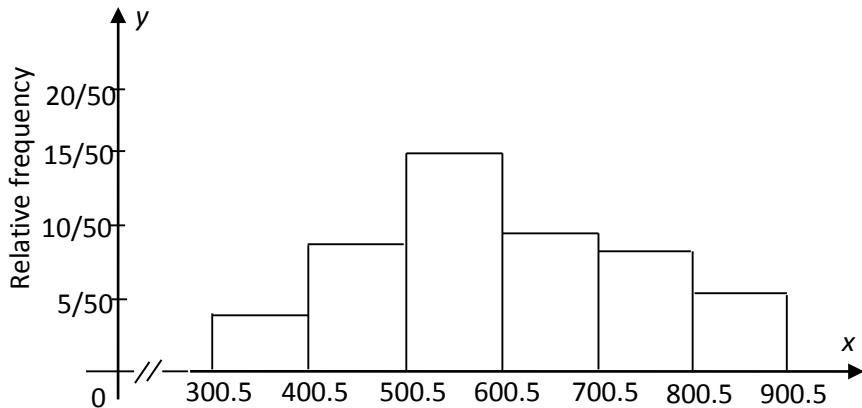


Fig. 1.2 Relative frequency for monthly earnings salaries

As shown in the figure 1.2., we see, for example, that 16/50 of all employees monthly earnings are between 500.5 and 600.5.

The cumulative relative frequencies are the cumulated sums of the relative frequencies. For the first class, the cumulative relative frequency is the same as the relative frequency. For subsequent classes, the cumulative relative frequency for the class to the cumulative relative frequency is obtained by adding the relative frequency for the class to the cumulative relative frequency of the previous class.

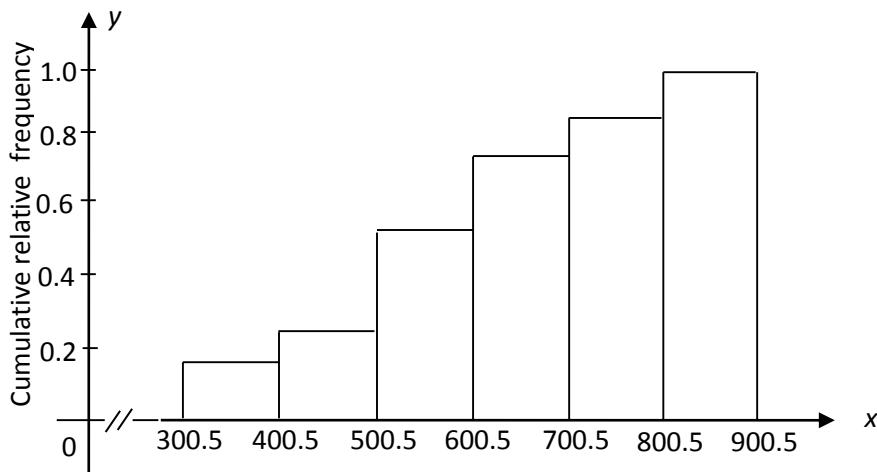


Fig. 1.3 Cumulative relative frequency for monthly-earnings salaries

The interpretation of these quantities is very valuable. For example, 38/50 of all employees' monthly earnings are less than 700.5. The information contained in the cumulative relative frequency can also be presented pictorially, as in Fig. 1.3

1.7.1. *Less than* method for writing classes

The classes in frequency distribution given in table 1.9 for the data on monthly-earning salaries for 50 employees were written as 301-400,

401-500, etc. Alternatively, we can write the classes in a frequency distribution table using the *less than* method. The technique for writing classes in previous topic is more commonly used for data sets that do not contain fractional values. The *less than* method is more appropriate when a data set contains fractional values.

Example:

The following data give the hourly wage rates for a sample of 30 employees selected from a population.

12.25	9.20	13.90	8.10	7.30	7.25	8.75
5.20	15.85	11.20	10.20	14.50	10.50	8.25
7.45	10.20	12.20	10.80	9.25	14.35	16.50
6.40	15.20	10.30	11.75	12.45	13.25	10.80
10.35	9.75					

Construct a frequency distribution table. Find the relative frequency distribution table. Find the relative frequency and cumulative frequencies.

Solution:

The minimum value in data set is 5.20 and the maximum value is 16.50.

Suppose we decide to group these data using six classes of equal width.

Then

$$\text{Approximate width of class} = \frac{16.50 - 5.20}{6} = 1.883$$

We round this number to a more convenient number, say 2.

Then we take 2 as the width of each class. If we start the first class at 5, the classes will be written as 5 to less than 7, 7 to less than 9, and so on as it shown in table 1.10.

Table 1.10

Hourly wage rate (in dollars)	<i>f</i>	Relative frequencies	Cumulative relative frequencies
5 but less than 7	2	$2/30=0.067$	$2/30=0.067$
7 but less than 9	6	$6/30=0.2$	$8/30=0.267$
9 but less than 11	10	$10/30=0.333$	$18/30=0.6$
11 but less than 13	5	$5/30=0.167$	$23/30=0.767$
13 but less than 15	4	$4/30=0.133$	$27/30=0.9$
15 but less than 17	3	$3/30=0.1$	$30/30=1$
	30	Sum=1.00	Sum=1

A histogram for frequencies can be drawn in the same way as for the data of table 1.10. (Fig.1.4; Fig.1.5; Fig.1.6)

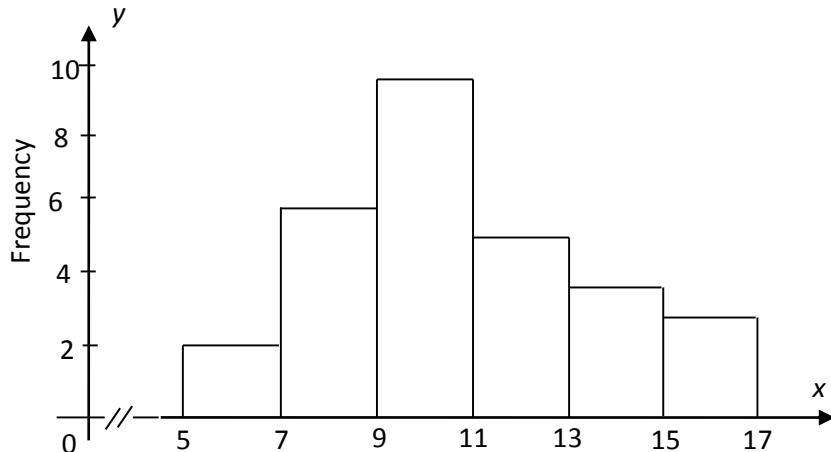


Fig. 1.4 Hourly wage rate

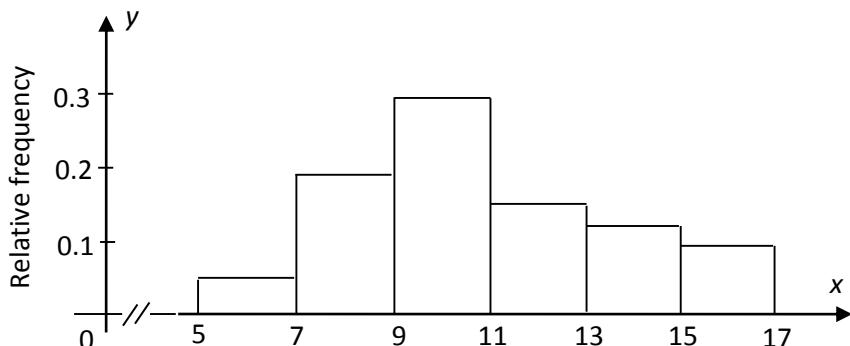
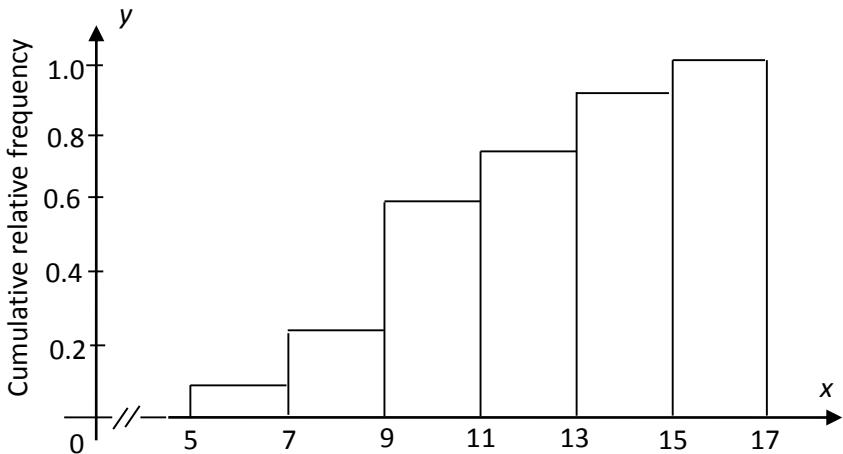


Fig. 1.5 Relative frequency for hourly wage rate



**Fig. 1.5 Cumulative relative frequency for hourly wage rate
Exercises**

1. In survey of 20 patients who smoked, the following data were obtained. Each value represents the number of cigarettes the patient smoked per day. Construct a frequency distribution, using six classes.

10	8	6	14	22	13	17	19	11
9	18	14	13	12	15	15	5	11
16	11							

Draw frequency, relative frequency and cumulative relative frequency histograms.

2. For 75 employees of a large department store, the following distribution for years of service was obtained. Construct histograms for frequency, relative frequency, and cumulative relative frequency.

Class limits	Frequency
1-5	21
6-10	25
11-15	15
16-20	0
21-25	8
26-30	6

3. In a study of _____ 32 student grade point averages (GPA), the following data were obtained.

3.2	2.0	3.3	2.7	2.1	3.9	1.1	3.5	1.9
1.7	0.8	2.6	0.6	4.0	3.5	2.3	1.6	2.8
2.6	1.6	1.6	2.4	2.6	2.3	3.8	2.1	2.9
3.0	1.7	4.0	1.2	3.1				

Construct histograms for frequency, relative frequency and cumulative relative frequency.

4. To determine their lifetimes, 80 randomly selected batteries were tested. The following frequency distribution was obtained. The data values are in hours given in table 1.11

Construct histograms for frequency, relative frequency and cumulative relative frequency.

Table 1.11

Class boundaries	Frequency
63.5-74.5	10
74.5-85.5	15
85.5-96.5	22
96.5-107.5	17
107.5-118.5	11
118.5-129.5	5

1.8. Mean for grouped data

The statistical measures we have presented for the central location and dispersion of data sets are computed using the individual data values. The computational procedures we have discussed provide the most common methods for computing measures of central location and dispersion. However, in some situations the data available only in grouped or frequency distribution form. In these cases special procedures are used in order to obtain approximations to the common measures of central location and dispersion.

The formulas used to calculate the mean for grouped data are as follows:

Mean for the population data:

$$\mu = \frac{\sum_{i=1}^k m_i \cdot f_i}{N}$$

Mean for the sample data:

$$\bar{x} = \frac{\sum_{i=1}^k m_i \cdot f_i}{n}$$

Where m_i – is the midpoint of i^{th} class, f_i – is the frequency of i^{th} – class, k – is the total number of classes.

To calculate the mean for grouped data, first find the midpoint of each class and then multiply by the frequencies of the corresponding classes. The sum of these products, denoted by $\sum_{i=1}^k f_i \cdot m_i$, gives an approximation for the sum of all values. To find the value of the mean, divide this sum by the total number of observations in the data.

Example:

The following table gives the frequency distribution of daily commuting time (in minutes) from home to work for all 25 employees of a company

Daily commuting time (minutes)	Number of employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean of daily commuting time.

Solution:

Note that because the data set includes all 25 employees of the company, it represents the population. Table 1.12 shows the calculation of $\sum_{i=1}^5 m_i \cdot f_i$.

In table 1.12- m_i denotes the midpoint of the classes.

Table 1.12

Daily commuting time (minutes)	f_i	m_i	$m_i \cdot f_i$
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
	$N = 25$		$\sum_{i=1}^5 m_i \cdot f_i = 535$

To calculate the mean, we first find the midpoint of each class. The class midpoints are recorded in the third column of Table 1.12. The products of the midpoints and the corresponding frequencies are listed in the fourth column of that table. The sum of column, denoted by $\sum m \cdot f$, gives the approximate total daily commuting time(in minutes) for all 25 employees.

The mean is obtained by dividing this sum by the total frequency. Therefore

$$\mu = \frac{\sum_{i=1}^5 m_i f_i}{N} = \frac{535}{25} = 21.40 \text{ minutes}$$

Thus, the employees of this company spend an average of 21.40 minutes a day commuting from home to work.

1.9. The Median for grouped data

Recall that the median is different for odd and for even numbers of observations when the data are not in the grouped form. However, if the n data are written in grouped form, then median is simply defined as the $(n/2)^{th}$ observation.

Thus, if we have the frequency distribution of 100 observations, then the 50^{th} observation in order of size would be the median; if we have 101 observations then the " 50.5^{th} " observation would be the median.

To find median, first, we need to find the class which contains the middle observation. Let M denotes the number of this class, where M is the some integers from 1 to k . If the median occurs in the fifth class then $M=5$; if it occurs in the seventh class, then $M=7$; and so on.

Let the frequency of the M^{th} class be denoted by f_M . Next, note how many observations are in $(M - 1)$ classes preceding the median class; denote this cumulative frequency by F_{M-1} .

The general formula for median is

$$\text{Median} = L_M + \frac{\frac{n}{2} - F_{M-1}}{f_M} \cdot C$$

where

- L_M – lower boundary of the median class
 n – number of observations
 f_M – the number of observations in the median class
 F_{M-1} – the number of observations in the $(M - 1)$ classes preceding the median class
 C – width of the median class

Example: Find the median of the frequency distribution

Starting monthly salary(in dollars)	Frequency
900-1000	2
1000-1100	4
1100-1200	3
1200-1300	1
1300-1400	1
1400-1500	0
1500-1600	1
	$n=12$

Solution:

First of all, let us divide n (the number of all observations) to find the halfway point.

$$\text{Median} = \left[\frac{n}{2} \right]^{\text{th}} \text{ observation} = \left[\frac{12}{2} \right]^{\text{th}} = 6^{\text{th}} \text{ observation}$$

To find the class that contains 6^{th} observation it is necessary to form cumulative frequency distribution. This class is called the median class; it contains the median:

Starting monthly salary(in dollars)	Frequency	Cumulative frequency
900-1000	2	2
1000-1100	4	6
1100-1200	3	9
1200-1300	1	10
1300-1400	1	11
1400-1500	0	11
1500-1600	1	12

6^{th} observation is in 2^{nd} class. So, the median class is 1000-1100.
Now let us apply

$$\text{Median} = L_M + \frac{\frac{n}{2} - F_{M-1}}{f_M} \cdot C$$

In our case

$$L_M = 1000; \quad n = 12; \quad F_{M-1} = 2; \quad f_M = 4; \quad C = 100$$

After substituting we get

$$\text{Median} = 1000 + \frac{6 - 2}{4} \cdot 100 = 1100$$

The median is 1100. In other words, median as a measure of center indicates that average value of monthly salaries of 12 employees is 1100\$.

1.10. Modal class

The mode for grouped data is the modal class. The modal class is the class with the largest frequency.

Example:

Find the mode of the frequency distribution

Solution:

The modal class is 20-25, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence the mode could be given as 22.5.

Class	Frequency
5-10	1
10-15	2
15-20	3
20-25	7
25-30	4
30-35	3
	$n=20$

Exercises

1. Dinner check amounts at the restaurant have the following frequency distribution:
Compute the mean, median, and mode for the above data.

Dinner check (dollars)	Frequency
4-8	4
8-12	5
12-16	7
16-20	2
20-24	1
24-28	1

2. The following table gives the frequency distribution of entertainment expenditures (in dollars) incurred by 50 families during the past week. Find the mean, median and mode.

Entertainment expenditure (dollars)	Number of families
0-10	5
10-20	10
20-30	15
30-40	12
40-50	5
50-60	3

3. The following table gives the frequency distribution of total hours studying during the semester for sample of 40 university students enrolled in an introductory business statistics course .
Find the mean, median, and mode

Hours of study	Number of students
24-40	3
40-56	5
56-72	10
72-88	12
88-104	5
104-120	5

4. This frequency distribution represents the data obtained from sample of 75 copying machine service technicians. The values represent the days between service calls for various copying machines.

Find the mean, median, and mode.

Class boundaries	Frequency
15.5-18.5	14
18.5-21.5	12
21.5-24.5	18
24.5-27.5	10
27.5-30.5	15
30.5-33.5	6

5. For 35 antique car owners, the following distribution of cars' age was obtained. Find the mean, median, and mode.

Class limits	Frequency
13-19	2
20-26	7
27-33	12
34-40	5
41-47	6
48-54	1
55-61	0
62-68	2
	$n=35$

Answers

1. 12.8; 12.57; 14; **2.** 27.20; 26.7; 25; **3.** 74.40; 74.67; 80; **4.** 23.72; 23.417; 23; **5.** 33.8; 31.5; modal class =27-33.

1.11. Variance and standard deviation for grouped data

Suppose that we have data grouped into K classes, with frequencies f_1, f_2, \dots, f_k . If the midpoints of these classes are m_1, m_2, \dots, m_k , the mean and variance of the grouped data are estimated by using following formulas

1. For a population of N observations, so that

$$N = \sum_{i=1}^k f_i$$

The variance is

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (m_i - \mu)^2}{N} = \frac{\sum_{i=1}^k f_i \cdot m_i^2}{N} - \mu^2$$

The standard deviation is $\sigma = \sqrt{\sigma^2}$.

2. For a sample of n observations, so that

$$n = \sum_{i=1}^k f_i$$

The variance is

$$s^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \left(\bar{x} \right)^2 \right]$$

The standard deviation is $s = \sqrt{s^2}$.

Example:

The following table gives the distribution of the number of days for which all 40 employees of a company were absent during the last year

Number of days absent	Number of employees
0-2	13
3-5	14
6-8	6
9-11	4
12-14	3

Calculate the variance and standard deviation.

Solution:

Let us apply

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

First we need to find m_i and μ

Class i	Number of days absent	Number of employees (f_i)	Class mark (m_i)	$m_i \cdot f_i$
1	0-2	13	1	13
2	3-5	14	4	56
3	6-8	6	7	42
4	9-11	4	10	40
5	12-14	3	13	39
		40		$\sum_{i=1}^5 m_i f_i = 190$

$$\mu = \frac{\sum_{i=1}^5 m_i \cdot f_i}{N} = \frac{190}{40} = 4.75$$

Now we need to find $\sum_{i=1}^k f_i \cdot (m_i - \mu)^2$.

In order to find products $f_i \cdot (m_i - \mu)^2$ we must first find the square quantities $(m_i - \mu)^2$. We need three columns to display the computation of

the quantities $(m_i - \mu)^2$ – a column for m_i , a column for $(m_i - \mu)$, and a column for the $(m_i - \mu)^2$. We also need a column for f_i and a final column for the products $f_i \cdot (m_i - \mu)^2$. The necessary computations are shown below in the table 1.13.

Table 1.13.

Class <i>i</i>	Class mark (m_i)	f_i	$(m_i - \mu)$	$(m_i - \mu)^2$	$f_i(m_i - \mu)^2$
1	1	13	1-4.75=-3.75	14.06	182.81
2	4	14	4-4.75=-0.75	0.56	7.88
3	7	6	7-4.75=2.25	5.06	30.38
4	10	4	10-4.75=5.25	27.56	110.25
5	13	3	13-4.75=8.25	68.06	204.19
					535.51

In the end,

$$\sigma^2 = \frac{535.51}{40} = 13.39 \text{ and } \sigma = \sqrt{\sigma^2} = 3.66$$

Example:

The following table gives the frequency distribution of the number of orders received each day during the past 50 days at office of a mail-order company. Calculate variance and standard deviation.

Solution:

Because the data includes only 50 days, it represents a sample. Hence, we will use sample formulas to calculate the variance and standard deviation. Let us apply

Number of orders	f_i	the
10-12	4	
13-15	12	
16-18	20	
19-21	14	

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \left(\bar{x} \right)^2 \right]$$

All the information required for the calculation of the variance and standard deviation appears in the following table

Number of orders	f_i	m_i	$m_i f_i$	m_i^2	$f_i \cdot m_i^2$
10-12	4	11	44	121	484
13-15	12	14	168	196	2352
16-18	20	17	340	289	5780
19-21	14	20	280	400	5600
	$N=50$		832		14216

$$\bar{x} = \frac{832}{50} = 16.64$$

By substituting the values in the formula for the sample variance, we obtain

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^k f_i \cdot m_i^2 - n \cdot \bar{x}^2 \right] = \frac{1}{50-1} [14216 - 50 \cdot (16.64)^2] = 7.582$$

Hence, the standard deviation is $s = \sqrt{s^2} = \sqrt{7.580} = 2.75$.

Thus, the standard deviation of the number of orders received at office of this mail-order company during the past 50 days is 2.75.

1.12. Interquartile range for grouped data

Suppose that a class, with lower boundary L and upper boundary U , contains f observations. If these observations were to be arranged in ascending order, the j^{th} observation is estimated by

$$L + (j - \frac{1}{2}) \cdot \frac{(U - L)}{f} \quad \text{for } j = 1, 2, 3, \dots, f.$$

where

L – is the lower limit of class containing j^{th} observation

U – is the upper limit of class containing j^{th} observation

f – is the frequency of class containing j^{th} observation

j – is the location of j^{th} observation in that class.

For interquartile range we need to find

$$Q_1 = \left[\frac{N+1}{4} \right]^{th} \text{ and } Q_3 = \left[\frac{3 \cdot (N+1)}{4} \right]^{th}$$

As we know I.Q.R. = Interquartile range = $Q_3 - Q_1$.

Example: The following table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company
Calculate the interquartile range.

Number of orders	f_i
10-12	4
13-15	12
16-18	20
19-21	14

Solution:

First of all, let us write cumulative frequency distribution

Number of orders	f_i	Cumulative frequency
10-12	4	4
13-15	12	16
16-18	20	36
19-21	14	50

Since there are $N=50$ observations, we have

$$Q_1 = \left[\frac{N+1}{4} \right]^{th} = \left[\frac{51}{4} \right]^{th} = \left[12 \frac{3}{4} \right]^{th} = 12^{th} + \frac{3}{4}(13^{th} - 12^{th})$$

Hence the first quartile is the three-quarters way from the 12^{th} observation to 13^{th} . From cumulative distribution we see that the 12^{th} value is the 8^{th} value in the class 13-15. In our notation then

$$j = 8; \quad f = 12; \quad L = 13; \quad U = 15$$

The 12th observation is estimated by

$$L + \left(j - \frac{1}{2}\right) \cdot \frac{(U - L)}{f} = 13 + \left(8 - \frac{1}{2}\right) \cdot \frac{15 - 13}{12} = 14 \frac{1}{4} = 14.25$$

Similarly, the 13th observation is the 9th value in the same class, so now, with $j = 9$, we have

$$L + \left(j - \frac{1}{2}\right) \cdot \frac{(U - L)}{f} = 13 + \left(9 - \frac{1}{2}\right) \cdot \frac{15 - 13}{12} = 14 \frac{5}{12} = 14.41$$

Since the first quartile is three-quarters of the way from the twelves observation to the thirteens observation, we have

$$\begin{aligned} \text{First quartile } Q_1 &= 12^{\text{th}} + \frac{3}{4}(13^{\text{th}} - 12^{\text{th}}) = \\ &= 14 \frac{1}{4} + \frac{3}{4}(14 \frac{5}{12} - 14 \frac{1}{4}) = 14.375. \end{aligned}$$

To find third quartile, we have

$$Q_3 = \frac{3 \cdot (N+1)}{4} = \left[\frac{150}{4} \right]^{\text{th}} = \left[37 \frac{1}{2} \right]^{\text{th}} = 37^{\text{th}} + \frac{1}{2}(38^{\text{th}} - 37^{\text{th}})$$

Therefore, when the observations are arranged in ascending order, the third quartile is half of the way from thirty-seventh to thirty-eighth.

Looking at table, we see that the thirty-seventh observation is the first value in class the 19-21, which contains 14 observations. We have then

$$j = 1; \quad f = 14; \quad L = 19; \quad U = 21$$

Thus, the thirty-seventh observation us estimated by

$$L + \left(j - \frac{1}{2}\right) \cdot \frac{(U - L)}{f} = 19 + \left(1 - \frac{1}{2}\right) \cdot \frac{21 - 19}{14} = 19 \frac{1}{14} = 19.07$$

Similarly, the thirty-eighth observations the second value in the same class, so with $j = 2$, we estimate 38th observation by

$$L + \left(j - \frac{1}{2}\right) \cdot \frac{(U - L)}{f} = 19 + \left(2 - \frac{1}{2}\right) \cdot \frac{21 - 19}{14} = 19 \frac{3}{14} = 19.21$$

Hence, since the third quartile is half of the way from the 37th to 38th, we have

$$\text{Third quartile} = Q_3 = 37^{\text{th}} + \frac{1}{2}(38^{\text{th}} - 37^{\text{th}}) = \\ = 19\frac{1}{14} + \frac{1}{2}(19\frac{3}{14} - 19\frac{1}{14}) = 19\frac{2}{14} = 19.14$$

Finally, then the interquartile range is the difference between the third and first quartiles, so

$$\text{Interquartile range} = 19\frac{2}{14} - 14\frac{1}{32} = 19.07 - 14.375 = 4.695$$

Thus, if the interquartile range is to be used as a measure of dispersion, we estimate it by 4.695.

Exercises

- 1.** For 50 airplanes that arrived late at an airport during a week, the time by which they were late was observed. In the following table, x denotes the time (in minutes) by which an airplane was late and f denotes the number of airplanes.

x	f
0 to less than 20	14
20 to less than 40	18
40 to less than 60	9
60 to less than 80	5
80 to less than 100	4

- a) Find the mean
- b) Find the median
- c) Find the mode
- d) Find the variance and standard deviation
- e) Find the interquartile range.

- 2.** The following table gives information on the amount (in dollars) of the electric bills for a sample of 40 families.

Amount of electric bill (dollars)	Number of families
4 to less than 8	2
8 to less than 12	9
12 to less than 16	16
16 to less than 20	8
20 to less than 24	5

- a) Estimate the sample mean

- b) Estimate the median
- c) Estimate the mode
- d) Estimate the variance and standard deviation.
- e) Estimate the interquartile range.

3. A population of all twenty financial analysts was asked to provide forecasts of earnings per share of a corporation for next year. The results are summarized in the table.

Forecast \$ per share	Number of analysts
0.5-10.5	2
10.5-20.5	4
20.5-30.5	9
30.5-40.5	5

- a) Find the relative frequencies.
- b) Find the cumulative frequencies.
- c) Find the cumulative relative frequencies.
- d) Estimate the population mean.
- e) Estimate the population variance.
- f) Estimate the population standard deviation.
- g) Estimate the population mode.
- h) Estimate the population median.
- i) Estimate the interquartile range.
- j) Which class is modal class?

4. A sample was taken of flights arriving at a major airport to study the problem of air traffic delays. The table shows numbers of minutes late for a sample of 100 flights.

Minutes late	0-10	10-20	20-30	30-40	40-50	50-60
Number of flights	29	23	17	14	11	6

- a) Draw the histogram
- b) Find the sample relative frequencies
- c) Find and interpret the sample cumulative relative frequencies
- d) Estimate the sample mean number of minutes
- e) Estimate the sample variance and standard deviation
- f) Estimate the sample median number of minutes late

- g) Estimate the interquartile range
 h) Which is the modal class for this sample?

5. The following table gives the frequency distribution of the number of computers sold during the past 25 weeks at a computer store.

Computers sold	Frequency
4 to 9	2
10 to 15	5
16 to 21	10
22 to 27	5
28 to 33	3

Calculate the mean, variance, and standard deviation.

6. Eighty randomly selected light bulbs were tested to determine their lifetimes (in hours). The following frequency distribution was obtained

Find the variance and standard deviation.

Class limits	Frequency
52.5-63.5	6
63.5-74.5	12
74.5-85.5	25
85.5-96.5	18
96.5-107.5	14
107.5-118.5	5

7. The following data represent the scores (in words per minute) of 25 typists on a speed test.

Find the variance and standard deviation.

Class limits	Frequency
54-58	2
59-63	5
64-68	8
69-73	0
74-78	4
79-83	5
84-88	1

8. For a sample of fifty new full-size cars, fuel consumption figures were obtained and summarized in the accompanying table

Fuel consumption	14-16	16-18	18-20	20-22	22-24
Number of cars	3	6	13	20	8

- a) Draw the histogram.
- b) Find the sample relative frequencies.
- c) Find and interpret the sample cumulative relative frequencies
- d) Estimate the sample mean fuel consumption.
- e) Estimate the sample standard deviation of fuel consumption.
- f) Estimate the sample median fuel consumption.
- g) Estimate the sample interquartile range.
- h) Which is the modal class for this sample?

9. The fuel capacity in gallons of 30 randomly selected cars is shown below.

Find

- a) Mean
- b) Median
- c) Modal class
- d) Variance
- e) Standard deviation

Class	Frequency
12.5-27.5	6
27.5-42.5	3
42.5-57.5	5
57.5-72.5	8
72.5-87.5	6
87.5-102.5	2

10. Twelve batteries were tested after being used for one hour. The output (in volts) is shown below.

Find each of the following

- a) Mean
- b) Median
- c) Mode
- d) Range
- e) Variance
- f) Standard deviation.

Volts	Frequency
2	1
3	4
4	5
5	1
6	1

11. For a sample of twenty-five students from a large class, the accompanying table shows the amount of time students spent studying for a test

Study time (hours)	0-2	2-4	4-6	6-8	8-10
Number of students	3	4	8	7	3

- a) Draw the histogram.
- b) Find and interpret the cumulative relative frequencies.
- c) Estimate the sample mean study time.
- d) Estimate the sample median.
- e) Find the modal class.
- f) Estimate the sample variance.
- g) Estimate the sample standard deviation study time.
- h) Estimate the sample interquartile range.

Answers

- 1.** a) $\bar{x} = 36.80$; b) 32.22; c) 30; d) $s^2 = 597.714$; $s = 24.45$;
 e) $I.R. = 35.28$; **2.** a) $\bar{x} = 14.50$; b) 14.25; c) 14; d) $s^2 = 18.205$;
 $s = 4.27$; e) $I.R. = 6.556$; **3.** a) 2/20; 4/20; 9/20; 5/20; b) 2; 6; 15; 20;
 c) 2/20; 6/20; 15/20; 20/20; d) $\mu = 480/20 = 24$; $\sigma^2 = 82.75$;
 e) $\sigma = 9.097$; f) median=24.944; i) I.Q.R. =13.735; j) modal class:
 20.5-30.5; **4.** b) 0.29; 0.23; 0.17; 0.14; 0.11; 0.06; c) 0.29; 0.52; 0.69; 0.83;
 0.94; 1.0; d) $\bar{x} = 2230/100 = 22.3$; e) $s^2 = 246.176$; $s = 15.69$; f)
 median=19.13; g) 25.931; h) modal class 0-10; **5.** $\bar{x} = 18.98$;
 $s^2 = 44.760$; $s = 6.69$; **6.** $s^2 = 211.2$; $s = 14.5$; **7.** $s^2 = 80.3$; $s = 9.0$;
8. b) 3/50; 6/50; 13/50; 20/50; 8/50; c) 3/50; 9/50; 22/50; 42/50; 50/50; d)
 $\bar{x} = 19.96$; e) $s = 2.185$; f) median=20.3; g) I.Q.R.= 3.075; h) modal class:
 20-22; **9.** a) $\bar{x} = 55.5$; b) median=59.4; c) modal class;
 57.5-72.5; d) $s^2 = 566.1$; e) $s = 23.8$; **10.** a) $\bar{x} = 3.8$; b) median=4;

c) mode = 4; d) range = 4; e) $s^2 = 1.11$; f) $s = 1.05$; **11.** b) $3/25$; $7/25$; $15/25$; $22/25$; $25/25$; c) 5.24; d) 5.375; e) modal class = 4-6; f) 5.773; g) 2.403; h) 3.64.

Chapter 2. Probability

2.1. Introduction

The idea of probability is a familiar one to everyone. Statements such as the following are heard frequently: “You have better to take an umbrella because it is likely to rain”, “It is not likely to snow today”, “He will probably read at least three books during the next two weeks”, “ I am almost certain that we will go home for the holidays”.

These examples illustrate that most of us use the concept of probability in our everyday speech. They also illustrate that there is a great deal of imprecision involved in such statement. For instance, a family has been arguing about whether they will go to the wife’s home for holidays, and have not reached a decision. Even though the wife may say, “It is almost certain that we will go home for the holidays”, the husband might not even think that such an occurrence is likely. This sort of imprecision is intolerable in mathematics. These difficulties can be avoided if we restrict our discussion of probability to events which are outcomes of experiments that can be repeated, and if we deal with idealized situations.

One can think of probability as the language in which we discuss uncertainty. Before we can communicate with one another in this language, we need to acquire a common vocabulary. Moreover, as in any other language, rules of grammar are needed so that a clear statement can be made with our vocabulary.

2.2. Random experiment, outcomes, and sample space

Definition:

A random experiment is a process that, when performed, results in one and only one of at least two observations.

Definition:

The possible outcomes of a random experiment are called the **basic outcomes**.

Definition:

The set of all basic outcomes is called the **sample space**.

A sample space will be denoted by **S**.

Table 2.1 illustrates list of a few examples of random experiments, their outcomes, and their sample spaces.

Table 2.1

Experiment	Outcomes	Sample space
Toss a coin once	Head, Tail	$S=\{\text{Head, Tail}\}$
Roll a die once	1,2,3,4,5,6	$S=\{1,2,3,4,5,6\}$
Toss a coin twice	HH,HT,TH,TT	$S=\{\text{HH,HT,TH,TT}\}$
Take a test	Pass, Fail	$S=\{\text{Pass, Fail}\}$
Select a student	Male, Female	$S=\{\text{Male, Female}\}$

The sample space for an experiment can also be described by drawing a

Venn diagram. A Venn diagram is a picture (a closed geometric shape such as a rectangle, a square, or circle) that depicts all the possible outcomes for an experiment.

Example:

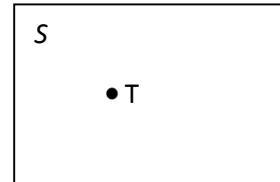
Draw the Venn diagram for the experiment of tossing a coin.

Solution:

This experiment has two possible outcomes: head and tail. Sample space is given by $S = \{H, T\}$;

where H =head, T =Tail.

To draw Venn diagram for this example, we draw a rectangle and mark two points inside this rectangle that represent two outcomes, head and tail.



The rectangle is labelled by S because it represents the sample space. (Fig. 2.1)

Example:

Suppose we randomly select two employees from a company and observe whether the employee selected each time is a male or female. Write all the outcomes for this experiment and draw the Venn diagram.

Solution:

Let us denote selection of male by M and that of female by F . There are four final possible outcomes: MM , MF , FM , and FF .

We can write sample space as

$$S = \{ MM, MF, FM, FF \}$$

The Venn diagram is given in Fig.2.2

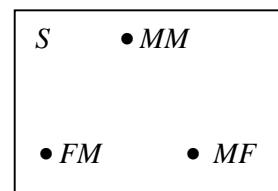


Fig.2.2.

Definition:

An event is a collection of one or more of the outcomes from the sample space.

Usually, an event is denoted by E_1, E_2, E_3 , and so forth. We can denote it by any other letters too, that is, by A, B, C and so forth.

Definition:

An event that includes one and only one of the outcomes for an experiment is called a **simple event**.

Definition:

An event is called a **compound event** if it contains more than one outcome for an experiment.

Definition:

Let A and B be two events defined in sample space S . The **intersection** of A and B represents the collection of all outcomes that are common to both A and B and is denoted by either $A \cap B$ or AB .

Hence, $A \cap B$ or AB occurs if and only if both A and B occur.

More generally, given N events E_1, E_2, \dots, E_N , their intersection, $E_1 \cap E_2 \cap \dots \cap E_N$, is the set of all basic outcomes that belong to every E_i , where $i = 1, 2, \dots, N$.

Example:

Let A =the event that a family owns washing machine

B = the event that family owns a VCR. Then intersection of these events includes all the families who own both washing machine and VCR.

Definition:

Let A and B be two events defined in sample space S . Their **union**, denoted by $A \cup B$ is the set of all basic outcomes in S that belong to at least one of these two events. The union $A \cup B$ occurs if and only if either A or B (or both) occurs.

More generally, given N events E_1, E_2, \dots, E_N , their union,

$E_1 \cup E_2 \cup \dots \cup E_N$, is the set of all basic outcomes that belong to at least one of these N events.

Definition:

Two events A and B are called **mutually exclusive** events if they have no common basic outcomes and their intersection $A \cap B$ is said to be empty set.

Definition:

Let E_1, E_2, \dots, E_N be N events in the sample space S .

If $E_1 \cup E_2 \cup \dots \cup E_N = S$, then these N events are said to be **collectively exhaustive** events.

Definition:

Let A be an event in the sample space. The **complement of event A** , denoted by \bar{A} and read as “ A bar” or “ A complement” is the event that includes all the outcomes for an experiment that are not in A .

Example:

A statistical experiment has eight equally likely outcomes that are denoted by 1, 2, 3, 4, 5, 6, 7, and 8. Let $A = \{2, 5, 7\}$ and $B = \{2, 4, 8\}$

a) Find $A \cap B$

- b) Find $A \cup B$
- c) Are events A and B mutually exclusive events?
- d) Are events A and B collectively exhaustive events?
- e) Find \bar{A} and \bar{B}

Solution:

- a) These two events have only one common element $A \cap B = \{2\}$
- b) $A \cup B = \{2, 4, 5, 7, 8\}$
- c) Events A and B are not mutually exclusive events, because they have common element, and their intersection is not empty set.
- d) Events A and B are not collectively exhaustive events, because their union does not equal to sample space.
- e) $\bar{A} = \{1, 3, 4, 6, 8\}; \bar{B} = \{1, 3, 5, 6, 7\}$

Example:

At a busy international airport arriving planes land a first come, first served basis. Let A , B , and C be the events that there are at least five, at most three, and exactly two planes waiting to land, respectively. Then

1. \bar{A} is the event that at most four planes are waiting to land.
2. \bar{B} is the event that at least four planes are waiting to land.
3. A is a subset of \bar{B} ; that is if A occurs, then \bar{B} occurs.

Therefore, $A \cap \bar{B} = A$.

4. C is a subset of B ; that is if C occurs, then B occurs.

Therefore, $B \cap C = C$.

5. A and B are mutually exclusive; that is, $A \cap B = \emptyset$.

A and C also mutually exclusive since $A \cap C = \emptyset$.

6. $B \cap \bar{C}$ is the event that number of planes waiting to land is zero, one or three.

2.3. Three conceptual approaches to probability

There are three conceptual approaches to probability:

1. Classical probability
2. Relative frequency concept of probability
3. Subjective probability.

2.3.1. Classical probability

Classical probability assumes that all outcomes in the sample space are equally likely to occur. It was developed originally in the analysis of gambling problems, where the assumption of equally likely outcomes often is reasonable. When the assumption of equally likely outcomes is used a basis for assigning probabilities, the approach is referred to as the classical method.

According to the classical method, the probability of a single event is equal to one divided by the total number of outcomes for experiment. On the other hand, the probability of a compound event A is equal to number of outcomes favourable to event A , divided by the total number of outcomes for the experiment.

$$P(A) = \frac{\text{Number of outcomes favorable to } A}{\text{Total number of outcomes for the experiment}}$$

2.3.2. Relative frequency concept of probability

The difference between classical probability and relative frequency probability is that classical method assumes that certain outcomes are equally (such as the outcomes when a die is rolled) while relative frequency method relies on actual experience to determine the likelihood of outcomes. In relative frequency method, one might actually roll a given die 1000 times and observe the relative frequencies and use these frequencies to determine the probability of an outcome. This method of assigning a probability to an event is called the **relative frequency concept of probability**.

Definition:

If an experiment is repeated n times and an event A is observed f times, then, according to the relative frequency concept of probability:

$$P(A) = \frac{f}{n}$$

Because relative frequencies are determined by performing an experiment, the probabilities calculated using relative frequencies may change almost each time an experiment is repeated. But the variation in probabilities will be small if the sample size is large.

2.3.3. Subjective probability

Subjective probability uses a probability value based on an educated guess or estimate, employing opinions.

In subjective probability, a person or group makes an educated guess at the chance that an event will occur. This guess is based on person's experience and evaluation of solution. For example, a sportswriter may say that there is a 65% probability that the Milan will win championship cup next year. A doctor might say that on the basis of his diagnosis, there is a 40% chance the patient will need an operation. A seismologist might say there is a 60% probability that an earthquake will occur in a certain area.

Definition:

Subjective probability is the probability assigned to an event based on subjective judgment, experience, information, and belief.

All three types of probability (classical, relative frequency, and subjective) are used to solve variety of problems in business, economics, engineering, and other fields.

2.4. Probability and its postulates

Probability, which gives the likelihood of occurrence of an event, is denoted by P .

Let S denote the sample space of a random experiment, E_i the basic outcomes, and A an event. The probability that an event A will occur is denoted by $P(A)$.

The probability has the following important properties:

1. If A is any event in the sample space S , then $0 \leq P(A) \leq 1$.

An event that can not occur has zero probability; such event is called **an impossible** event. An event that is certain to occur has a probability equal to 1 and is called **sure** event.

For impossible event M : $P(M) = 0$

For a sure event C : $P(C) = 1$

2. Let A be an event in sample space S , and let E_i denote the basic outcomes. Then

$$P(A) = \sum_A P(E_i),$$

where the notation implies that the summation extends over all the basic outcomes in A .

3. The sum of the probabilities for all the basic outcomes in the sample space always is 1. Thus

$$P(S) = \sum_{i=1}^n P(E_i) = P(E_1) + P(E_2) + \dots + P(E_n) = 1.$$

2.5. Formula for classical probability

Classical probability uses sample spaces to determine the numerical probability that an event will happen.

Classical probability assumes that all outcomes in the sample space are equally likely to occur. For example, when a single die is rolled, each outcome has the same probability of occurring. Since there are six outcomes, each outcome has a probability of $1/6$. When a card is selected from an ordinary deck of 52 cards, we assume that the deck has been shuffled, and each card has the same probability of being selected. In this case, it is $1/52$.

Definition:

If the sample space S contains n equally likely basic outcomes and the event A consists of m of these outcomes ($m \leq n$) , then

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

In words, “The probability of event A equals to number of basic outcomes in A , divided by the total number of outcomes in the sample space”.

We can write definition above as $P(A) = \frac{n(A)}{n(S)}$.

Example:

For a card drawn from an ordinary deck find the probability of getting a queen

Solution:

Let A - be an event getting a queen. Since there are four queens then

$$n(A) = 4. \text{ Hence, } P(A) = \frac{4}{52} = \frac{1}{13}.$$

Example:

Find the probability of obtaining an even number in one roll of a die.

Solution:

In this experiment $S= \{1, 2, 3, 4, 5, 6\}$. Let A - be an event that an even number is observed on the die. Event A has three outcomes: 2, 4, and 6.

If any one of these three numbers is obtained, event A is said to occur. Hence,

$$P(A) = \frac{\text{number of outcomes included in } A}{\text{total number of outcomes}} = \frac{3}{6} = 0.5$$

2.6. Consequences of the postulates

1. Let A and B be mutually exclusive events. Then the probability of their union is the sum of their individual probabilities;

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

More generally, if E_1, E_2, \dots, E_N are mutually exclusive events, then

$$P(E_1 \cup E_2 \cup \dots \cup E_N) = P(E_1) + P(E_2) + \dots + P(E_n)$$

2. If E_1, E_2, \dots, E_N are collectively exhaustive events, then the probability of their union is $P(E_1 \cup E_2 \cup \dots \cup E_N) = 1$.

Since the events are collectively exhaustive, their union is the whole sample space S and $P(S) = 1$.

Example:

A drawer contains three pairs of red socks, two pairs of black socks and four pairs of brown socks. If a person in a dark room selects a pair of socks, find probability that the pair will be either black or brown. (Note: The socks are folded together in matching pairs).

Solution:

Let us define the following events

A= the selected socks are black

B= the selected socks are brown.

Since there are nine pairs of socks,

$$P(\text{black}) = P(A) = \frac{2}{9}; \quad P(\text{brown}) = P(B) = \frac{4}{9}$$

$$P(\text{black or brown}) = P(A \cup B) = P(A) + P(B) = \frac{2}{9} + \frac{4}{9} = \frac{2}{3}.$$

Example:

A day of the week is selected at random. Find the probability that it is a weekend day.

Solution:

Let

A = the selected day is Saturday

B = the selected day is Sunday

$$P(A)=\frac{1}{7}; \quad P(B)=\frac{1}{7} \text{ and } P(A \cup B) = P(A) + P(B) = \frac{2}{7}$$

Exercises

1. A box contains three red and five blue balls. Define the sample space for experiment of recording the colours of three balls that are drawn from the box one by one, with replacement.

2. Define a sample space for the experiment of putting three different books on a shelf in random order. If two of these three books are a two-volume dictionary, describe the event that these volumes stand in increasing order side by side (i.e., volume I precedes volume II).

3. A simple card is drawn from an ordinary pack of playing cards. What is the probability that the card is

- a) An ace
- b) A five
- c) A red card
- d) A club

4. There are 15 slips of paper in a hat, numbered from 1 to 15. If one of slip is drawn at random, find the probability that

- a) The number drawn is 5
- b) The number drawn is even

- c) The number drawn is odd
- d) The number drawn is divisible by 3

5. Two events, A and B , are mutually exclusive: $P(A) = \frac{1}{5}$ and $P(B) = \frac{1}{3}$.

Find the probability that

- a) Either A or B will occur
- b) Both A and B will occur
- c) Neither A nor B will occur

6. The manager of a furniture store sells from zero to four sofas each week. Based on past experience, the following probabilities are assigned to sales of zero, one, two, three, or four sofas:

$$P(0) = 0.08$$

$$P(1) = 0.18$$

$$P(2) = 0.32$$

$$P(3) = 0.30$$

$$\underline{P(4) = 0.12}$$

$$1.00$$

- a) Are these valid probability assignments? Why or why not?
- b) Let A be the event that two or fewer are sold in one week. Find $P(A)$.
- c) Let B be the event that four or more are sold in one week. Find $P(B)$.
- d) Are A and B mutually exclusive? Find $P(A \cap B)$ and $P(A \cup B)$.

7. Bektur, Janat, and Linar are the finalists in the spelling contest of a local school. The winner and the first runner-up will be sent to a city-wide competition.

- List the sample space of concerning the out comes of the local contest.
- Give the composition of each of the following events

A =Linar wins the local contest

B = Bektur does not go to the city-wide contest.

8. In a large department store, there are two managers, four department heads, 16 clerks, and four stokers. If a person selected at random, find the probability that the person is either a clerk or a manager.

9. On a small college campus, there are five English professors, four mathematics professors, two science professors, three psychology professors, and three history professors. If a professor is selected at random, find the probability that the professor is the following

- An English or psychology professor.
- A mathematics or science professor.
- A history, science, or mathematics professor.
- An English, mathematics, or history professor.

10. A hospital has monitored the length of time a patient spends in a hospital. The probability for numbers of days a patient spends in the hospital, are shown in following table:

Number of days	0	1-3	4-6	7-9	10-12	More than 12
Probability	0.14	0.39	0.23	0.15	0.06	0.03

Let A be the event “The patient spends at least one days in the hospital”, and B be the event “The patient spends less than 10 days in the hospital”.

- a) Find the probability of event A
- b) Find the probability of event B .
- c) Find the probability of the complement of A .
- d) Find the probability of the union of A and B .
- e) Find the probability of the intersection of A and B .
- f) Are A and B mutually exclusive events?
- g) Are A and B collectively exhaustive events?

Answers

1. { $RRR, RRB, RBR, RBB, BRR, BRB, BBR, BBB$ }; 2. { $d_1d_2a, d_1ad_2,$
 $, d_2d_1a, d_2ad_1, ad_1d_2, ad_2d_1$ }; { d_1d_2a, ad_1d_2 }; 3. a) $\frac{1}{13}$; b) $\frac{1}{13}$; c) $\frac{1}{2}$;
d) $\frac{1}{4}$; 4. a) $\frac{1}{15}$; b) $\frac{7}{15}$; c) $\frac{8}{15}$; d) $\frac{1}{3}$; 5. a) $\frac{8}{15}$; b) 0; c) $\frac{7}{15}$; 6. a) Yes;
b) 0.58; c) 0.12; d) Yes; $P(A \cap B) = 0$; $P(A \cup B) = 0.70$; 7. a) { $BJ, BL,$
 JB, JL, LB, LJ }; b) $A = \{ LB, LJ \}$; $B = \{ JL, LJ \}$; 8. $\frac{9}{13}$; 9. a) $\frac{8}{17}$; b) $\frac{6}{17}$;
c) $\frac{9}{17}$; d) $\frac{12}{17}$; 10. a) 0.86; b) 0.91; c) 0.14; d) 1; e) 0.77; f) no, because
 $P(A \cap B) \neq \emptyset$; g) Yes, because $P(A \cup B) = 1$.

2.7. Counting principle. Permutation and combination

Counting principle:

If the set E contains n elements and the set F contains m elements, there are $n \times m$ ways in which we can choose first an element of E and then element of F .

Example:

We toss a coin two times. This experiment has two steps: the first step toss, the second toss. Each step has two outcomes: a head and a tail. Thus, total outcomes for two tosses of a coin = $2 \times 2 = 4$.

The four outcomes for this experiment are: HH, HT, TH, TT

Generalized counting principle:

Let $E_1, E_2, E_3, \dots, E_k$ be sets with $n_1, n_2, n_3, \dots, n_k$ elements, respectively. Then there are $n_1 \times n_2 \times n_3 \times \dots \times n_k$ ways in which we can first choose an element of E_1 , then an element of E_2, \dots , and finally an element of E_k .

Example:

How many outcomes does the experiment of throwing five dice have?

Solution:

Let $E_i, 1 \leq i \leq 5$ be set of all possible outcomes of i^{th} die. Then

$E_i = \{1, 2, 3, 4, 5, 6\}$. The number of the outcomes of the experiment of throwing five dice equals the number of ways we can first choose an element of E_1 , then an element of E_2, \dots , and finally an element of E_5 .

That is $6 \times 6 \times 6 \times 6 \times 6 = 6^5$

2.7.1. Permutation

Definition:

An n -element permutation of a set with n objects is simply called a **permutation**, denoted by P_n^n . The number of permutations of a set containing n elements is $P_n^n = n(n - 1)(n - 2) \dots \dots 2 \cdot 1 = n!$

Example:

The 3 digits 1, 2, 3 can be arranged in $3! = 6$ different orders:

123, 132, 213, 231, 312, 321

Definition:

The number of permutations, P_r^n of r objects chosen from n is the number of possible arrangements when r objects are to be selected from a total of n and arranged in order. This number is

$$P_r^n = \frac{n!}{(n - r)!}$$

Remark:

Instead of P_r^n , the symbols ${}_nP_r$ and $P(n, r)$ are frequently used to denote the number of permutations of n objects taken r at a time. Different authors frequently use different symbols.

Example:

Three students, Kanat, Askhat, and Marat must be scheduled for a job interviews. In how many different orders can this be done?

Solution:

The number of different orders is equal to the number of permutations of the set {Kanat, Askhat, Marat}. So there are $3! = 6$ possible orders for the interviews.

Example:

If 5 persons are to pose for a photograph by standing in a row, how many different arrangements are possible?

Solution:

Since we are arranging 5 “objects” 5 at a time, the number of different arrangements is given by $P_5^5 = 5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$.

Example:

If five boys and five girls sit in a row in a random order, what is the probability that no two children of the same sex sit together?

Solution:

There are $10!$ ways for 10 persons to sit in a row. In order that no two of the same sex sit together, boys must occupy positions 1, 3, 5, 7, 9 and girls positions 2, 4, 6, 8, 10, or vice versa. In each case there are $5! \times 5!$

possibilities. So, the desired possibility is $\frac{2 \cdot 5! \cdot 5!}{10!} \approx 0.008$.

Theorem: The number of distinguishable permutations of n objects of k different types, where n_1 are alike, n_2 are alike, ..., n_k are alike and $n = n_1 + n_2 + \dots + n_k$ is

$$\frac{n!}{n_1! \times n_2! \times \dots \times n_k!}.$$

Example:

How many different 10-letter codes can be made using three a 's, four b 's, and three c 's?

Solution:

By theorem, the number of such codes is $\frac{10!}{3!4!3!} = 4200$.

2.7.2. Combination

If order is of no importance, then we have a combination rather a permutation. A combination of n objects taken r at a time is a selection of r objects taken from the n , without regard to the order in which they are selected or arranged. Order is irrelevant.

Definition:

The number of combinations, C_r^n , of r objects chosen from n is the number of possible selections that can be made. This number is

$$C_r^n = \frac{n!}{r!(n-r)!}$$

Remark:

Some other symbols which are used to denote the number of combinations of n objects taken r at a time are ${}_n C_r$, $C(n, r)$ and $\binom{n}{r}$.

Example:

If a club has a membership of ten, then how many three-man committees are possible?

Solution:

Order is not important, so this is combination problem. The number of possible committees is equal to the number of ways three persons can be selected from ten persons, namely C_3^{10} .

We have:

$$C_3^{10} = \frac{10!}{3!7!} = 120.$$

Example:

In how many ways can two math and three biology books be selected from eight math and six biology books?

Solution:

There are C_2^8 possible ways to select two math books and C_3^6 possible ways to select three biology books. Therefore, by counting principle,

$$C_2^8 \cdot C_3^6 = \frac{8!}{2!6!} \cdot \frac{6!}{3!3!} = 560$$

is the total number of ways in which two math

and three biology books can be selected.

Example:

A box contains 24 transistors, four of which are defective. If four are sold at random, find the following probabilities:

- a) Exactly two are defective
- b) None is defective
- c) All are defective
- d) At least one is defective

Solution:

There are C_4^{24} ways to sell four transistors, so the denominator in each case will be $C_4^{24} = 10626$

- a) Two defective transistors can be selected as C_2^4 and two nondefective ones as C_2^{20} . Hence

$$P(2 \text{ defective}) = \frac{C_2^4 \cdot C_2^{20}}{C_4^{24}} = \frac{1140}{10626} = \frac{190}{1771}$$

- b) The number of ways to choose no defective is C_4^{20} .

Hence

$$P(\text{no defective}) = \frac{C_4^{20}}{C_4^{24}} = \frac{4845}{10626} = \frac{1615}{3542}$$

- c) The number of ways to choose four defectives from four is C_4^4 , or 1.

Hence $P(\text{all defective}) = \frac{C_4^4}{C_4^{24}} = \frac{1}{10626}$

- d) To find the probability of at least one defective transistor, find the probability that there are no defective transistors, and then subtract that probability from 1.

$$P(\text{at least 1 defective}) = 1 - P(\text{no defective}) = 1 - \frac{C_4^{20}}{C_4^{24}} = \frac{1927}{3542}.$$

Exercises

- 1.** How many permutations of the set $\{a, b, c, d, e\}$ begin with a and end with c ?
- 2.** How many different messages can be sent by five dashes and three dots?
- 3.** Roman has eight guests, two of whom are Jane and John. If the guests will arrive in a random order, what is the probability that John will not arrive right after Jane?
- 4.** Find the number of distinguishable permutations of the letters MISSISSIPPI.
- 5.** There are 20 chairs in a room numbered 1 through 20. If eight girls and 12 boys sit on these chairs at random, what is the probability that the thirteenth chair is occupied by a boy?
- 6.** If we put five math, six biology, eight history, and three literature books on a bookshelf at random, what is the probability that all the math books are together?
- 7.** Five boys and five girls sit in a row at random. What is the probability that the boys are together and the girls are together?
- 8.** A man has 20 friends. If he decides to invite six of them to his birthday party, how many choices does he have?
- 9.** A panel consists of 20 men and 20 women. How many choices do we have for a jury of six men and six women from this panel?

10. In a company there are seven executives: four women and three men. Three are selected to attend a management seminar. Find the following probabilities:

- a) All three selected will be women.
- b) All three selected will be men.
- c) Two men and one woman will be selected.
- d) One man and two women will be selected.

11. In a class of 18 students, there are 11 men and seven women. Four students are selected to present a demonstration on the use of the calculator. Find the probability that the group consists of the following:

- a) All men
- b) All women
- c) Three men and one woman
- d) One man and three women
- e) Two men and two women.

12. A committee of four people is to be formed from six doctors and eight dentists. Find the probability that the committee will consist of the following:

- a) All dentists
- b) Two dentists and two doctors
- c) All doctors
- d) Three doctors and dentist
- e) One doctor and three dentists

13. From a faculty of six professors, six associate professors, 10 assistant professors, and 12 instructors, a committee of size 6 is formed randomly. What is probability that

- a) There are exactly two professors on the committee?
- b) All committee members are of the same rank?

14. Almas has three sets of classics in literature, each set having four volumes. In how many ways can he put them in a bookshelf so that books of each set are not separated?

Answers

- 1.** 6; **2.** 40320; **3.** 0.875; **4.** 34 650; **5.** 0.6; **6.** 0.00068; **7.** 0.0079; **8.** 38 760;
9. 1502337600; **10.** a) $4/35$; b) $1/35$; c) $12/35$; d) $18/35$; **11.** a) $11/102$;
b) $7/612$; c) $77/204$; d) $77/612$; e) $77/204$; **12.** a) $10/143$; b) $60/143$;
c) $15/1001$; d) $160/1001$; e) $48/143$; **13.** a) 0.228; b) 0.00084; **14.** 82944.

2.8. Basic theorems

Theorem 1:

Let A be an event and \bar{A} its complement. Then the **complement rule** is:

$$P(\bar{A}) = 1 - P(A)$$

In words, the probability of the occurrence of any event equals one minus the probability of the occurrence of its complementary event.

Proof: Since events A and \bar{A} are mutually exclusive. Thus

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}), \text{ But } A \cup \bar{A} = S \text{ and } P(S) = 1, \text{ so}$$

$$1 = P(S) = P(A) + P(\bar{A}). \text{ Therefore, } P(\bar{A}) = 1 - P(A)$$

Example:

A club has a membership of six men and four women. A three-person committee is chosen at random. What is the probability that at least one woman will be selected?

Solution:

Let A - be the event “at least one woman will be selected”. We will start solution by computing the probability of the complement: \bar{A} -“no woman is selected”, and then using the complement rule will compute the probability of A .

$$P(\bar{A}) = \frac{C_3^6}{C_3^{10}} = \frac{1}{6}$$

And therefore the required probability is $P(\bar{A}) = 1 - P(A) = 1 - \frac{1}{6} = \frac{5}{6}$.

Theorem 2. If $A \subseteq B$, then

$$P(B - A) = P(B \cap \bar{A}) = P(B) - P(A)$$

Proof. If $A \subseteq B$, then it is clear that, $B = (B - A) \cup A$. But $(B - A) \cap A = \emptyset$, so events $(B - A)$ and A are mutually exclusive, and

$$P(B) = P((B - A) \cup A)) = P(B - A) + P(A). \text{ It gives}$$

$$P(B - A) = P(B) - P(A)$$

Theorem 3. (The addition rule of probabilities)

Let A and B be two events. The probability of their union is

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

Proof: $A \cup B = A \cup (B - (A \cap B))$, But $A \cap (B - (A \cap B)) = \emptyset$, so events A and $(B - (A \cap B))$ are mutually exclusive events and

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

Now using Theorem 2, we obtain, that

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

Example:

The probability that a randomly selected student from a university is a senior is 0.18, a business major is 0.14, and a senior and a business major is 0.04. Find the probability that a student selected at random from this university is a senior or a business major.

Solution:

Let A - be the event “Chosen student is a senior student” and B the event “Chosen student is a business major student”. Thus we have

$$P(A) = 0.18, \quad P(B) = 0.14 \quad \text{and} \quad P(A \cap B) = 0.04$$

The required probability is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 $= 0.018 + 0.014 - 0.04 = 0.28$.

Example:

Suppose that in a community of 400 adults, 300 bike or swim or do both, 160 swim and 120 swim and bike. What is the probability that an adult selected at random from this community bikes?

Solution:

Let A be the event that person swims and B be the event that he or she bikes, then $P(A \cup B) = 300/400$; $P(A) = 160/400$; $P(A \cap B) = 120/400$. Hence the relation $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ implies that

$$P(B) = P(A \cup B) + P(A \cap B) - P(A) = \frac{300}{400} + \frac{120}{400} - \frac{160}{400} = 0.65.$$

Exercises

- 1.** The probability that a randomly selected elementary school teacher from a city is a female is 0.68, holds a second job is 0.42, and is a female and holds a second job is 0.29. Find the probability that an elementary school teacher selected at random from this city is a female or holds a second job?
- 2.** It was estimated that 35% of all students were seriously concerned about employment prospects, 20% were seriously concerned about grades, and 15% were seriously concerned about both. What is the probability that a

randomly chosen student is seriously concerned about at least one of these two things?

3. It was found that 45% of students think that professors must be “more tolerant” to the students. If a student is selected randomly what is the probability that he or she will disagree or have no opinion on the issue.

4. In a statistics class there are 18 juniors and 10 seniors; 6 of the seniors are females, and 12 of the juniors are males. If a student is selected at random, find the probability of the following:

- a) A junior or a female
- b) A senior or a female
- c) A junior or a senior

5. If a die is rolled three times, find the probability of getting at least one 6.

6. If a die is rolled three times, find the probability of getting at least one even number.

7. A number is selected at random from the set of natural numbers

{1,2,3,.....,1000} . What is the probability that it is divisible by 4 but neither by 5 nor by 7?

8. There are four tickets numbered 1, 2, 3, and 4. Suppose a two-digit number will be formed by first drawing one ticket at random and then drawing a second ticket at random from remaining three.(For instance, if the first ticket drawn shows 4 and the second shows 1, the number recorded is 41.) List the sample space and determine the following probabilities

- a) What is the probability of getting an even number?
- b) What is the probability of getting a number larger than 20?
- c) What is the probability that obtained number is between 22 and 30?

9. A three-digit number is formed by arranging the digits 1, 5, and 6 in a random order.

- List the sample space.
- Find the probability of getting a number larger than 400.
- What is the probability that an even number is obtained?

Answer

1. 0.81; **2.** 0.40; **3.** 0.55; **4.** a) 6/7; b) 4/7; c) 1; **5.** 91/216; **6.** 7/8; **7.** 0.172; **8.** a) 0.5; b) 0.75; c) 0.167; **9.** b) 2/3; c) 1/3;

2.8.1. Conditional probability

Definition:

Conditional probability is the probability that an event will occur given that another event has already occurred. If A and B are two events, then the conditional probability of A is written as $P(A / B)$ and read as “the probability of A given that B has already occurred”.

If A and B are two events, then

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

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

given that $P(A) > 0$ and $P(B) > 0$.

Example:

A box contains black chips and white chips. A person selects two chips without replacement. If the probability of selecting a black chip and a white chip is $15/56$, and the probability of selecting a black chip on the first draw is $3/8$, find the probability of selecting the white chip on the second draw, given that the first chip selected was a black chip.

Solution:

Let B =selecting a black chip

W =selecting a white chip.

$$\text{Then } P(W / B) = \frac{P(W \cap B)}{P(B)} = \frac{15/56}{3/8} = \frac{5}{7}.$$

Hence, the probability of selecting a white chip on the second draw given that the first chip selected was black is $5/7$.

Example:

In a certain region of Kazakhstan, the probability that a person lives at least 80 years is 0.75 and the probability that he or she lives at least 90 years is 0.63. What is the probability that randomly selected 80-year old person from this region will survive to become 90?

Solution:

Let A and B be the events that the person selected survives to become 90 and 80 years old, respectively. We are interested in $P(A / B)$. By definition,

$$P(A / B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.63}{0.75} = 0.84$$

(Note that in this case $P(A \cap B) = P(A)$).

2.8.2. The multiplication rule of probability

Let A and B be two events. The probability of their intersection is

$$P(A \cap B) = P(A / B) \cdot P(B)$$

Also $P(A \cap B) = P(B / A) \cdot P(A)$

Example:

Suppose that seven nondefective and three defective goods have been mixed up. To find defective goods, we test them one by one, at random, and without replacement. What is the probability that we are lucky and find both of the defective goods in the first two tests?

Solution:

Let D_1 and D_2 be the events of finding defective goods in the first and second tests respectively. We are interested in

$$P(D_1 \cap D_2) = P(D_2 / D_1) \cdot P(D_1)$$

As we know, there are three defective goods in total 10 goods.

Consequently, the probability of selecting a defective good at the first selection is $P(D_1) = \frac{3}{10}$. To calculate the probability $P(D_2 / D_1)$, we know

that the first good is defective because D_1 has already occurred. Because the selections are made without replacement, there are 9 total goods and 2 of them are defective at the time of the second selection. Therefore

$$P(D_2 / D_1) = 2/9. \text{ Hence the required probability is}$$

$$P(D_1 \cap D_2) = P(D_2 / D_1) \cdot P(D_1) = \frac{3}{10} \cdot \frac{2}{9} = \frac{1}{15}.$$

Remark: Multiplication rule can be generalized for calculating the probability of the joint occurrence of several events.

For example, if $P(A \cap B) > 0$, then

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

2.8.3. Multiplication rule for independent events

Two events A and B are **independent** if $P(A / B) = P(A)$

Equivalent conditions are $P(B / A) = P(B)$ or $P(A \cap B) = P(A) \cdot P(B)$.

Example:

An urn contains three red balls, two blue balls, and five white balls. A ball is selected and its colour is noted. Then it is replaced. A second ball is selected and its colour is noted. Find the probability of

- a) Selecting two blue balls
- b) Selecting a blue and then white ball
- c) Selecting a red ball and then a blue ball

Solution:

a) $P(\text{blue and blue}) = P(\text{blue}) \cdot P(\text{blue}) = \frac{2}{10} \cdot \frac{2}{10} = \frac{1}{25}$

b) $P(\text{blue and white}) = P(\text{blue}) \cdot P(\text{white}) = \frac{2}{10} \cdot \frac{5}{10} = \frac{1}{10}$

c) $P(\text{red and blue}) = P(\text{red}) \cdot P(\text{blue}) = \frac{3}{10} \cdot \frac{2}{10} = \frac{3}{50}$

Example:

An urn contains five red and seven blue balls. Suppose that two balls are selected at random with replacement. Let A and B be the events that the first and the second balls are red, respectively. Then we get $P(A \cap B) = \frac{5}{12} \cdot \frac{5}{12}$.

Now $P(A \cap B) = P(A)P(B)$ since $P(A) = \frac{5}{12}$ and $P(B) = \frac{5}{12}$.

Thus A and B are independent.

If we do the same experiment without replacement, then $P(B/A) = \frac{4}{11}$

while $P(B) = P(B/A) \cdot P(A) + P(B/\bar{A}) \cdot P(\bar{A}) = \frac{4}{11} \frac{5}{12} + \frac{5}{11} \frac{7}{12} = \frac{5}{12}$

as expected. Thus $P(B/A) \neq P(B)$, implying that A and B are dependent.

Remark:

Multiplication rule for independent events can also be extended to three or more independent events by using the formula

$$P(A \cap B \cap C \cap \dots \cap K) = P(A) \cdot P(B) \cdot P(C) \cdot \dots \cdot P(K).$$

Example:

The probability that a specific medical test will show positive is 0.32. If four people are tested, find the probability that all four will show positive.

Solution:

Let T_i ($i=1, 2, 3, 4$) be the symbol for a positive test result.

$$P(T_1 \cap T_2 \cap T_3 \cap T_4) = P(T_1) \cdot P(T_2) \cdot P(T_3) \cdot P(T_4) = 0.32^4 = 0.010$$

Exercises

1. Suppose that $P(A) = 0.30$, $P(B) = 0.25$, and $P(A \cap B) = 0.20$

a) Find $P(A \cup B)$, $P(A/B)$, $P(B/A)$.

b) Are the events A and B independent? Why or why not?

2. Suppose that $P(A) = 0.68$, $P(B) = 0.55$, and $P(A \cap B) = 0.32$. Find

a) The conditional probability that B occurs, given that A occurs.

b) The conditional probability that B does not occur given that A occurs.

c) The conditional probability that B occurs given that A does not occur.

3. Concerning the events A and B , the following probabilities are given

$$P(B) = \frac{1}{3}; P(A/B) = \frac{2}{3}; P(A/\bar{B}) = \frac{3}{7}.$$

Determine

a) $P(A \cap \bar{B})$; b) $P(A)$; c) $P(\bar{B}/A)$

4. In a study of television viewing habits among married couples, a researcher found that for a popular Saturday night program 25% of the husbands viewed the program regularly and 30% of the wives viewed the program regularly. The study found that for couples where the husband watches the program regularly 80% of the wives also watch regularly.

- a) What is the probability that both husband and wife watch the program regularly?
- b) What is the probability that at least one-husband or wife-watches the program regularly?
- c) What percentage of married couples do not have at least one regular viewer of the program?
- 5.** Of 20 rats in a laboratory, 12 are males and 9 are infected with a virus. Of the 12 male rats, 7 infected with the virus. One rat is randomly selected from the laboratory.
- a) If the selected rat is found to be infected, what is the probability that it is a female?
- b) If the selected rat is found to be a male, what is the probability that it is infected?
- c) Are the events “the selected rat is infected” and “the selected rat is male” independent? Why or why not?
- 6.** Suppose $P(A) = 0.50$, $P(B) = 0.22$.
- Determine $P(A \cup B)$ if A and B are independent.
 - Determine $P(A \cup B)$ if A and B are mutually exclusive.
 - Find $P(A / \bar{B})$ if A and B are mutually exclusive.
- 7.** An urn has three red and five blue balls. Suppose that 8 balls are selected at random and with replacement. What is the probability that the first three are red and the rest are blue balls?

Answer

- 1.** a) $P(A \cup B) = 0.35$; $P(A / B) = 0.8$; $P(B / A) = 0.67$; b) No;

2. a) 0.471; b) 0.529; c) 0.719; 3. a) 2/7; b) 32/63; c) 9/16; 4. a) 0.2;
 b) 0.35; c) 65%; 5. a) 2/9; b) 7/12; c) No; 6. a) 0.61; b) 0.72; c) 0.641;
7. 0.00503;

2.8.4. The law of total probability

Theorem:

Let B be an event with $P(B) > 0$ and $P(\bar{B}) > 0$. Then for any event A ,

$$P(A) = P(A/B)P(B) + P(A/\bar{B})P(\bar{B}).$$

Example:

An urn contains 10 white and 6 red balls. Two balls are selected at random without replacement. What is the probability that second selected ball is red?

Solution:

Let A be the event that second selected ball is red, B be event that the first ball is white. Then $P(B) = \frac{10}{16}$, $P(\bar{B}) = \frac{6}{16}$, $P(A/B) = \frac{6}{15}$, $P(A/\bar{B}) = \frac{5}{15}$.

Then by the law of total probability:

$$P(A) = P(A/B)P(B) + P(A/\bar{B})P(\bar{B}) = \frac{6}{15} \frac{10}{16} + \frac{5}{15} \frac{6}{16} = \frac{3}{8}.$$

Theorem:

Let $\{B_1, B_2, \dots, B_n\}$ be a set of nonempty, mutually exclusive subsets of the sample space S and $P(B_i) > 0$ for $i = 1, 2, \dots, n$, then for any event A of S ,

$$P(A) = P(A/B_1)P(B_1) + P(A/B_2)P(B_2) + \dots + P(A/B_n)P(B_n) = \\ = \sum_{i=1}^n P(A/B_i)P(B_i).$$

Example:

Suppose that 70% of seniors, 60% of juniors, 55% of the sophomores, and 40% of the freshmen of a university use the library frequently. If 35% of all students are freshmen, 30% are sophomores, 20% are juniors, and 15% are seniors, what percent of all students use the library frequently?

Solution:

Let A be the event that a randomly selected student is using library frequently. Let F , O , J , and E be the events that he or she is a freshmen, sophomore, junior, or senior respectively. Thus

$$P(A) = P(A/F)P(F) + P(A/O)P(O) + P(A/J)P(J) + \\ + P(A/E)P(E) = 0.4 \cdot 0.35 + 0.55 \cdot 0.3 + 0.6 \cdot 0.2 + 0.7 \cdot 0.15 = 0.53.$$

Therefore, 53% of these students use the library frequently.

Exercises

- 1.** In a country men constitute 58% of the labour force. The rates of unemployment are 6.2% and 4.3% among males and females respectively.
- What is the overall rate of unemployed in the country?
 - If a worker selected at random is found to be unemployed, what is the probability that the worker is a woman?

2. In a shipment of 15 air conditioners, there are 4 with defective thermostats. Two air conditioners will be selected at random and inspected one after another. Find the probability that

- a) The first is defective.
- b) The first is defective and the second good.
- c) Both are defective.
- d) The second air conditioner is defective.
- e) Exactly one is defective.

3. Suppose that 40% of the students are girls. If 25 % of the girls and 15% of the boys of this university are A students, what is the probability that randomly selected student is A student?

4. A factory produces all its products by three machines. Machines I; II; and III produces 40%; 40% and 20% of the output, where 5%, 4%, and 2% of their outputs are defective, respectively. What percentage of the total product is defective?

5. A box contains 18 tennis balls, of which eight are new. Suppose that three balls are selected randomly, played with, and after play are returned to the box. If another three balls are selected for a second play, what is the probability that they are all new?

6. In an economical college all students are required to take calculus and economics course. Statistics shows that 37 % of the students of this college get A's in calculus and 25 % of them get A's in both economics and calculus. If randomly selected student of this college has passed calculus with an A, what is the probability that he or she got A in economics?

7. Suppose that 12 % of the population of a country are unemployed women and 17 % of population are unemploymed. What percentage of the unemployed are women?

Answer

1. a) 5.4%; b) 0.334; 2. a) 4/15; b) 22/105; c) 2/35; d) 4/15; e) 44/105;
3. 0.19; 4. 4%; 5. 0.148; 6. 0.676; 7. 70.6 %.

2.9. Bayes' theorem

Often, we begin our analysis with initial or *prior* probability estimates for specific events of interest. Then, from sources such as a sample, a special report, a product test, etc., we obtain some additional information about the events. Given this new information, we want to revise and update the prior probability values. The new or revised probabilities for the events are referred to as *posterior* probabilities. ***Bayes' theorem***, which will be presented shortly, provides a means of computing these revised probabilities. To introduce Bayes' formula, let us consider the following example:

Example 1:

In the factory 40%, 30%, and 30% of the goods is produced by machines I, II, and III, respectively. If 5%, 4%, and 3% of the outputs of these machines is defective, what is the probability that a randomly selected good that is found to be defective is produced by machine III?

Solution:

Let A be the event that a randomly selected good is defective and B_3 be the event that it is produced by machine III. We are asked to find $P(B_3 / A)$. We know that

$$P(B_3 / A) = \frac{P(B_3 \cap A)}{P(A)}$$

To find $P(B_3 \cap A)$, note that since $P(A / B_3)$ and $P(B_3)$ are known, we can use relation $P(B_3 \cap A) = P(A / B_3) \cdot P(B_3)$.

To calculate $P(A)$, we use the law of total probability. Let B_1 and B_2 be the events that the good is produced by machines I and II, respectively. Hence,

$$P(A) = P(A / B_1)P(B_1) + P(A / B_2)P(B_2) + P(A / B_3)P(B_3)$$

By substituting we obtain

$$\begin{aligned} P(B_3 / A) &= \frac{P(B_3 \cap A)}{P(A)} = \\ &= \frac{P(A / B_3) \cdot P(B_3)}{P(A / B_1)P(B_1) + P(A / B_2)P(B_2) + P(A / B_3)P(B_3)} = \\ &= \frac{0.03 \cdot 0.3}{0.05 \cdot 0.4 + 0.04 \cdot 0.3 + 0.03 \cdot 0.3} \approx 0.22. \end{aligned}$$

The formula for $P(B_3 / A)$ is a particular case of Bayes' formula.

To write formula for $P(B_3 / A)$ we can use tree diagram. (Fig. 2.1).

Letter D stands for “defective” and N for “not defective”.

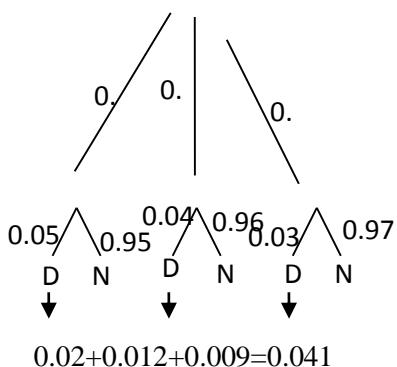


Fig.2.1. For example 1.

Theorem: (Bayes' theorem)

Let A and B be two events. Then Bayes' theorem states that

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)} \quad \text{and}$$
$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Theorem (Bayes'theorem, general form):

Let $B_1, B_2, B_3, \dots, B_n$ be n mutually exclusive and collectively exhaustive events of the sample space S . Then for any other event A of S with $P(A) > 0$

$$P(B_n/A) = \frac{P(A/B_n) \cdot P(B_n)}{P(A/B_1)P(B_1) + P(A/B_2)P(B_2) + \dots + P(A/B_n)P(B_n)}$$

Example 2:

A box contains 8 red and 11 blue balls. Two balls are selected at random without replacement and without their colour being seen. If the third ball is drawn randomly and observed to be red, what is the probability that both of previous selected balls were blue?

Solution: Let BB , BR , and RR be the events that first two selected balls are blue and blue, blue and red, and red and red. Let R be the event that the third ball drawn is red. We need to find $P(BB/R)$. Using Bayes' formula:

$$P(BB/R) = \frac{P(R/BB) \cdot P(BB)}{P(R/BB)P(BB) + P(R/BR)P(BR) + \dots + P(R/RR)P(RR)}.$$

Now $P(BB) = \frac{11}{19} \cdot \frac{10}{18} = \frac{55}{171}$

$$P(RR) = \frac{8}{19} \cdot \frac{7}{18} = \frac{28}{171}$$

and

$$P(BR) = \frac{11}{19} \cdot \frac{8}{18} + \frac{8}{19} \cdot \frac{11}{18} = \frac{88}{171},$$

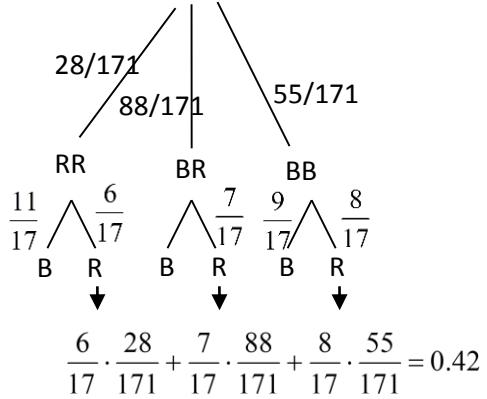


Fig.2.2. for example 2.

where BR is the union of two events:

namely, the first ball was blue, the second was red, and vice versa.

Thus,

$$P(BB/R) = \frac{\frac{8}{17} \cdot \frac{55}{171}}{\frac{8}{17} \cdot \frac{55}{171} + \frac{7}{17} \cdot \frac{88}{171} + \frac{6}{17} \cdot \frac{28}{171}} \approx 0.36.$$

This can be found easily from tree diagram on Fig. 2.2. as well.

Exercises

1. Given that $P(A_1) = 0.66$, $P(A_2) = 0.34$, $P(B/A_1) = 0.57$,

$P(C/A_1) = 0.43$, $P(B/A_2) = 0.61$, $P(C/A_2) = 0.39$, find the following probabilities: $P(A_1/B)$, $P(A_2/B)$, $P(A_1/C)$, and $P(A_2/C)$.

2. A store purchases electric irons from two companies. From company *A*, 500 irons are purchased, and 2% are defective. From company *B*, 850 irons are purchased, and 2% are defective. Given that an iron is defective, find the probability that it came from company *B*.

3. A store owner purchases telephones from two companies. From company *A*, 350 telephones are purchased, and 2% are defective. From company *B* 1050 telephones are purchased, and 4% are defective. Given that a phone is defective, find the probability that it came from company *B*.

4. A certain cancer is found in 1 person in 5000. If a person does have the disease, in 92% of the cases the diagnostic procedure will show that he or she actually has it. If a person does not have the disease, the diagnostic procedure in 1 out of 500 cases gives a false positive result. Determine the probability that a person with a positive test result has the cancer.

5. Suppose that 5% of the men and 2% of the women working for a corporation make over 4000\$ a year. If 30% of the employees of the corporation are women, what percent of those who make over 4000\$ a year are women?

6. Company purchases a certain part from three suppliers A, B, and C. Supplier A supplies 60% of the parts, B supplies 30% and C supplies 10%. The quality of parts is known to vary among suppliers, with A, B, and C parts having 0.25%, 1%, and 2% defective rates, respectively.

- What percent of product of the company has a defect?
- When a defective part is found, which supplier is the likely source?

7. At a department store, 20% of all customers spend \$50 or less and 80% spend more than \$50 per visit. Of those who spend \$50 or less, 75% pay by cash or check and 25% pay by credit card. Of those who spend more than \$50, 30% pay by cash or check and 70% pay by credit card. One randomly selected customer, who made a purchase at this store, paid by credit card. What is the probability that this customer spent more than \$50 at his store?

8. Each day companies introduce thousands of new products in the market. Usually new products are test marketed before they are introduced for sale. The probability is 0.65 that a new product introduced by a company will be successful. For an eventually successful product, probability is 0.95 that 50% or more of the people included in the test like it. However, for an eventually unsuccessful product, probability is 0.20 that 50% or more of the people included in the test like it. A company recently introduced a new product. What is the probability that this product will be successful if less than 50% of the people included in the test like it?

9. Professor classifies students as most accurate (in calculations and writing), moderately accurate, or poorly accurate, and finds 50%, 40%, and 10% respectively of all students fall into these categories. Professor found that A grade was got by 70% of the most accurate, by 50% of the moderately accurate, and by 30% of the poorly accurate students.

- a) What is the probability that a randomly chosen student is A grade student?
- b) If the randomly selected student is A grade student, what is the probability that the student is from group of most accurate student?
- c) If randomly selected student is A grade student, what is the probability that the student is not from the group of most accurate students?

Answers

- 1.** $P(A_1 / B) = 0.645$; $P(A_2 / B) = 0.355$; $P(A_1 / C) = 0.682$;
 $P(A_2 / C) = 0.318$; **2.** 0.063; **3.** 0.857; **4.** 0.084; **5.** 14.63%; **6.** a) 0.65%;
b) B; **7.** 0.91; **8.** 0.104; **9.** a) 0.58; b) 0.6034; c) 0.3966.

2.10. Bivariate probabilities

Joint and marginal probabilities

Suppose that a problem involves two distinct sets of events that we label $A_1, A_2, A_3, \dots, A_n$, $B_1, B_2, B_3, \dots, B_k$. The events A_i and B_j are mutually exclusive and collectively exhaustive within their sets, but intersections $A_i \cap B_j$ can occur between all events from the two sets. These intersections can be considered as a basic outcome of a random experiment. Two sets of events considered jointly in this way, are called **bivariate**, and the probabilities are called **bivariate probabilities**.

Table 2.1.

	B_1	B_2	B_3	B_k
A_1	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$	$P(A_1 \cap B_3)$	$P(A_1 \cap B_k)$
A_2	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	$P(A_2 \cap B_3)$	$P(A_2 \cap B_k)$
.
.
.
A_n	$P(A_n \cap B_1)$	$P(A_n \cap B_2)$	$P(A_n \cap B_3)$	$P(A_n \cap B_k)$

Definition:

In the table 2.1., the intersection probabilities $P(A_i \cap B_j)$ are called joint probabilities. Marginal probability is the probability of a single event without consideration of any other event. Marginal probability can be computed by summing the corresponding row or column.

Example:

All the 420 employees of a company were asked if they smoke or not and whether they are university graduates or not. Based on this information, the following two-way classification table was prepared.

Table 2.2.

	University graduate	Not a university graduate
Smoker	35	80
Nonsmoker	130	175

In table 2.2. each box that contains a number is called a cell. There are four cells in table 2.2. Each cell gives the frequency for two characteristics.

For example, 35 employees in this group possess two characteristics: They are university graduates and smoke. We can interpret the number in other cells the same way.

By adding the row of totals and the column of totals to table 2.2, we write table 2.3.

Table 2.3.

	University graduate	Not a university graduate	Total
Smoker	35	80	115
Nonsmoker	130	175	305
Total	165	255	420

Suppose one employee is selected at random from these 420 employees. This employee may be classified either on the basis of smoker or non-smoker alone or on the basis of university graduate or not. If only one characteristic is considered at a time, the employee selected can be a smoker, non-smoker, a university graduate, or not a university graduate. The probability of each of

these four characteristics or events is called marginal probabilities because they calculated by dividing the corresponding row margins (totals for rows) or column margins (totals for the columns) by the grand total. For table 2.3., the marginal probabilities are calculated as follows:

$$P(\text{smoker}) = \frac{\text{Number of smoker}}{\text{Total number of employees}} = \frac{115}{420} = 0.274$$

$P(\text{smoker})=0.274$ can be interpreted as “The probability that randomly selected employee is a smoker is 0.274”. Similarly

$$P(\text{nonsmoker}) = \frac{305}{420} = 0.726$$

$$P(\text{university graduate}) = \frac{165}{420} = 0.393$$

$$P(\text{not a university graduate}) = \frac{255}{420} = 0.607$$

Now, suppose that one employee is selected at random from these 420 employees. Furthermore, assume that it is known that this (selected) employee is a smoker. In other words, the event that the employee selected is a smoker has already occurred. What is the probability that the employee selected is a university graduate?

This probability, $P(\text{university graduate/smoker})$, as we know, is called the conditional probability, and it is read as “the probability that the employee selected is a university graduate given that this employee is a smoker”. The required conditional probability is calculated as follows:

$$P(\text{university graduate/smoker}) =$$

$$= \frac{\text{Number of smokers who are university graduate}}{\text{total number of smokers}} = \frac{35}{115} = 0.304.$$

Example:

For the data of table 2.3. calculate the conditional probability that a randomly selected employee is a non-smoker given that this employee is not a university graduate.

Solution:

We are to compute the probability $P(\text{non-smoker} / \text{not a university graduate})$.

$$P(\text{non-smoker}/\text{not a university graduate}) = \\ = \frac{\text{Number of not a university graduates who do not smoke}}{\text{total number of not university graduates}} = \frac{175}{255} = 0.686.$$

The probability that randomly selected employee who is not a university graduate, does not smoke is 0.686.

Example:

Refer to the information on 420 employees given in table 2.3., are the events “smoker (S)” and “university graduate (U)” independent?

Solution:

If the occurrence of one event affects the probability of the occurrence of the other event then the two events are said to be dependent events. Using probability notation, the two events will be dependent if either

$$P(A / B) \neq P(A) \text{ or } P(B / A) \neq P(B).$$

Events S and U will be independent if $P(S) = P(S / U)$, otherwise they will be dependent.

Using the information given in table 2.3., we compute the following two probabilities

$$P(S) = \frac{115}{420} = 0.274; \text{ and } P(S / U) = \frac{35}{165} = 0.212.$$

Because these two probabilities are not equal, the two events are dependent. Here, dependence of events means that percentage of smokers is different from percentage between a university graduates.

Example:

A recent survey asked 100 people if they thought women should be permitted to participate in weightlifting competition. The results of the survey are shown in the table.

Gender	Yes	No	Total
Male	32	18	50
Female	8	42	50
Total	40	60	100

Find the probabilities

- That a randomly selected person is a male.
- The respondent answered “yes”, given that the respondent was a female
- The respondent was a male, given that the respondent answered “no”.

Solution:

Let M =respondent was a male; Y = respondent answered “Yes”

F =respondent was a female; N =respondent answered “No”.

- We need to compute the probability $P(M)$. The probability that randomly selected respondent is a male is obtained by dividing total number of row labelled “Male” (50) by the total number of respondents (100).

$$P(\text{male}) = \frac{50}{100} = \frac{1}{2}$$

b) The problem is to find $P(Y / F)$. The rule states

$$P(Y / F) = \frac{P(F \cap Y)}{P(F)}$$

The probability $P(F \cap Y)$ is the number of females who responded “yes” divided by the total number of respondents $P(F \cap Y) = \frac{8}{100}$.

The probability $P(F)$ is the probability of selecting a female:

$$P(F) = \frac{50}{100}.$$

Then

$$P(Y / F) = \frac{P(F \cap Y)}{P(F)} = \frac{8/100}{50/100} = \frac{4}{25}$$

c) The problem is to find $P(M / N)$

$$P(M / N) = \frac{P(M \cap N)}{P(N)} = \frac{18/100}{60/100} = \frac{3}{10}$$

Exercises

- 1.** The following table shows the probabilities concerning two events A and B .
- a) Determine the missing entries.

	B	\bar{B}
A	0.25	0.12
\bar{A}		
	0.40	

- b) What is the probability that A occurs and B does not occur?
- c) Find the probability that either A or B occurs.
- d) Find the probability that one of these events occurs and other does not.

2. A woman's clothing store owner buys from three companies:

A, B, and C. The purchases are shown below:

Product	Company A	Company B	Company C
Dresses	24	18	12
Blouses	13	36	15

If one item is selected at random, find the following probabilities:

- a) It is purchased from company A or it is a dress.
- b) It was purchased from company B or company C.
- c) It is a blouse or was purchased from company A.

3. In a statistics class there are 18 local and 10 foreign students: 6 of the foreign students are females, and 12 of the local students are males. If a student is selected at random, what is the probability that

- a) randomly selected student is a local or a female?
- b) randomly selected student is foreign or a female student?
- c) randomly selected student is a local or a foreign student?

4. Two thousand randomly selected adults were asked if they think they are financially better off than their parents. The following table gives the two-way classification of the responses based on the education levels of

the persons included in the survey and whether they are financially better off, the same, or worse off than their parents.

		<u>Education level</u>		
	Less than High school	High school	More than high school	
Better off	140	450	420	
Same	60	250	110	
Worse off	200	300	70	

Suppose one adult is selected at random from these 2000 adults, find the probability that this adult is

- a) financially better off his (her) parents or high school student;
- b) more than high school student or financially worse off than his (her) parents;
- c) financially better off his (her) parents or financially worse off his (her) parents;
- d) financially better off than his (her) parents
- e) financially better off than his (her) parents given that he (she) has less than high school education;
- f) financially worse off than his (her) parents given that he (she) has hight school education
- g) financially the same as his (her) parents given that he (she) has more than high school eduction.
- h) Are the events “better off ” and ”hight school education” mutually exclusive? What about the events “less than high school” and “more than high school”? Why or why not?
- i) Are the events “worse off “and “more than high school” independent? Why or why not?

5. Eighty students in a university cafeteria were asked if they favoured a ban on smoking in the cafeteria. The results of the survey are shown in the table.

Class	Favour	Oppose	No opinion
Freshmen	15	27	8
Sophomore	23	5	2

If a student is selected at random, find these probabilities:

- a) He or she opposes the ban, given that the student is a freshman.
- b) Given that the student favours the ban, the student is a sophomore.

6. The following table gives a two-way classification of 200 randomly selected purchases made at department store.

	Paid by cash/check	Paid by credit card
Male	24	46
Female	77	53

If one of these 200 purchases is selected at random, find the probability that it is

- a) made by a female
- b) paid by cash/check
- c) paid by credit card given that the purchase is made by a male
- d) made by a female given that it is paid by cash/check
- e) made by a female and paid by a credit card
- f) paid by cash/check or made by a male

g) Are the events “female” and “paid by credit card” independent? Are they mutually exclusive? Explain why or why not.

7. Three cable channels (6, 8, and 10) have quiz shows, comedies, and dramas. The table gives proportions in the nine joint classifications

Type of show \ Channels	Channel 6	Channel 8	Channel 10
Quiz show	0.21	0.11	0.06
Comedy	0.08	0.21	0.08
Drama	0.01	0.07	0.17

- What proportion of shows is quiz show?
- What proportion of shows does Channel 6 have?
- If randomly selected show is quiz show, what is the probability that it was shown on channel 6?
- If the show was shown on channel 10, what is the probability that it was comedy?
- What is the probability that randomly chosen show is drama, or shown on channel 8, or both?

8. A supermarket manager classified customers according to whether their visits to the store as frequent or infrequent and whether they often, sometimes, or never make a purchase. The accompanying table gives the proportions of people surveyed in each of six joint classifications:

Making purchase	Often	Sometimes	Never
Frequency of visit			
Frequent	0.12	0.48	0.19
Infrequent	0.07	0.06	0.08

- a) What is the probability that a customer is both a frequent shopper and often purchases?
- b) What is the probability that a customer who never makes purchase visits the store frequently?
- c) Are the events “Never makes a purchase” and “Visits the store frequently” independent?
- d) What is the probability that a customer who infrequently visits the store often makes a purchase?
- e) Are the events “Often makes a purchase” and “Visits the store infrequently” independent?
- f) What is the probability that a customer frequently visits the store?
- g) What is the probability that customer never makes a purchase in this store?
- h) What is the probability that a customer either frequently visits the store or never makes a purchase, or both?

9. The accompanying table shows proportions of salespeople classified according to marital status and whether or not they own stocks.

Own stocks	Yes	No
Martial status		
Married	0.64	0.13
Single	0.17	0.06

- a) What is the probability that a randomly chosen salesperson was married?
- b) What is the probability that a randomly chosen salesperson does not own stocks?
- c) What is the probability that randomly chosen single salesperson does not own stocks?
- d) What is the probability that a randomly chosen salesperson who owns stocks was married?

10. Forty-two percent of employees in a large corporation were in favour of a modified health care plan, and 22% of the corporation employees favoured a proposal to change the work schedule. Thirty-four percent of those favouring the health plan modification favoured the work schedule change.

- a) What is the probability that a randomly selected employee is in favour of both modified health care plan and the changed work schedule?
- b) What is the probability that a randomly chosen employee is in favour of at least one of these two changes?
- c) What is the probability that a randomly selected employeeavouring the work schedule change also favours the modified health plan?

Answers

- 1.** b) $P(A \cap \bar{B}) = 0.12$; c) $P(A \cup B) = 0.52$; d) $P(A \cap \bar{B} \cup \bar{A} \cap B) = 0.27$;
- 2.** a) 67/118; b) 81/118; c) 88/118; **3.** a) 6/7; b) 4/7; c) 1; **4.** a) 0.78; b) 0.55; c) 0.79; d) 0.505; e) 0.350; f) 0.300; g) 0.183; h) “Better off” and “high school education” are not mutually exclusive, “Less than high school” and “more than high school” are mutually exclusive events; i) “Worse off” and “more than high school” are not independent events; **5.** a) 0.54; b) 0.61; **6.** a) 0.65; b) 0.51; c) 0.66; d) 0.76; e) 0.27; f) 0.74; g) no and no; **7.** a) 0.38; b) 0.3;

- c) 0.55; d) 0.26; e) 0.57; **8.** a) 0.12; b) 0.704; c) No; d) 0.333; e) No; f) 0.79;
g) 0.27; h) 0.87; **9.** a) 0.77; b) 0.19; c) 0.2609; d) 0.7901; **10.** a) 0.1428;
b) 0.4972; c) 0.6491.

Chapter 3

Discrete random variables and probability distributions

3.1. Random variables

Suppose that experiment of rolling two fair dice to be carried out. Let X be the sum of outcomes, then X can only assume the values 2, 3, 4, ..., 12 with the following probabilities:

$$P(X = 2) = P\{(1,1)\} = \frac{1}{36}$$

$$P(X = 3) = P\{(1,2);(2,1)\} = \frac{2}{36}$$

$$P(X = 4) = P\{(1,3);(3,1);(2,2)\} = \frac{3}{36} \text{ and so on.}$$

The numerical value of random variable depends on the outcomes of the experiment. In this example, for instance, if it is (3, 2), then X is 5, and if it is (6, 6) then X is 12. In this example X is called a random variable.

Definition:

A **random** variable is a variable whose value is determined by the outcome of a random experiment.

Notationally, we use capital letters, such as X , to denote the random variable and corresponding lowercase x to denote a possible value.

Set of possible values of a random variable might be finite, infinite and countable, or uncountable.

Definition:

A random variable X is called a **discrete random variable** if it can take on no more than a countable number of values.

Some examples of discrete random variable:

1. The number of employees working at a company.
2. The number of heads obtained in three tosses of a coin.
3. The number of customers visiting a bank during any given day.

A random variable whose values are not countable is called a **continuous random variable**.

Definition:

A random variable X is called a **continuous** if it can take any value in an interval.

Here are some examples of continuous random variables:

1. Prices of houses
2. The amount of oil imported.
3. Time taken by workers to learn a job.

3.2. Probability distributions for Discrete Random Variables

Let X be a discrete random, and x be one of its possible values. The probability that the random variable X takes the value x is denoted by $P(X = x)$.

Definition:

The **probability distribution function**, $P(x)$, of a discrete random variable X indicates that this variable takes the value x , as a function of x . That is

$$P(x) = P(X = x), \text{ for all values of } x.$$

Example 1:

In the experiment of tossing a fair coin three times, let X be number of heads obtained. Determine and sketch the probability function of X .

Solution:

First, X is a variable and the number of heads in three tosses of a coin can have any of the values 0, 1, 2, or 3.

We can make a list of the outcomes and the associated values of X . (Table 3.1).

Note that, for each basic outcome there is only one value of X . However, several basic outcomes may yield the same value. We identify the events

(i.e., the collections of the distinct values of X). (Table 3.2)

Table 3.1

Outcome	Value of X
HHH	3
HHT	2
HTH	2
HTT	1
THH	2
THT	1
TTH	1
TTT	0

Table 3.2.

Numerical value of X as an event	Composition of the event
[X=0]	{TTT}
[X=1]	{HTT, THT, TTH}
[X=2]	{HHT, HTH, THH}
[X=3]	{HHH}

The model of a fair coin entails that 8 basic outcomes are equally likely, so each is assigned the probability 1/8.

The event [X=0] has a single outcome TTT, so its probability is 1/8.

Similarly, the probabilities of [X=1], [X=2], and [X=3] are found to be 3/8, 3/8, and 1/8, respectively. Collecting these results, we obtain the probability distribution of X shown in table 3.3.

Table 3.3. The probability distribution of X , the number of heads in 3 tosses of a coin.

Value of X	Probability
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

Remark: When summed over all possible values of X , these probabilities must add up to 1.

The graphical representation of the probability of X , the number of heads in three tosses of a coin is shown in Fig. 3.1.

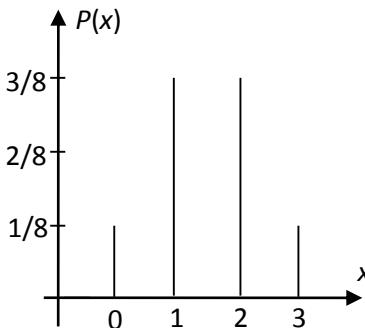


Fig. 3.1 Probability function for example 1.

In the development of the probability distribution for a discrete random variable, the following two conditions must always be satisfied:

Properties of probability function of discrete random variables:

Let X be a discrete random variable with probability $P(x)$. Then

1. $P(x) \geq 0$ for any value x .

2. The individual probabilities sum to 1; that is $\sum_x P(x) = 1$, where the

notation \sum_x indicates summation over all possible values of x .

Another representation of discrete probability distribution is also useful.

Cumulative probability function $F(x_0)$:

The cumulative probability function, $F(x_0)$ of a random variable X expresses the probability that X does not exceed the value x_0 as a function of x_0 . That is

$$F(x_0) = P(X \leq x_0),$$

where the function is evaluated of all values x_0 .

$$F(x_0) = \sum_{x \leq x_0} P(x)$$

Properties of cumulative probability functions for discrete random variables:

Let X be a discrete random variable with cumulative probability function $F(x_0)$. Then we can show that

1. $0 \leq F(x_0) \leq 1$ for every number x_0 .
2. If x_1 and x_2 are two numbers with $x_1 < x_2$, then $F(x_1) < F(x_2)$.

Example 2:

In the experiment of rolling a balanced die twice, let X be the minimum of the two numbers obtained. Determine and sketch the probability function and cumulative probability function of X .

Solution:

The possible values of X are 1, 2, 3, 4, 5, and 6. The sample space of this experiment consists of 36 basic outcomes. Hence the probability of any of them is $1/36$. In our experiment

$$P(X=1)=P\{(1,1)(1,2)(2,1)(1,3)(3,1)(1,4)(4,1)(1,5)(5,1)(1,6)(6,1)\}=11/36$$

$$P(X=2)=P\{(2,2)(2,3)(3,2)(2,4)(4,2)(2,5)(5,2)(2,6)(6,2)\}=9/36$$

$$P(X=3)=P\{(3,3)(3,4)(4,3)(3,5)(5,3)(3,6)(6,3)\}=7/36$$

$$P(X=4)=P\{(4,4)(4,5)(5,4)(4,6)(6,4)\}=5/36$$

$$P(X=5)=P\{(5,5)(5,6)(6,5)\}=3/36$$

$$P(X=6)=P\{(6,6)\}=1/36$$

The graphical representation of $P(x)$ is shown in Fig.3.2.

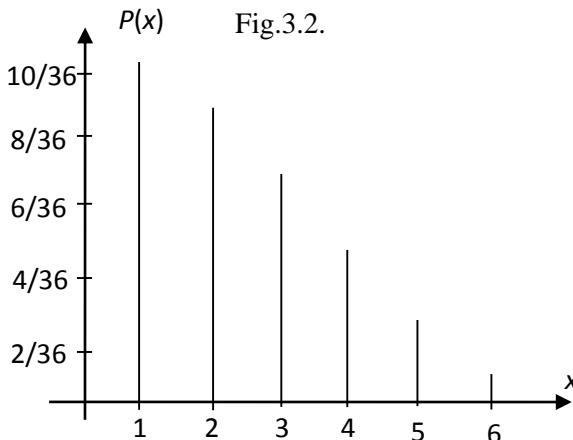


Fig. 3.2. Probability function for example 2.

Now let us form cumulative probability function.

If x_0 is some number less than 1, X can not be less than x_0 , so

$$F(x_0) = P(X \leq x_0) = 0 \text{ for all } x_0 < 1$$

If x_0 is greater than or equal to 1 but strictly less than 2, the only one number less than 2, the only way for X to be less than or equal to x_0 is if $X=1$. Hence

$$F(x_0) = P(X \leq x_0) = P(1) = 11/36 \text{ for all } 1 \leq x_0 < 2$$

If x_0 is greater than or equal to 2 but strictly less than 3, X is less than or equal to the x_0 if and only if either $X=1$ or $X=2$, so

$$F(x_0) = P(X \leq x_0) = P(1) + P(2) = 20/36 \text{ for } 2 \leq x_0 < 3.$$

Continuing in this way we can write cumulative probability function as

$$F(x_0) = \begin{cases} 0 & \text{if } x_0 < 1 \\ 11/36 & \text{if } 1 \leq x_0 < 2 \\ 20/36 & \text{if } 2 \leq x_0 < 3 \\ 27/36 & \text{if } 3 \leq x_0 < 4 \\ 32/36 & \text{if } 4 \leq x_0 < 5 \\ 35/36 & \text{if } 5 \leq x_0 < 6 \\ 1 & \text{if } x_0 \geq 6 \end{cases}$$

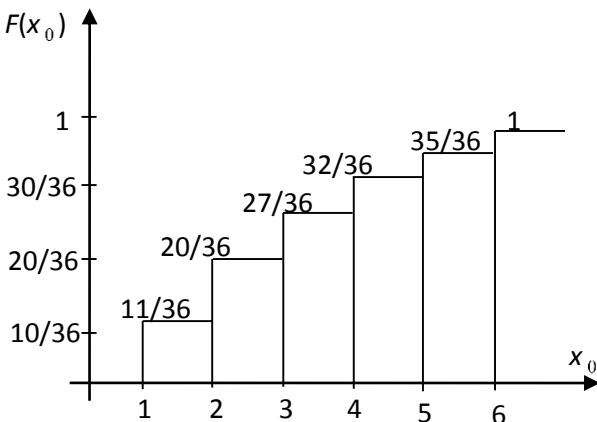


Fig. 3.3. Cumulative probability function for example 2.

The cumulative distribution function of X , $F(x_0)$, is plotted in Fig. 3.3. It can be seen that the cumulative probability function increases in steps until the sum is 1.

Example 3:

A consumer agency surveyed all 2500 families living in a small town to collect data on the number of TV sets owned by them. The following table lists the frequency distribution of the data collected by this agency

Number of TV sets owned	0	1	2	3	4
Number of families	120	970	730	410	270

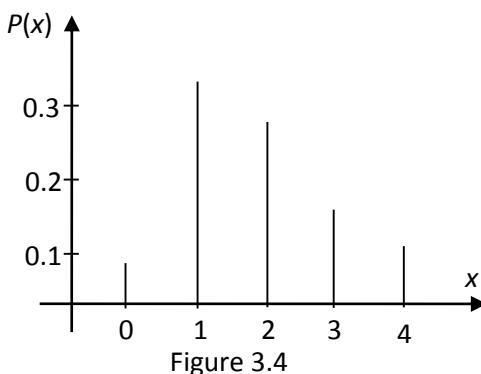
- a) Construct a probability distribution table. Draw a graph of the probability distribution.
- b) Calculate and draw the cumulative probability function.
- c) Find the probabilities: $P(X=1)$, $P(X>1)$, $P(X \leq 1)$, $P(1 \leq X \leq 3)$.

Solution:

- a) In a chapter 2 we learned that the relative frequencies obtained from an experiment or a sample can be used as approximate probabilities. Using the relative frequencies, we can write the probability distribution on the discrete random variable X in the following table.

Number of TV sets owned, x	Probability $P(x)$
0	120/2500=0.048
1	970/2500=0.388
2	730/2500=0.292
3	410/2500=0.164
4	270/2500=0.108

Figure 3.4 shows the graphical presentation of the probability distribution.



b) Let us form cumulative probability distribution function.

If x_0 is less than 0, then

$$F(x_0) = P(X \leq x_0) = 0 \text{ for } x_0 < 0$$

If x_0 is less than 1, then

$$F(x_0) = P(X \leq x_0) = P(0) = 0.048 \text{ for } 0 \leq x_0 < 1$$

Continuing in this way, we obtain

$$F(x_0) = \begin{cases} 0 & \text{if } x_0 < 0 \\ 0.048 & \text{if } 0 \leq x_0 < 1 \\ 0.436 & \text{if } 1 \leq x_0 < 2 \\ 0.728 & \text{if } 2 \leq x_0 < 3 \\ 0.892 & \text{if } 3 \leq x_0 < 4 \\ 1 & \text{if } x_0 \geq 4 \end{cases}$$

This function is plotted in Fig. 3.5.

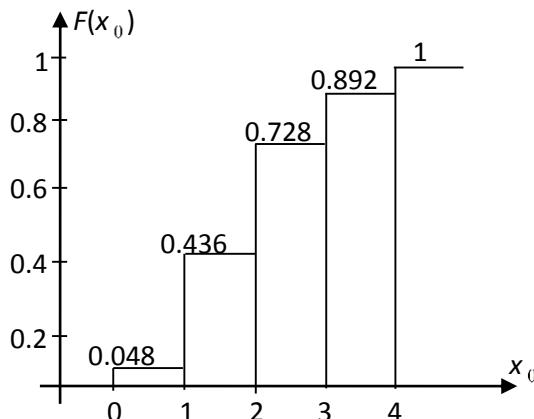


Fig. 3.5. Cumulative probability function for example 3.

c) $P(X = 1) = 0.388$

$$P(X > 1) = 1 - F(1) = 1 - 0.436 = 0.564$$

$$P(X \leq 1) = F(1) = 0.436$$

$$P(1 \leq X \leq 3) = F(3) - F(0) = 0.892 - 0.048 = 0.844.$$

Exercises

1. Each of the following tables lists certain values of X and their probabilities. Determine if each of them satisfies the two conditions required for a valid probability distribution.

a)

b)

c)

X	$P(x)$
5	-0.39
6	0.67
7	0.31
8	0.28

X	$P(x)$
2	0.22
3	0.23
5	0.65

X	$P(x)$
0	0.16
1	0.00
2	0.43
3	0.41

2. For each case, list the values of x and $P(x)$ and examine if the specification represents a probability distribution. If does not, state what properties are violated

a) $P(x) = \frac{1}{10}(x - 2)$ for $x = 3, 4, 5, 6$

b) $P(x) = \frac{1}{2}(x - 2)$ for $x = 1, 2, 3, 4$

c) $P(x) = \frac{1}{20}(2x + 4)$ for $x = -2, -1, 0, 1, 2$

d) $P(x) = 3/2^x$ for $x = 2, 3, 4, 5$

3. The following table gives the probability distribution of a discrete random variable X .

X	0	1	2	3	4	5
$P(x)$	0.03	0.13	0.22	0.31	0.19	0.12

- a) Draw the probability function.
- b) Calculate and draw the cumulative probability function.
- c) Find $P(X = 1)$; $P(X \leq 1)$; $P(X \geq 3)$; $P(0 \leq X \leq 2)$
- d) Find the probability that x assumes a value less than 3
- e) Find the probability that x assumes a value in the interval 2 to 4.

- 4.** Despite all safety measures, accidents do happen at the factory. Let X denote the number of accidents that occur during a month at this factory. The following table lists the probability distribution of X .

X	0	1	2	3	4
$P(x)$	0.25	0.30	0.20	0.15	0.10

- a) Draw the probability function.
- b) Calculate and draw the cumulative probability function.
- c) Determine the probability that the number of accidents that will occur during a given month at this company is exactly 4.
- d) What is the probability that number of accidents will be at least 2?
- e) What is the probability that number of accidents will be less than 3?
- f) What is the probability that number of accidents will be between 2 to 4?
- g) Two month are chosen at random. What is the probability that on both of these months there will be fewer than two accidents?

- 5.** Let X be the number of shopping trips made by family during a month. The following table lists the frequency distribution of X of 1000 families.

X	4	5	6	7	8	9	10
f	70	180	240	210	170	90	40

- a) Draw the probability function.
- b) Calculate and draw the cumulative probability function.
- c) Find the following probabilities $P(X = 5)$; $P(X > 6)$; $P(4 \leq X \leq 7)$; $P(X \leq 6)$.

- 6.** The following table lists the probability distribution of the number of phone calls received per 10-minute period at an office.

Number of phone calls	0	1	2	3	4
Probability	0.12	0.26	0.34	0.18	0.10

Let X denote the number of phone calls received during a certain 10-minute period at this office.

- a) Find the probabilities: $P(X = 1)$; $P(X < 2)$; $P(X > 2)$; $P(1 \leq X \leq 3)$.
- b) Two 10-minute period are chosen at random. What is the probability that at least one of them there will be at least one received call?
- 7.** In successive rolls of a fair die, let X be the number of rolls until the first 6. Determine the probability function.
- 8.** In a tennis championship, player A competes against player B in consecutive sets and the game continues until one player wins three sets. Assume that, for each set $P(\text{A wins}) = 0.4$, $P(\text{B wins}) = 0.6$, and the outcomes of different sets are independent. Let X stand for the number of sets played.
- a) List the possible values of X and identify the basic outcomes associated with each value.
- b) Obtain the probability distribution of X .

Answers

- 1.** a) this is not a valid probability distribution; b) this is not a valid probability distribution; c) this is a valid probability distribution;
- 2.** a) this is a valid probability distribution; b) this is not a valid probability distribution; c) this is a valid probability distribution;
- d) this is not a valid probability distribution; **3.** c) 0.13; 0.16; 0.62; 0.38; d) 0.38; e) 0.72; **4.** c) 0.10; d) 0.45; e) 0.75; f) 0.45; g) 0.3025; **5.** c) 0.18; 0.51; 0.70; 0.49; **6.** a) 0.26; 0.38; 0.28; 0.78; b) 0.9856;
- 7.** $P(x) = \left(\frac{5}{6}\right)^{x-1} \cdot \frac{1}{6}; x \geq 1$; **8.** b) $P(3) = 0.2800$; $P(4) = 0.3744$;
 $P(5) = 0.3456$;

3.3. Expected (mean) value and variance for discrete random variables

3.3.1. Expected value

Once we have constructed the probability distribution for a random variable, we often want to compute the mean or expected value of the random variable. The mean of discrete random variable X , denoted either μ_X or $E(X)$, is actually the mean of its probability distribution. The mean (or expected) value of a discrete random variable is the value that we expect to observe per repetition, on average, if we perform an experiment a large number of times. For example, we may expect a house salesperson to sell on average, 3.50 houses per month. It does not mean that every month this salesperson will sell exactly 3.50 houses. (Actually he (or she) can not sell exactly 3.50 houses). This simply means that if we observe for many months, this salesperson will sell a different number of houses different months. However, the average of all sold houses in these months will be 3.50.

Definition:

The mean (or expected value) of discrete random variable X is defined as

$$\mu_X = E(X) = \sum (\text{value} \cdot \text{probability}) = \sum_x x \cdot P(x)$$

Here the sum extends over all distinct values x of X .

In order to compute the expected value of a discrete random variable we must multiply each value of the random variable by the corresponding value of its probability function. We then add the resulting terms.

Example:

Sales show that five is the maximum number of cars sold on a given day at car selling company. Table 3.4 shows probability distribution of cars sold per day. Find the expected number of cars sold.

Table 3.4

Solution:

To find the expected number (or mean) of cars sold, we multiply each value of x by its probability and add these results.

x	$P(x)$	$x \cdot P(x)$
0	0.18	0.00
1	0.39	0.39
2	0.24	0.48
3	0.14	0.42
4	0.04	0.16
5	0.01	0.05

x	$P(x)$
0	0.18
1	0.39
2	0.24
3	0.14
4	0.04
5	0.01

$$\mu_x = E(X) = \sum_x x \cdot P(x) = 1.50$$

In fact, it is impossible for company to sell exactly 1.50 cars in any given day. But we examine selling cars at this company for many days into the future, and see that, the expected value of 1.50 cars provides a good estimate of the mean or average daily sales volume. The expected value can be important to the managers from both planning and decision making points of view.

For example, suppose that this company will be open 40 days during next 2-month. How many cars should the owner expect to be sold during this time?

While we can not specify the exact value of 1.50 cars, it provides an expected sale of $40 \cdot 1.50 = 60$ cars for the next 2-month period.

3.3.2. Variance and standard deviation of discrete random variable

While the expected value gives us an idea of the average or central value for the random variable, often we would also like to measure the dispersion or variability of the possible values of the random variable. The variance of discrete random variable X , denoted by σ_x^2 , measures the spread of its probability distribution. In defining the variance of a random variable, a

weighted average of the squares of its possible discrepancies about the means is formed; the weight associated with $(x - \mu_x)^2$ is the probability that the random variable takes the value x . The variance can be viewed as the average value that will be taken by the function $(X - \mu_x)^2$ over a very large number of repeated trials.

The mathematical expression for the variance of a discrete random variable is

$$\sigma_x^2 = E[(X - \mu_x)^2] = \sum_x (x - \mu_x)^2 P(x).$$

The standard deviation, σ_x , is the positive square root of the variance.

In some particular cases, an alternative but equivalent (sometimes called shortcut formula) formula for the variance can be used:

$$\sigma_x^2 = E(X^2) - \mu_x^2 = \sum_x x^2 P(x) - \mu_x^2$$

Example:

Find the variance for the example in previous topic, for the number of cars sold per day at a car selling company.

Solution:

Let us apply $\sigma_x^2 = E[(X - \mu_x)^2] = \sum_x (x - \mu_x)^2 P(x)$.

The calculations are shown in the table 3.5:

Table 3.5

x	$(x - \mu)$	$(x - \mu)^2$	$P(x)$	$(x - \mu)^2 \cdot P(x)$
0	0-1.50=-1.50	2.25	0.18	2.25·0.18=0.4050
1	1-1.50=-0.50	0.25	0.39	0.25·0.39=0.0975
2	2-1.50=0.50	0.25	0.24	0.25·0.24=0.0600
3	3-1.50=1.50	2.25	0.14	2.25·0.14=0.3150
4	4-1.50=2.50	6.25	0.04	6.25·0.04=0.2500
5	5-1.50=3.50	12.25	0.01	12.25·0.01=0.1225
				$\sum (x - \mu)^2 \cdot P(x) = 1.25$

We see that the variance for the number of cars sold per day is 1.25.

The standard deviation of the number of cars sold per day is

$$\sigma = \sqrt{1.25} = 1.118.$$

Remark:

For the purpose of easier managerial interpretation the standard deviation may be preferred over the variance because it is measured in the same units as the random variable.

Example:

The following table gives the probability distribution of X .

x	0	1	2	3	4	5
$P(x)$	0.02	0.20	0.30	0.30	0.10	0.08

Compute the standard deviation of x .

Solution:

Let us apply equivalent (shortcut formula) formula for the variance

$$\sigma_x^2 = \sum_x x^2 P(x) - \mu_x^2$$

The following table shows all the calculations required for the computation of the standard deviation of x .

x	$P(x)$	$x P(x)$	x^2	$x^2 \cdot P(x)$
0	0.02	0.00	0	0.00
1	0.20	0.20	1	0.20
2	0.30	0.60	4	1.20
3	0.30	0.90	9	2.70
4	0.10	0.40	16	1.60
5	0.08	0.40	25	2.00
		$\sum xP(x) = 2.50$		$\sum x^2 P(x) = 7.70$

We perform the following steps to compute the standard deviation by shortcut formula:

Step1: Compute the mean of discrete random variable:

$$\mu = \sum xP(x) = 2.50$$

Step2: Compute the value of $\sum x^2 P(x)$.

Step3: Substitute the values of μ and $\sum x^2 P(x)$ in the shortcut formula for the variance

$$\sigma_x^2 = \sum_x x^2 P(x) - \mu^2 = 7.70 - (2.50)^2 = 1.45$$

Step4: Take positive square root of variance.

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{1.45} = 1.20.$$

Example:

A farmer will earn a profit of \$30 thousand in case of heavy rain next year, \$60 thousand in case of a moderate rain, and \$15 thousand in case of little rain. A meteorologist forecasts that the probability is 0.35 for heavy rain, 0.40 for moderate rain, and 0.25 for little rain next year. Let X be the random variable that represents next year's profit in thousands of dollars for this farmer. Write the probability distribution of x . Find the mean and standard deviation of x . Give a brief interpretation of the values of the mean and standard deviation.

Solution:

The table 3.6 lists the probability distribution of x

Table 3.6

x	$P(x)$
30	0.35
60	0.40
15	0.25

The table 3.7 shows all calculations needed for the computation of the mean and standard deviation.

Table 3.7

x	$P(x)$	$x P(x)$	x^2	$x^2 \cdot P(x)$
30	0.35	10.5	900	315
60	0.40	24	3600	1440
15	0.25	3.75	225	56.25
		$\sum xP(x) = 38.25$		$\sum x^2 P(x) = 1811.25$

The mean of x is $\mu_x = \sum xP(x) = \$38.25$ thousand. The standard deviation is $\sigma_x = \sqrt{\sum x^2 P(x) - \mu_x^2} = \sqrt{1811.25 - (38.25)^2} = \18.660 .

Thus, it is expected that a farmer will earn an average of \$38.25 thousand profits in next year with a standard deviation of \$18.660 thousand.

3.3.3. Mean and variance of linear function of a random variable

Let X be a random variable that takes the value x with probability $P(x)$ and consider a new random variable Y , defined by $Y=a+bX$.

Suppose that random variable X has mean μ_X , and variance σ_X^2 .

Then mean and variance of Y are

$$\mu_Y = E(a + bX) = a + b \cdot \mu_X \quad \text{and}$$

$$\sigma_Y^2 = \text{Var}(a + bX) = b^2 \sigma_X^2 \text{ so that standard deviation of } Y \text{ is}$$

$$\sigma_Y = |b| \cdot \sigma_X .$$

Example:

A car salesman estimates the following probabilities for the number of cars that he will sell in next month.

Number cars	0	1	2	3	4
Probability	0.12	0.20	0.25	0.25	0.18

- a) Find the expected number of cars that will be sold in the next month.
- b) Find the standard deviation of the number of cars that will be sold in next month.
- c) The salesperson receives for the month a salary of \$300, plus an additional \$200 for each car sold. Find the mean and standard deviation of his total monthly salary.

Solution:

- a) The random variable X has mean

$$\begin{aligned}\mu_X &= E(X) = \sum_x xP(x) = \\ &= 0 \cdot 0.12 + 1 \cdot 0.20 + 2 \cdot 0.25 + 3 \cdot 0.25 + 4 \cdot 0.18 = 2.17.\end{aligned}$$

$$\begin{aligned}
 \text{b) Variance } \sigma_X^2 &= \sum_x (x - \mu_x)^2 \cdot P(x) = \\
 &= (0 - 2.17)^2 (0.12) + (1 - 2.17)^2 (0.20) + (2 - 2.17)^2 (0.25) + (3 - 2.17)^2 (0.25) + \\
 &\quad + (4 - 2.17)^2 (0.18) = 1.621 \\
 \sigma_X &= \sqrt{\sigma_X^2} = \sqrt{1.621} = 1.273.
 \end{aligned}$$

c) Total monthly salary of salesperson can be written as $Y = 300 + 200X$. Then

$$\begin{aligned}
 \mu_Y &= E(Y) = E(300 + 200X) = 300 + 200 \cdot \mu_X = 300 + 200 \cdot 2.17 = \$734. \\
 \text{Var}(\mu_Y) &= \text{Var}(300 + 200X) = 200^2 \cdot 1.621 = 64840. \\
 \sigma_{\mu_Y} &= \$254.64.
 \end{aligned}$$

Summary results for the mean and variance of special linear functions:

a) Let $b=0$ in the linear function, $Y = a + bX$. Then $Y = a$ for any constant a . $E(a) = a$ and $\text{Var}(a) = 0$

If a random variable always takes the value a , it will have a mean a and a variance 0.

b) Let $a=0$ in the linear function, $Y = a + bX$. Then $Y = bX$.

$$E(bX) = b \cdot \mu_X \text{ and } \text{var}(bX) = b^2 \sigma_X^2.$$

Exercises

1. Find the mean and standard deviation for each of the following probability distributions.

a)

X	$P(x)$
0	0.12
1	0.27
2	0.43
3	0.18

b)

X	$P(x)$
6	0.36
7	0.26
8	0.21
9	0.17

2. Given the following probability distribution.

Find $E(X)$, σ^2 , σ

X	$P(x)$
0	0.4
1	0.3
2	0.2
3	0.1

3.

Let x be the number of errors that a randomly selected page of a book contains. The following table lists the probability distribution of x .

X	0	1	2	3	4
$P(x)$	0.73	0.16	0.06	0.04	0.01

Find the mean and standard deviation.

4. Suppose the probability function of a random variable X is given by the formula $P(x) = \frac{60}{77} \cdot \frac{1}{x}$ for $x = 2, 3, 4, 5$

Calculate the mean and standard deviation of this distribution.

5. Given the two probability distributions

X	$P(x)$	X	$P(x)$
1	0.2	0	0.1
2	0.6	1	0.2
3	0.2	2	0.4
		3	0.2
		4	0.1

a) Verify that both distributions have the same mean.

b) Compare the two standard deviations.

6. An instant lottery ticket costs \$2. Out of a total of 10 000 tickets printed for this lottery, 1000 tickets contain a prize of \$5 each, 100 tickets have a prize of \$10 each, 5 tickets have a prize of \$1000 each, and 1 ticket has a prize of \$5000. Let x be the random variable that denotes the net amount a player wins by playing this lottery. Write the probability distribution of x .

Determine the mean and standard deviation of x . How will you interpret the values of the mean and standard deviation of x ?

7. A TV repairer estimates the probabilities for the number of hours required to complete some job as follows:

Time taken (Hours)	1	2	3	4	5
Probability	0.05	0.2	0.35	0.3	0.1

- Find the expected time to complete the job.
- The TV repairer's service is made up of two parts- a fixed cost of \$20, plus \$2 for each hour taken to complete the job. Find the mean and standard deviation of total cost.

8. Consider the following probability distribution for the random variable X .

X	$P(x)$
10	0.20
20	0.40
30	0.25
40	0.15

- Find the expected value of X .
- Find the variance and standard deviation.
- If $Y = 3X + 5$, find the expected value, variance, and standard deviation for Y .

Answers

1. a) 1.67; 0.906; b) 7.19; 1.102; 2. 1; 1; 1; 3. 0.44; 0.852; 4. 3.12; 1.09;
5. a) $\mu_1 = \mu_2 = 2$; b) $\sigma_1 = 0.63$; $\sigma_2 = 1.1$; 6. 1.6; 54.78; 7. a) 3.2;
b) 26.4; 2.06; 8. a) 23.5; b) 92.75 and 9.63; c) 75.5; 834.75; 28.89.

3.4. Jointly distributed discrete random variable

Although the probability distributions studied so far have involved only one random variable, many decisions are based upon an analysis of two or more random variables. In problem situations that involve two or more random variables, the resulting probability distribution is referred to as a **joint probability distribution**.

Example:

The number of between-meal snacks eaten by students in a day during final examinations week depends on the number of tests a student had to take on that day. The accompanying table shows joint probabilities, estimated from a survey.

Table 3.8

Number of snacks (Y)	Number of tests (X)			$P(y)$
	0	1	2	
0	0.05	0.08	0.09	0.22
1	0.07	0.09	0.11	0.27
2	0.11	0.04	0.10	0.25
3	0.08	0.07	0.11	0.26
$P(x)$	0.31	0.28	0.41	1.00

Definition:

Let X and Y be a pair of discrete random variables. Their joint probability function expresses the probability that simultaneously X takes the specific value x and Y takes the value y , as a function of x and y .

The notation used is $P_{X,Y}(x,y)$ so,

$$P_{X,Y}(x,y) = P(X = x \cap Y = y)$$

For example, $P_{X,Y}(2,3) = 0.11$. It means that, the probability that randomly chosen student has 2 tests and eats 3 snacks is 0.11.

Definition:

Let X and Y be a pair of jointly distributed random variables. The probability function of the random variable X is called its **marginal** probability function, denoted by $P_X(x)$, and is obtained by summing the joint probabilities over all possible values; that is

$$P_X(x) = \sum_y P(x,y).$$

Similarly, the **marginal** probability function of the random variable Y is

$$P_Y(y) = \sum_x P(x, y).$$

Marginal probability functions $P_X(x)$ and $P_Y(y)$ are shown in the lower row and the right column of the table 3.8.

For example, $P_X(x=0)=0.31$ expresses the probability that, randomly chosen student has no tests is 0.31.

$P_Y(y=2)=0.25$, expresses the probability that randomly chosen student eats 2 snacks is 0.25.

Properties of joint probability functions of discrete random variables

Let X and Y be a discrete random variables with joint probability function $P_{X,Y}(x, y)$. Then

1. $P_{X,Y}(x, y) \geq 0$ for any pairs of x and y .
2. The sum of the joint probabilities $P_{X,Y}(x, y)$ over all possible pairs of values must be 1.

3.4.1. Conditional probability function

Let X and Y be a pair of jointly distributed discrete random variables. The **conditional** probability function of the random variable Y , given that the random variable X takes the value x , expresses the probability that Y takes the value y , as a function of y , when the value x is specified for X .

This is denoted by $P_{Y/X}(y/x)$, and defined as

$$P_{Y/X}(y/x) = \frac{P_{X,Y}(x, y)}{P_X(x)}$$

Similarly, the conditional probability function of X , given $Y=y$ is

$$P_{X/Y}(x/y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}.$$

For example, using the table 3.8, we can compute the conditional probability of $y=2$, given that $x=1$ as

$$P_{Y/X}(2/1) = \frac{P_{Y,X}(2,1)}{P_X(1)} = \frac{0.04}{0.28} = \frac{1}{7}$$

It means, the probability that randomly chosen student who has 1 test eats 2 snacks is 1/7.

The probability of $x=2$, given that $y=3$ is

$$P_{X/Y}(2/3) = \frac{P_{X,Y}(2,3)}{P_Y(3)} = \frac{0.11}{0.26} = \frac{11}{26}$$

It means, the probability that randomly chosen student who eats 3 snacks has 2 tests is 11/26.

3.4.2. Independence of jointly distributed random variables

Definition:

Let X and Y be a pair of jointly distributed discrete random variables. They are said to be **independent** if and only if their joint probability function is the product of their marginal probability functions:

$$P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y)$$

for all possible pairs of values x and y . Otherwise they are said to be dependent.

As an example, from table 3.8, let $x=1$, $y=2$.

Then

$$P_{X,Y}(x,y) = P(1,2) = 0.04; \quad P_X(1) = 0.28; \quad P_Y(2) = 0.25.$$

$$0.04 \neq 0.28 \cdot 0.25,$$

so number of eaten snacks and number of tests are not independent.

3.4.3. Expected value of the function of jointly distributed random variables

Let X and Y be a pair of discrete random variables with probability function $P_{X,Y}(x,y)$.

The mean of random variable X is

$$\mu_X = E(X) = \sum_x x \cdot P(x)$$

The mean of random variable Y is

$$\mu_Y = E(Y) = \sum_y y \cdot P(y)$$

The mean, or expectation of any function $g(X, Y)$ of the random variables X and Y is defined as:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) \cdot P(x, y).$$

As an example let us calculate means of X , Y , and $g(X, Y)$ for the Example above.

The mean of X is:

$$\mu_X = E(X) = \sum_x x \cdot P(x) = 0 \cdot 0.31 + 1 \cdot 0.28 + 2 \cdot 0.41 = 0.28 + 0.82 = 1.1.$$

It means, on average we expect that each student eats 1.1 snacks per day during final examination week.

The mean of Y is:

$$\mu_Y = E(Y) = \sum_y y \cdot P(y) =$$

$$= 0 \cdot 0.22 + 1 \cdot 0.27 + 2 \cdot 0.25 + 3 \cdot 0.26 = 0.27 + 0.5 + 0.78 = 1.55$$

It means, on average, we expect that each student has 1.55 tests per day during final examination week.

3.4.4. Covariance

Suppose that X and Y are pair of random variables and they are dependent. We use covariance to measure the nature and strength of the relationship between them.

Definition:

Let X be a random variable with mean μ_X , and let Y be a random variable with mean μ_Y . The expected value of $(X - \mu_X)(Y - \mu_Y)$ is called the covariance between X and Y , denoted $Cov(X, Y)$, defined as

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) \cdot P(x, y).$$

An equivalent expression for $Cov(X, Y)$ is:

$$Cov(X, Y) = E(XY) - \mu_X \cdot \mu_Y = \sum_x \sum_y x \cdot y \cdot P(x, y) - \mu_X \cdot \mu_Y .$$

If $Cov(X, Y)$ is a positive, then there is a positive linear association between X and Y , if $Cov(X, Y)$ is a negative value, then there is a negative linear association between X and Y . An expectation of 0 for $Cov(X, Y)$ would imply an absence of linear association between X and Y .

Let us calculate $Cov(X, Y)$ for probability distribution shown in the table 3.8.

Using an equivalent expression for $Cov(X, Y)$ yields:

$$\begin{aligned} Cov(X, Y) &= \sum_x \sum_y x \cdot y \cdot P(x, y) - \mu_X \cdot \mu_Y \\ \sum_x \sum_y x \cdot y \cdot P(x, y) &= 0 \cdot 0 \cdot 0.05 + 0 \cdot 1 \cdot 0.08 + 0 \cdot 2 \cdot 0.09 + \\ + 1 \cdot 0 \cdot 0.07 + 1 \cdot 1 \cdot 0.09 + 1 \cdot 2 \cdot 0.11 + 2 \cdot 0 \cdot 0.11 + 2 \cdot 1 \cdot 0.04 + 2 \cdot 2 \cdot 0.10 + \\ + 3 \cdot 0 \cdot 0.08 + 3 \cdot 1 \cdot 0.07 + 3 \cdot 2 \cdot 0.11 &= 0.09 + 0.22 + 0.08 + \\ + 0.40 + 0.21 + 0.66 &= 1.66 \\ Cov(X, Y) &= \sum_x \sum_y x \cdot y \cdot P(x, y) - \mu_X \cdot \mu_Y = \\ &= 1.66 - 1.1 \cdot 1.55 = 1.66 - 1.705 = -0.045 \end{aligned}$$

It means that there is a weak negative association between number of tests taken a day during a final examination week and number of eaten snacks.

Exercises

- 1.** Shown below is the joint probability distribution for two random variables X and Y .

X	Y		\sum
	5	10	
10	0.12	0.08	0.20
20	0.30	0.20	0.50
30	0.18	0.12	0.30
\sum	0.60	0.40	1.00

- a) Find $P_{X,Y}(10,10)$, $P_{X,Y}(30,5)$, and $P_{X,Y}(20,5)$.
- b) Specify the marginal probability distributions for X and Y .
- c) Compute the mean and variance for X and Y .
- d) Are X and Y independent random variables? Justify your answer.
- 2.** There is a relationship between the number of lines in a newspaper advertisement for an apartment and the volume of interest from the potential renters. Let volume of interest be denoted by the random variable X , with the value 0 for little interest, 1 for moderate interest, and 2 for heavy interest. Let Y be the number of lines in a newspaper. Their joint probabilities are shown in the table

Number of lines (Y)	Volume of interest (X)		
	0	1	2
3	0.09	0.14	0.07
4	0.07	0.23	0.16
5	0.03	0.10	0.11

- a) Find and interpret $P_{X,Y}(2,4)$.
- b) Find the joint cumulative probability function at $X=2$, $Y=4$, and interpret your result.
- c) Find and interpret the conditional probability function for Y , given $X=0$.
- d) Find and interpret the conditional probability function for X , given $Y=4$.
- e) If the randomly selected advertisement contains 5 lines, what is the probability that it has heavy interest from the potential renters?
- f) Find expected number of volume of interest.
- g) Find and interpret covariance between X and Y .
- h) Are the number of lines in the advertisement and volume of interest independent of one another?
- 3.** Students at a university were classified according to the years at the university (X) and number of visits to a museum in the last year. ($Y=0$ for no visits, 1 for one visit, 2 for two visits, 3 for more than two visits). The accompanying table shows joint probabilities.

Number of visits (Y)	Years at the university (X)			
	1	2	3	4
0	0.06	0.08	0.07	0.02
1	0.08	0.07	0.06	0.01
2	0.05	0.05	0.12	0.02
3	0.03	0.06	0.18	0.04

- a) Find and interpret $P_{X,Y}(4,3)$
- b) Find and interpret the mean number of X .
- c) Find and interpret the mean number of Y .
- d) If the randomly selected student is a 2^{nd} year student, what is the probability that he or she) visits museum at least 3 times?
- e) If the randomly selected student has 1 visit to a museum, what is the probability that he (or she) is a 3^{rd} year student?
- f) Are number of visits to a museum and years at the university independent of each other?

4. It was found that 20% of all people both watched the show regularly and could correctly identify the advertised product. Also, 27% of all people regularly watched the show and 53% of all people could correctly identify the advertised product. Define a pair of random variables as follows:

$$\begin{array}{ll} X=1 \text{ if regularly watch the show; } & X=0 \text{ otherwise} \\ Y=1 \text{ if product correctly identified; } & Y=0 \text{ otherwise.} \end{array}$$

- a) Find the joint probability function of X and Y .
- b) Find the conditional probability function of Y , given $X=0$.
- c) If randomly selected person could identify the product correctly, what is the probability that he (or she) regularly watch the show?
- d) Find and interpret the covariance between X and Y .

Answers

- 1.** a) 0.08; 0.18; 0.30; b) $P_X(10)=0.20$; $P_X(20)=0.50$; $P_X(30)=0.30$;
 $P_Y(5)=0.60$; $P_Y(10)=0.40$; c) $\mu_X = 21$; $\sigma_X^2 = 49$; $\mu_Y = 7$; $\sigma_Y^2 = 6$; d) Yes;
- 2.** a) 0.16; b) 0.76; c) $P(3/0)=9/19$; $P(4/0)=7/19$; $P(5/0)=3/19$;
d) $P(0/4)=7/46$; $P(1/4)=1/2$; $P(2/4)=8/23$; e) 11/24; f) 1.15;
g) 0.109; h) No; **3.** a) 0.04; b) 2.39; c) 1.63; d) 3/13; e) 3/11; f) No;
- 4.** a) $P_{X,Y}(0,0)=0.40$; $P_{X,Y}(0,1)=0.33$; $P_{X,Y}(1,0)=0.07$; $P_{X,Y}(1,1)=0.20$;

b) $P_{Y/X}(0/0) = 40/73$; $P_{Y/X}(1/0) = 33/73$; c) 20/53; d) 0.057.

3.5. The binomial distribution

An experiment that satisfies the following four conditions is called a binomial experiment:

1. There are n identical trials. In other words, the given experiment is repeated n times. All these repetitions are performed under identical conditions.
2. Each trial has two and only two outcomes. These outcomes are usually called a success and a failure.
3. The probability of success is denoted by p and that of failure by q , and $p + q = 1$. The probabilities p and q remain constant for each trial.
4. The trials are independent. In other words, the outcome of one trial does not affect the outcome of another trial.

It is important to note that one of the two outcomes of a trial is called a **success** and the other a **failure**. Note that a success does not mean that the corresponding outcome is considered favorable or desirable. Similarly, a failure does not necessarily refer to an unfavorable or undesirable outcome. Success and failure simply the names used to denote the two possible outcomes of a trial.

The random variable x that represents the number of successes in n trial for a binomial experiment is called a binomial random variable.

Binomial formula:

For a binomial experiment, the probability of exactly x successes in n trials is given by the binomial formula:

$$P(x) = C_x^n \cdot p^x \cdot q^{n-x}$$

where

n = total number of trials

p = probability of success

$q = 1 - p$ = probability of failure

x = number of success in n trials

$n - x$ = number of failures in n trials.

To find the probability of x successes in n trials for a binomial experiment, the only values needed are those of n and q . These are called the parameters of the binomial distribution or simply the binomial parameters.

Example:

A certain drug is effective in 30 per cent of the cases in which it has been prescribed. If a doctor is now administering this drug to four patients, what is the probability that it will be effective for at least three of the patients?

Solution:

We can consider the administration of the drug to each patient as a trial. Thus, this experiment has four trials. There are only two outcomes for each trial: the drug is effective or the drug is not effective. The event “effective for at least three” can be broken down into two mutually exclusive events (outcomes), “effective for three or effective for four”. If we use the term “success” instead of “effective” we can say that

$$P(\text{at least 3 successes}) = P(3 \text{ successes or } 4 \text{ successes}) = \\ = P(3 \text{ successes}) + P(4 \text{ successes}) = P(x=3) + P(x=4).$$

Now we can find $P(x=3)$ and $P(x=4)$ separately. Since the drug is effective in 30% of the cases, we say that the probability that the drug is effective in any single case is $p=0.3$.

Hence, $q=1-p=0.7$, then the equation of the particular binomial distribution is

$$P(x) = C_x^4 \cdot 0.3^x \cdot 0.7^{4-x}$$

Hence, we have

$$P(x=3) + P(x=4) = C_3^4 \cdot 0.3^3 \cdot 0.7^1 + C_4^4 \cdot 0.3^4 \cdot 0.7^0 = 0.0837.$$

Practically interpreted, this number means that if the drug is administrated to 10 000 sets of four patients, in about 837 of the 10 000 sets will be effective for at least three patients out of four.

Example:

It is known from past data that despite all efforts, 2% of the packages mailed through post office do not arrive at their destinations within the specified time. A corporation mailed 10 packages through post office.

- a) Find the probability that exactly one of these 10 packages will not arrive at this destination within the specified time.
 b) Find the probability that at most one of these 10 packages will not arrive at this destination within the specified time.

Solution:

Let us call it a success if a package does not arrive at its destination within the specified time and a failure if it does arrive within the specified time.
 Then

$$n=10; \quad p=0.02; \quad q=0.98$$

- a) For this part,

$$x = \text{number of successes}=1$$

$$n-x = \text{number of failures}=10-1=9$$

Substituting all values in the binomial formula, we obtain:

$$P(x=1) = C_1^{10} \cdot (0.02)^1 \cdot (0.98)^9 = 0.1667.$$

Thus, there is a 0.1667 probability that exactly one of the 10 packages mailed will not arrive at its destination within the specified time.

- b) The probability that at most one of the 10 packages will not arrive at its destination within the specified time is given by the sum of the probabilities of $x=0$ and $x=1$. Thus,

$$\begin{aligned} P(x \leq 1) &= P(x=0) + P(x=1) = C_0^{10} \cdot (0.02)^0 \cdot (0.98)^{10} + \\ &\quad + C_1^{10} \cdot (0.02)^1 \cdot (0.98)^9 = 0.8171 + 0.1667 = 0.9838. \end{aligned}$$

Thus, the probability that at most one of the 10 packages will not arrive at its destination within the specified time is 0.9838.

3.5.1. Mean and standard deviation of the binomial distribution

The mean and standard deviation for a binomial distribution are

$$\mu = n \cdot p \text{ and } \sigma = \sqrt{n \cdot p \cdot q},$$

where

n -is the total number of trials,

p -is the probability of success, and

q -is the probability of failure.

Example:

The probability that a certain rifleman will get a hit on any given shot at the rifle range is $\frac{3}{10}$. If he fires one hundred shots, find the theoretical mean and standard deviation of x , the number of hits.

Solution:

We have $n=100$, $p=\frac{3}{10}$ and $q=\frac{7}{10}$.

Then, by formula, the mean is

$$\mu = n \cdot p = 100 \cdot \frac{3}{10} = 30$$

and standard deviation is

$$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{100 \cdot \frac{3}{10} \cdot \frac{7}{10}} = \sqrt{21} = 4.583.$$

Exercises

1. Let x be a discrete random variable that possesses a binomial distribution. Using binomial formula, find the following probabilities:

- a) $P(x=5)$ for $n=8$ and $p=0.60$
- b) $P(x=3)$ for $n=4$ and $p=0.30$
- c) $P(x=2)$ for $n=6$ and $p=0.20$

2. Determine the probability of getting:

- a) exactly three heads in 6 tosses of a fair coin;
- b) at least 3 heads in 6 tosses of a fair coin.

3. A card is drawn from an ordinary pack of playing cards, and its suit (clubs, diamonds, hearts, spades) noted, then it is replaced, the pack is shuffled, and another card is drawn. This is done until four cards have been drawn.

- a) What is the probability that two spades will be drawn in four draws?
- b) What is the probability that at least two spades will be drawn in four draws?
- c) What is the probability that two red cards will be drawn in four draws?

d) What is the probability that at most two red cards will be drawn in four draws?

4. At a particular university it has been found that 20% of the students withdraw without completing the business statistics course. Assume that 20 students have registered for the course.

a) What is the probability that two or fewer will withdraw?

b) What is the probability that exactly four will withdraw?

c) What is the probability that more than three will withdraw?

d) What is the expected number and standard deviation of withdrawals?

5. For the binomial distribution with $n=4$ and $p=0.25$ find the probability of

a) three or more successes

b) at most three successes

c) two or more failures.

6. Calculate the mean and standard deviation of the binomial distribution with

a) $n=16$; $p=0.5$

b) $n=25$; $p=0.1$

c) $n=25$; $p=0.9$

7. 19% of cars in the country were at least 12 years old. Find the probability that in a random sample of 10 cars

a) exactly 4 are at least 12 years old;

b) exactly 2 are at least 12 years old;

c) none are at least 12 years old;

d) exactly 5 are at least 12 years old.

8. Suppose that, for a particular type of a cancer, treatment provides a 5-or more years survival rate of 80% if the disease could be detected at an early stage. Among 18 patients diagnosed to have this form of cancer at an early stage who are just starting this treatment, find the probability that

a) fourteen will survive beyond 5-years;

b) six will die within 5-years;

c) the number of patients surviving beyond 5-years will be between 9 and 13(inclusive);

d) find the expectation and standard deviation of the number of 5-years survivors.

9. It is known that 3% of produced goods have some defects. Eight of these goods are selected randomly.

a) What is the probability that none of these goods are defective?

- b) What is the probability that one of these goods is defective?
c) What is the probability that at least two of these goods are defective?
- 10.** A certain type of infection is spread by contact with an infected person. Let the probability that a healthy person gets the infection, in one contact, be $p=0.4$.
- a) An infected person has contact with five healthy persons. Specify the distribution of $X = \text{number of persons who contact the infection}$.
b) Find $P[X \leq 3]$; $P[X = 0]$; and $E[X]$.
- 11.** A salesman of home computers will contact four customers during a week. Each contact can result in either a sale, with probability 0.20, or no sale with probability 0.80. Assume that customer contacts are independent. Let X denotes the number of computers sold during the week.
- a) Obtain the probability distribution of X .
b) Calculate the expected value of X .

Answers

- 1.** a) 0.279; b) 0.076; c) 0.246; **2.** a) $5/16$; b) $21/32$; **3.** a) 0.211; b) 0.262;
c) 0.375; d) 0.688; **4.** a) 0.2060; b) 0.2182; c) 0.5886; d) 4; 1.790;
5. a) 0.051; b) 0.996; c) 0.949; **6.** a) 8;2; b) 2.5; 1.5; c) 22.5; 1.5;
7. a) 0.0773; b) 0.3010; c) 0.1216; d) 0.0218; **8.** a) 0.215; b) 0.082; c) 0.283;
d) 14.4; 1.697; **9.** a) 0.784; b) 0.194; c) 0.022; **10.** a) binomial distribution
with $n=5$; $p=0.4$; b) 0.913; 0.078; 2; **11.** a) $P(0)=0.4096$; $P(1)=0.4096$;
 $P(2)=0.1536$; $P(3)=0.0256$; $P(4)=0.0016$; b) 0.8;

3.6. The hypergeometric probability distribution

In previous section, we have learned that one of the conditions required to apply the binomial probability distribution is that the trials are independent so that the probabilities of the two outcomes (success and failure) remains constant. If the trials are not independent, we can not apply the binomial probability distribution to find probability of x successes in n trials. In such cases we replace the binomial distribution by the **hypergeometric probability distribution**. Such a case occurs when a sample is drawn without replacement from a finite population.

Definition:

Let

N = total number of elements in the population

S = number of successes in the population

$N - S$ = number of failures in the population

n = number of trials (sample size)

x = number of successes in n trials

$n - x$ = number of failures in n trials.

The probability of x successes in n trials is given by

$$P(x) = \frac{C_x^S \cdot C_{n-x}^{N-S}}{C_n^N} = \frac{\frac{S!}{x!(S-x)!} \cdot \frac{(N-S)!}{(n-x)!(N-S-n+x)!}}{\frac{N!}{n!(N-n)!}}$$

Example:

A company has 12 employees who hold managerial positions. Of them, 7 are females and 5 are males. The company is planning to send 3 of these 12 managers to a conference. If 3 managers are randomly selected out of 12,

- find the probability that all 3 of them are female
- find the probability that at most 1 of them is a female.

Solution:

Let the selection of a female be called a success and the selection of a male be called a failure.

- From the given information,

$$N = \text{total number of managers in the population} = 12$$

S = number of successes (females) in the population=7

$N-S$ =number of failures (males) in the population=5

n = number of selections (sample size) =3

x = number of successes (females) in three selections =3 $n-x$ = number of failures (males) in three selections =0.

Using the hypergeometric formula, we find

$$P(x=3)=\frac{C_x^S \cdot C_{n-x}^{N-S}}{C_n^N}=\frac{C_3^7 \cdot C_0^5}{C_3^{12}}=\frac{35 \cdot 1}{220}=0.1591$$

Thus, the probability that all three of the selected managers are female is 0.1591.

b) The probability that at most one of them is a female is given by the sum of the probabilities that either none or one of the selected managers is a female.

To find the probability that none of the selected managers is a female:

$N=12$

$S=7$

$N-S=5$

$n=3$

$x=0$

$n-x=3$

$$P(x=0)=\frac{C_0^7 \cdot C_3^5}{C_3^{12}}=\frac{1 \cdot 110}{220}=0.0455$$

To find the probability that one of the selected managers is a female:

$N=12$

$S=7$

$N-S=5$

$n=3$

$x=1$

$n-x=2$

$$P(x=1)=\frac{C_1^7 \cdot C_2^5}{C_3^{12}}=\frac{7 \cdot 10}{220}=0.3182$$

In the end,

$$P(x \leq 1)=P(x=0)+P(x=1)=0.0455+0.3182=0.3637.$$

Exercises

1. Let $N=14$, $S=6$, and $n=5$. Using the hypergeometric probability distribution formula, find

a) $P(x=4)$; b) $P(x=5)$; c) $P(x \leq 1)$

2. Let $N=16$, $S=10$, and $n=5$. Using the hypergeometric probability distribution formula, find

a) $P(x=5)$; b) $P(x=0)$; c) $P(x \leq 1)$

3. There are 25 goods, 5 of which are defective. We randomly select 4 goods. What is the probability that three of those four goods are nondefective and one is defective?

4. A committee of two members is to be formed from the list of 8 candidates. Of the 8 candidates, 5 are management and 3 are economics department students. Find the probability that

- a) both candidates are managers
- b) neither of the candidates are managers
- c) at most one of the candidates is a manager.

5. A company buys keyboards from another company. The keyboards are received in shipments of 100 boxes, each box containing 20 keyboards. The quality control department first randomly selects one box from each shipment, and then randomly selects five keyboards from that box. The shipment is accepted if not more than one of the five keyboards is defective. The quality control inspector selects a box from a recently received shipment of keyboards. Unknown to the inspector, this box contains six defective keyboards.

- a) What is the probability that this shipment will be accepted?
- b) What is the probability that this shipment will not be accepted?

Answers

1. a) 0.0599; b) 0.0030; c) 0.2378; 2. a) 0.0577; b) 0.0014; c) 0.0357;

3. 0.4506; 4. a) 0.3571; b) 0.1071; c) 0.6429; 5. a) 0.5165; b) 0.4835.

3.7.The Poisson probability distribution

The Poisson probability distribution, named after the French mathematician Siemon D. Poisson, is another important probability distribution of a discrete random variable that has a large number of applications.

A Poisson probability distribution is modeled according to certain assumptions:

- 1.** x is a discrete random variable;
- 2.** The occurrences are random.
- 3.** The occurrences are independent.

In the Poisson probability distribution terminology, the average number of occurrences in an interval is denoted by λ (Greek letter lambda). The actual number of occurrences in that interval is denoted by x .

Poisson probability distribution formula:

According to the Poisson probability distribution, the probability of x occurrences in an interval is

$$P(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

where

$P(x)$ = the probability of x successes over an interval;

λ = is the mean number of occurrences in that interval

$e = 2.71828$ (the base of natural logarithms)

The **mean** and **variance** of the Poisson probability distribution are:

$$\mu_x = E(X) = \lambda \quad \text{and} \quad \sigma_x^2 = E[(X - \mu_x)^2] = \lambda .$$

Remark: As it is obvious from the Poisson probability distribution formula, we need to know only the value λ to compute the probability of any given value of x . We can read the value of $e^{-\lambda}$ for a given λ from Table1 of the Appendix.

Example:

A computer breaks down at an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this computer will have

- a) exactly three breakdowns;
- b) at most one breakdown.

Solution:

Let λ be the mean number of breakdowns per month and x be the actual number of breakdowns observed during the next month for this computer.

Then $\lambda=3$.

a) The probability that exactly three breakdowns will be observed during the next month is

$$P(x=3) = \frac{3^3 \cdot e^{-3}}{3!} = \frac{27 \cdot 0.049787}{6} = 0.224$$

b) The probability that at most one breakdown will be observed during the next month is given by the sum of the probabilities of zero and one breakdown. Then

$$P(\text{at most one breakdown}) = P(0 \text{ or } 1 \text{ breakdown}) = P(x=0) + P(x=1) =$$

$$= \frac{3^0 \cdot e^{-3}}{0!} + \frac{3^1 \cdot e^{-3}}{1!} = 0.049787 + 3 \cdot 0.049787 = 0.1991.$$

Example:

A car salesperson sells an average of 0.9 cars per day. Find the probability of selling

- a) exactly 2
- b) at least 3 cars per day
- c) find the mean, variance and standard deviation of selling cars per day.

Solution:

Let λ be the mean number of cars sold per day by this salesperson.

Let x be the number of cars sold by this salesperson. Hence, $\lambda=0.9$

$$\text{a) } P(x=2) = \frac{0.9^2 \cdot e^{-0.9}}{2!} = \frac{0.81 \cdot 0.406570}{2} = 0.1647.$$

$$\begin{aligned}\text{b) } P(\text{at least 3 cars sold}) &= P(x=3) + P(x=4) + \dots = \\ &= 1 - P(x=0) - P(x=1) - P(x=2) = 1 - 0.0406570 - 0.9 \cdot 0.406570 - \\ &\quad - 0.1647 = 0.0628.\end{aligned}$$

$$\text{c) } \mu = \lambda = 0.9$$

$$\sigma^2 = \lambda = 0.9 \quad \text{and} \quad \sigma = \sqrt{\lambda} = \sqrt{0.9} = 0.949.$$

Exercises

1. Using the Poisson formula, find the following probabilities

- a) $P(x < 2)$ for $\lambda = 3$
- b) $P(x = 8)$ for $\lambda = 5.3$

2. Let x be a Poisson random variable. Using the Poisson probabilities table, write the probability distribution of x for each of the following. Find the mean and standard deviation for each of these probability distributions.

- a) $\lambda = 0.6$;
- b) $\lambda = 1.8$

3. An average of 7.5 crimes are reported per day to police in a city. Use the Poisson formula to find the probability that

- a) exactly 3 crimes will be reported to a police on a certain day
- b) at least 2 crimes will be reported to a police on a certain day.

4. A mail-order company receives an average 1.3 complaints per day. Find the probability that it will receive

- a) exactly 3 complaints
- b) 2 to 3 complaints
- c) more than 3 complaints
- d) less than 3 complaints on a certain day.

5. An average of 4.5 customers come to the bank per half hour.

- a) Find the probability that exactly 2 customers will come to this bank during a given hour;
- b) Find the probability that during a given hour, the number of customers who will come to the bank is at most 2.

6. An average of 0.6 accidents occur per month at a large company.

- a) Find the probability that no accident will occur at this company during a given month.
- b) Find the mean, variance, and standard deviation of the number of accidents that will occur at this company during a given month.

Answers

1. a) 0.1991; b) 0.0771; 2. a) $\mu = 0.6$; $\sigma = 0.7746$ b) $\mu = 1.8$; $\sigma = 1.3416$

3. a) 0.03888; b) 0.9953; 4. a) 0.0998; b) 0.3301; c) 0.0431; d) 0.8569;

5. a) 0.0050; b) 0.0062; 6. a) 0.5488; b) $\mu = 0.6$; $\sigma^2 = 0.6$; $\sigma = 0.7746$.

Chapter 4

Continuous random variables and their probability distributions

4.1. Introduction

Up to this point, we have limited our discussion to probability distributions of discrete random variables. Recall that a discrete random variable takes on only some isolated values, usually integers representing a count. We now turn our attention to the probability distribution of a continuous random variable- one that can ideally assume any value in an interval. Variables measured on an underlying continuous scale, such as weight, strength, life length, and temperature, have this feature.

Figure 4.1 displays the histogram and polygon for some continuous data set. The smoothed polygon is an approximation of the probability distribution curve of the continuous random variable X . The probability distribution curve of a continuous random variable is also called its probability density function.

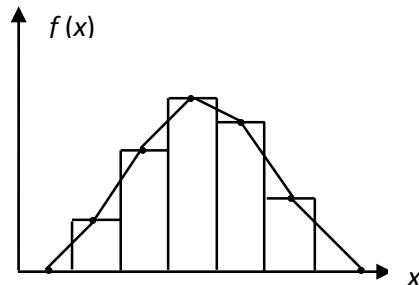


Fig.4.1. Histogram and polygon

The probability density function, denoted by $f(x)$ possesses the following characteristics:

1. $f(x) > 0$ for all x .
2. The area under the probability density function $f(x)$ over all possible values of the random variable X is equal to 1.

3. Let a and b be two possible values of the random variable X , with $a < b$. Then the probability that X lies between a and b is the area under the density function between a and b . (Fig.4.2)

4. The cumulative distribution

function $F(x_0)$ is the area under the probability density function $f(x)$ up to x_0

$$F(x_0) = \int_{x_a}^{x_0} f(x) dx$$

where x_a is the minimum value of the random variable X .

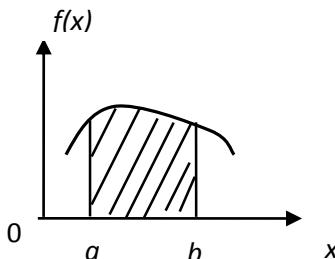


Fig.4.2. Shaded area is the probability that X lies between a and b .

4.2. Areas under continuous probability density functions

Let X be a continuous random variable with probability density function $f(x)$ and cumulative distribution function $F(x)$. Then:

1. The total area under the curve $f(x)$ is 1.

2. The area under the curve $f(x)$ to the left of x_0 is $F(x_0)$, where x_0 is any value that the random variable X can take.

The area under the probability distribution curve of a continuous random variable between any two points is between 0 and 1, as shown in Figure 4.3.

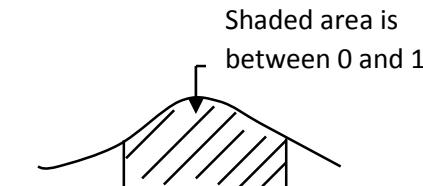


Fig.4.3. Area under a curve between two points

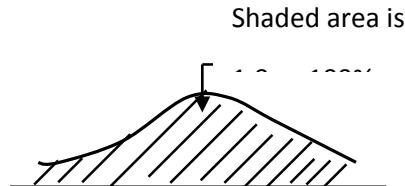


Fig.4.4. Total area under a probability distribution curve

The total area under the probability distribution curve of a continuous random variable is always 1.0 or 100% as shown in Figure 4.4.

Remark:

The probability that a continuous random variable x assumes a single value is always zero.

This is because the area of a line, which represents a single point, is zero. (Fig.4.5)

In general, if a and b are two of the values that X can assume, then,

$$P(a) = 0 \text{ and } P(b) : P(x = 2) = 0$$

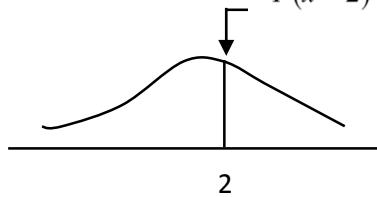


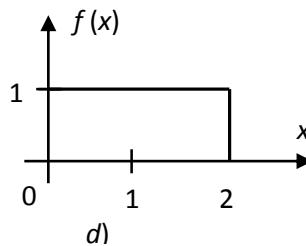
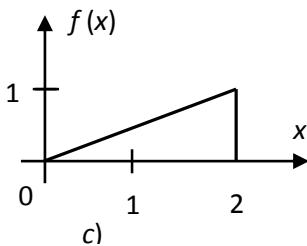
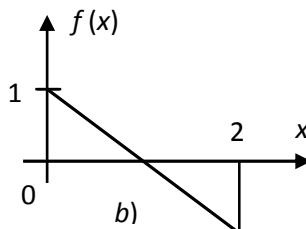
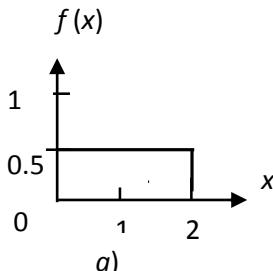
Fig.4.5 Probability of a single value of x is zero

When determining the probability of an interval a to b , we need not be concerned if either or both end points are included in the interval. Since the probabilities of $X = a$ and $X = b$ are both equal to 0,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Exercises

1. Which of the functions sketched in *a-d* could be a probability density function for a continuous random variable? Why or why not?



2. Determine the following probabilities from the curve $f(x)$ diagrammed in exercise 1(a).

- | | |
|---------------------|---------------------|
| a) $P(0 < X < 0.5)$ | b) $P(0.5 < X < 1)$ |
| c) $P(1.5 < X < 2)$ | d) $P(X = 1)$ |

3. For the curve $f(x)$ graphed in exercise 1(c) which of the two intervals $(0 < X < 0.5)$ or $(1.5 < X < 2)$ is assigned a higher probability?

4. The time it takes for a TV repair master to finish his job (in hours) has a density function of the form

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

- a) Determine the constant c .

b) What is the probability that a TV repair master will finish the job in less than 75 minutes? Between $1\frac{1}{2}$ and 2 hours?

5. Suppose that the loss in a certain investment, in thousands of dollars, is a continuous random variable X that has a density function of the form

$$f(x) = \begin{cases} k(2x - 3x^2) & \text{if } -1 < x < 0 \\ 0 & \text{otherwise} \end{cases}$$

a) Calculate the value of k .

b) Find the probability that the loss is at most \$500.

6. Let the random variable X has probability density function

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

a) Draw the probability density function

b) Show that the density function has the properties of a proper probability density function

c) Find the probability that X takes a value between 0.5 and 1.5.

Answers

2. a) 0.25; b) 0.25; c) 0.25; d) 0; **3.** The interval 1.5 to 2 has higher probability; **4.** a) -6; b) 5/32 ; 1/2; **5.** a) -1/2; b) 3/16; **6.** c) 0.75.

4.3. Joint Probability Density Functions

Let us consider the joint probability density function of two continuous random variables X and Y . From previous study we know that a single random variable X is called continuous if there exists a nonnegative real-valued function $f(x) : R \rightarrow [0, +\infty)$ such that for any subset A of real numbers that can be constructed from intervals by a countable number of set operations,

$$P(X \subset A) = \int_A f(x) dx$$

And we can generalize to the case when more than one variable is involved.

Definition:

Two random variables X and Y, defined on the same sample space, have a continuous joint distribution if there exists a nonnegative function of two variables, $f(x, y)$ on RxR $f(x, y)$, such that for any region R in the xy-plane that can be formed from rectangles by a countable number of set operations,

$$f(x, y) \subset R = \iint_R f(x, y) dx dy$$

Where A and B are any subsets of real numbers that can be constructed from intervals by a countable number of set operations. Then we obtain

$$P(X \subset A, Y \subset B) = \int_B \int_A f(x, y) dx dy$$

If we put $A = (-\infty, +\infty)$, $B = (-\infty, +\infty)$, we will get

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

which is the continuous analog of one variable case.

$$P(X = a, Y = b) = \int_b^b \int_a^a f(x, y) dx dy = 0$$

Therefore, for $a < b$ and $c < d$

$$P(a < X < b, c < Y < d) = P(a \leq X < b, c \leq Y < d) = \dots = \int_c^d \int_a^b f(x, y) dx dy$$

Let X and Y have a joint probability density function $f(x, y)$. Let f_Y be the probability density function of Y. To find f_Y in terms of f , for any subset B of R $P(Y \subset B) = \int_B f_Y(y) dy$ (1)

From the another side

$$P(Y \subset B) = P(X \subset (-\infty, \infty), Y \subset B) = \int_B \left(\int_{-\infty}^{\infty} f(x, y) dx \right) dy \quad (2)$$

Comparing (1) and (2) we obtain

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (3)$$

Similarly

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (4)$$

$f_X(x)$ and $f_Y(y)$ are called, respectively, the marginal probability density functions of X and Y.

Marginal probability cumulative distribution functions of X and Y are defined as respectively

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

$$F_Y(y) = \int_{-\infty}^y F_Y(t) dt$$

These relations show that if X and Y have joint probability density function $f(x, y)$, then X and Y are continuous random variables with density functions f_X and f_Y , and distribution functions F_X and F_Y , respectively. Therefore

$$F'_X(x) = f_X(x) \text{ and } F'_Y(y) = f_Y(y)$$

Expected values of random variables X and Y are defined as

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x, y) dx \quad \text{and}$$

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f(x, y) dy$$

Example

The joint probability density function of random variables X and Y is given by

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

- a) Determine the value of C
- b) Find the marginal probability density functions of X and Y .
- c) Calculate $E(X)$ and $E(Y)$.

Solution:

a) To find C , note that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$, gives

$$\int_0^2 \left(\int_x^2 C \cdot x^2 y dy \right) dx = 1$$

Therefore

$$1 = \int_0^2 \left(\int_x^2 C \cdot x^2 y dy \right) dx = \int_0^2 C x^2 \left(\int_x^2 y dy \right) dx = \int_0^2 C x^2 \left(2 - \frac{x^2}{2} \right) dx = C \cdot \frac{32}{15} = 1$$

and hence $C = \frac{15}{32}$

- b) To find f_X and f_Y , the respective marginal probability density functions of X and Y , we use (3), (4)

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_x^2 \frac{15}{32} x^2 y dy = \frac{15}{32} x^2 \left(2 - \frac{x^2}{2} \right) \quad \text{if } 0 < x < 2$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^y \frac{15}{32} x^2 y dx = \frac{5y^4}{32} \quad \text{if } 0 < y < 2$$

- c) To find $E(X)$ and $E(Y)$, we use the results obtained in part (b):

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x, y) dx = \int_0^2 x \cdot \frac{15}{32} x^2 \left(2 - \frac{x^2}{2}\right) dx = -\frac{5}{4}$$

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f(x, y) dy = \int_0^2 y \cdot \frac{5y^4}{32} dy = \frac{5}{3}$$

Let X and Y be independent continuous random variables. Then for all real-valued functions $g : R \rightarrow R$ and $h : R \rightarrow R$

$$E[g(X) \cdot h(X)] = E[g(X)] \cdot E[h(X)]$$

where, as usual, we assume that $E[g(X)]$ and $E[h(X)]$ are finite

Independence of Continuous Random Variables

Let X and Y be jointly continuous random variables with joint probability density function $f(x, y)$. Then X and Y are independent if and only if $f(x, y)$ is the product of their marginal density functions $f_x(x)$ and $f_y(y)$

$$f(x, y) = f_x(x) \cdot f_y(y) \quad (5)$$

Otherwise, it is said, that X and Y are dependent.

Example

Let the joint probability density function of random variables X and Y be given by

$$f(x, y) = \begin{cases} 4 \cdot xy & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Are X and Y independent? Why or why not?

Solution:

To check the validity of (5), we first calculate $f_x(x)$ and $f_y(y)$

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_x^1 4xy dy = 2x(1 - x^2) \quad \text{if } 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^y 4xy dx = 2y^3 \quad \text{if } 0 < y < 1$$

Since $f(x, y) \neq f_X(x) \cdot f_Y(y)$, X and Y are dependent.

Exercises

1. The joint probability density function of random variables X and Y is given by $f(x, y) = \begin{cases} k \cdot x^2 \cdot y & \text{if } 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$

- (a) Determine the value of k .
- (b) Find the marginal probability density functions of X and Y .
- (c) Calculate $E(X)$ and $E(Y)$.

2. Let the joint probability density function of random variables X and Y be given by $f(x, y) = \begin{cases} \frac{1}{2} \cdot y \cdot e^{-x} & \text{if } x \geq 0, 0 \leq y \leq 3 \\ 0 & \text{otherwise} \end{cases}$

- a) Find the marginal probability density functions of X and Y
- b) Find $F_X(x)$, $F_Y(y)$, $E(X)$, $E(Y)$

3. Let the joint probability density function of random variables X and Y be given by $f(x, y) = \begin{cases} 6 \cdot x \cdot y & \text{if } 0 \leq y \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$

- a) Calculate the marginal probability density functions of X and Y , respectively.
- b) Calculate $E(X)$ and $E(Y)$.

4. Let the joint probability density function of random variables X and Y be given by $f(x, y) = \begin{cases} 3 & \text{if } 0 \leq y \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Are X and Y independent? Why or why not?

5. Let the joint probability density function of random variables X and Y be given by $f(x, y) = \begin{cases} x^3 \cdot e^{-x(y+2)} & \text{if } x \geq 0, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Are X and Y independent? Why or why not?

6. Let the conditional probability density function of X given that $Y = y$ be given by

$$f(x, y) = \begin{cases} \frac{3(x^2 + y^2)}{3y^2 + 1} & \text{if } 1 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $P(1/3 \leq x \leq 1/2 / Y = 1/4)$

7. Let X and Y be continuous random variables with joint probability density function

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

Calculate $f_{X/Y}(x/y)$

Answers

1. 2.; 3.; 4. 5.; 6. 7. $\frac{2(x+y)}{2y+3}$

4.3. The normal distribution

The normal distribution is one of the many probability distributions that a continuous random variable can possess. The normal distribution is the most important and most widely used of all the probability distributions. A large number of phenomena in the real world are normally distributed either exactly or approximately.

The probability density function for a normally distributed random variable X is:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-(x-\mu)^2/(2\sigma^2)} \quad \text{for } -\infty < x < \infty$$

where μ and σ^2 are any number such that $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$, $e=2.71828\dots$ and $\pi=3.14159\dots$ are constants.

A normal probability distribution, when plotted, gives a bell-shaped curve such that

1. The total area under the curve is 1.0.
 2. The curve is symmetric about the mean.
 3. The two tails of the curve extend indefinitely.
1. The total area under a normal curve is 1.0 or 100%, as shown in Figure 4.6.

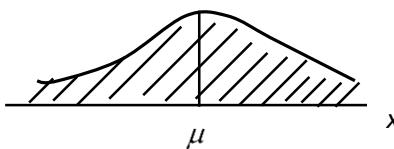


Fig.4.6. T
rea under

a normal curve

2. A normal curve is symmetric about the mean, as shown in Figure 4.7. Consequently 0.5 of the total area under a normal curve lies on the left side of the mean and 0.5 lies on the right side of the mean.

Each of the two shaded areas is 0.5

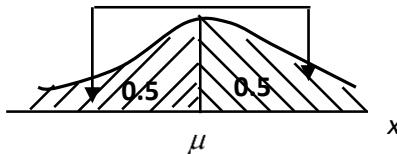


Fig.4.7. A normal curve is symmetric about the mean

3. The tails of a normal distribution curve extend indefinitely in both directions without touching or crossing the horizontal axis.

Each of the two shaded areas is very close to zero

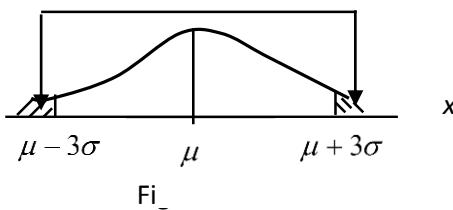


Fig. 4.8

Although a normal curve never meets the horizontal axis, beyond the points represented by $\mu - 3\sigma$ and $\mu + 3\sigma$ it becomes so close to this axis that the area under the curve beyond these points in both directions can be taken as virtually zero. These areas are shown in Figure 4.8.

Remark:

There is not just one normal distribution curve but rather a family of normal distribution curves. Each different set of values of μ and σ gives a different normal distribution.

The value of μ determines the center of a normal distribution on the horizontal axis. The three distribution curves drawn in Figure 4.9 have the same mean but different standard deviations.

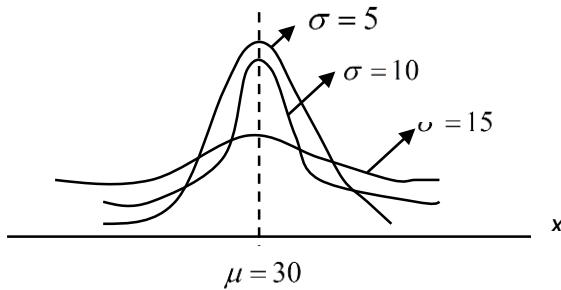


Fig.4.9. Three normal distribution curves with the same mean but different standard deviations

The value of σ gives the spread of the normal distribution curve. The three normal distribution curves in Figure 4.10 have different means but the same standard deviation.

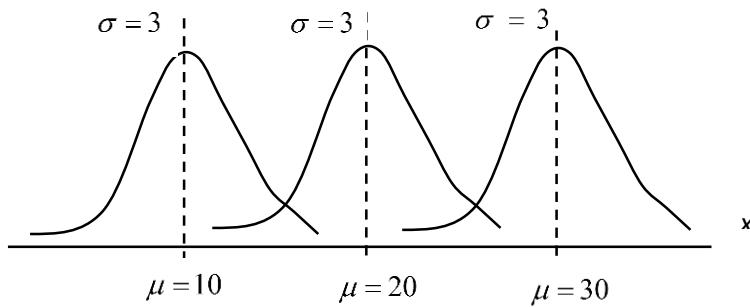


Fig.4.10. Three normal distribution curves with different means but the same standard deviation.

Properties of the normal distribution:

Suppose that the random variable X follows a normal distribution. Then the following properties hold:

1. The mean of the random variable is μ :

2. The variance of the random variable is σ^2

$$Var(X) = E(X - \mu)^2 = \sigma^2$$

3. By knowing the mean and standard deviation (or variance) we can define

the normal distribution by using the notation:

$$X \sim N(\mu, \sigma^2)$$

4.3.1. Cumulative distribution function of the normal distribution

Suppose that X is a normal random variable with mean μ , and variance σ^2 , that is $X \sim N(\mu, \sigma^2)$.

Then the cumulative distribution function

$$F(x_0) = P(X \leq x_0)$$

This is the area under the normal probability density function to the left of x_0 as illustrated in Figure 4.11.

As for any proper density function, the total area under the curve is 1, that is

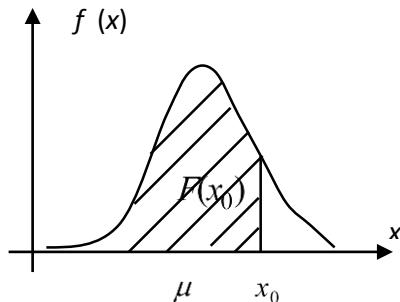


Fig.4.11

$$F(-\infty) = 0 \text{ and } F(\infty) = 1.$$

Let a and b be two possible values of X , with $a < b$.

Then

$$P(a < X < b) = F(b) - F(a)$$

The probability is the area under corresponding density function between a and b as shown in Figure 4.12.

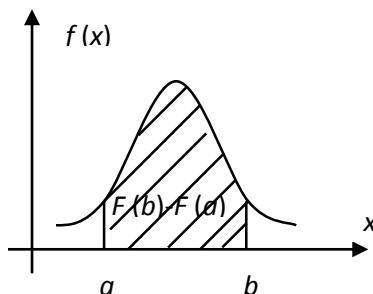


Fig.4.12. Shaded area indicates the probability that X lies between a and b

4.4. The standard normal distribution

The **standard normal distribution** is a special case of the normal distribution. The particular normal distribution that has a mean of 0 and a standard deviation of 1 is called the standard normal distribution. It is customary to denote the standard normal variable by Z .

Definition:

The standard normal distribution has a bell-shaped density with

$$\text{mean} = \mu = 0$$

$$\text{standard deviation} = \sigma = 1$$

The standard normal distribution is denoted by $Z \sim N(0, 1)$. (Fig.4.13)

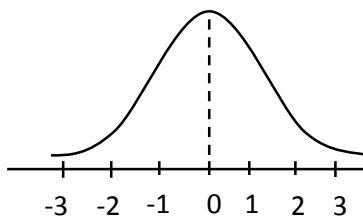


Fig.4.13. The standard normal curve

We will denote the cumulative distribution function of Z by $F(z)$, and for two numbers a and b with $a < b$

$$P(a < Z < b) = F(b) - F(a)$$

Now let us see the procedure for finding probabilities associated with a continuous random variable. We wish to determine the probability of a random variable having a value in a specified interval from a to b . Thus we have to find the area under the curve in the interval from a to b . Finding areas under the standard normal distribution curve appears at first glance to be much

more difficult. The mathematical technique for obtaining these areas is beyond the scope of the text, but fortunately tables are available which provide the areas or probability values for the standard normal distribution. The cumulative distribution function of the standard normal distribution is tabulated in Table 2 in the Appendix. This table gives values of

$$F(z) = P(Z \leq z)$$

for nonnegative values of z .

For example the cumulative probability for a Z value of 1.13 from Table 2

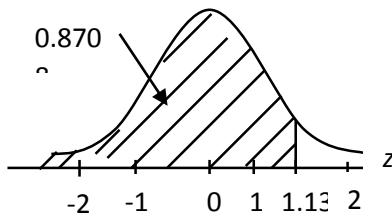


Fig.4.14

$$F(1.13) = 0.8708$$

This is the area, shown in Figure 4.14, for Z less than 1.13.

Because of the symmetry of the normal distribution the probability that $Z > -1.13$ is also equal to 0.8708.

In general, values of the cumulative distribution function for negative values of z can be inferred using the symmetry of the probability density function. To find the cumulative probability for a negative z (for example $Z = -2.25$) defined as

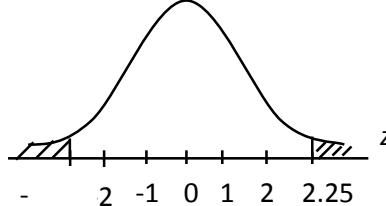


Fig.4.15

$$F(-z_0) = P(Z \leq -z_0) = F(-2.25)$$

We use the complement of the probability for $Z = -2.25$ as shown in Figure 4.15. From the symmetry we can see that

$$F(-2.25) = 1 - P(Z \leq 2.25) = 1 - F(2.25) = 1 - 0.9878 = 0.0122$$

We can see that the area under the curve to the left of $Z = -2.25$ is equal to the area to the right of $Z = 2.25$ because of the symmetry of the normal distribution.

Example: Find $P(Z > -1.35)$

Solution:

We see that because of the symmetry the probability or area to the right of -1.35 (Fig.4.16) is the same as the area to the left of 1.35 (Fig.4.17.)

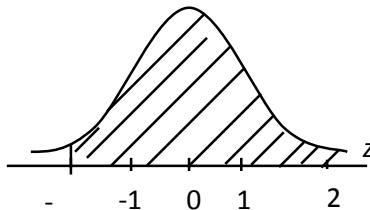


Fig.4.16

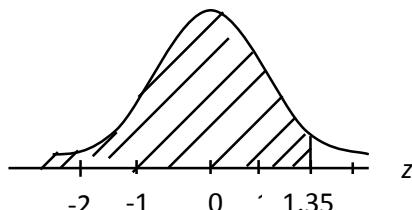


Fig.4.17

So,

$$P(z > -1.35) = P(z < 1.35) = F(1.35) = 0.9115.$$

Example:

Let the random variable Z follow a standard normal distribution. The probability is 0.25 that Z is greater than what number?

Solution:

We need to find such a point z_0 that $P(Z > z_0) = 0.25$. (Fig.4.18)

Area to the left of z_0 is $1 - 0.25 = 0.75$.

So,

$$F(z_0) = 0.75 \text{ and } z_0 = 0.675.$$

$$P(z > 0.675) = 0.25.$$

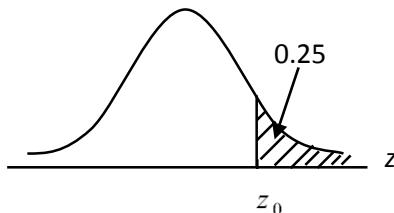


Fig.4.18

Exercises

1. Find the area under the standard normal curve to the left of

- | | |
|------------------|-----------------|
| a) $z = 1.17$; | b) $z = 0.16$; |
| c) $z = -1.83$; | d) $z = -2.3$ |

2. Find the area under the standard normal curve to the right of

- | | |
|------------------|-----------------|
| a) $z = 1.17$; | b) $z = 0.60$; |
| c) $z = -1.13$; | d) $z = 1.55$; |

3. Find the area under the standard normal curve over the interval

- | | |
|--------------------------------|--|
| a) $z = -0.65$ to $z = 0.65$; | b) $z = -1.04$ to $z = 1.04$; |
| c) $z = 0.32$ to $z = 2.65$; | d) $z = -0.755$ to $z = 1.254$ (interpolate) |

4. For the random variable Z follows a standard distribution, find

- | | |
|-----------------------------|-------------------------|
| a) $P(Z < 0.42)$; | b) $P(Z < -0.42)$; |
| c) $P(Z > 1.69)$; | d) $P(Z > -1.69)$; |
| e) $P(-1.2 < Z < 2.1)$; | f) $P(0.5 < Z < 0.8)$; |
| g) $P(-1.62 < Z < -0.51)$; | h) $P(Z < 1.64)$; |
| i) $P(-2.34 < Z < -1.09)$ | |

5. Find the z -value in each of the following cases:

- | | |
|----------------------------|-----------------------------|
| a) $P(Z < z) = 0.1736$; | b) $P(Z > z) = 0.20$; |
| c) $P(-z < Z < z) = 0.8$; | d) $P(-0.6 < Z < z) = 0.50$ |

6. Let the random variable Z follow a standard normal distribution.

- The probability is 0.80 that Z is less than what number?
- The probability is 0.35 that Z is less than what number?
- The probability is 0.3 that Z is greater than what number?

d) The probability is 0.75 that Z is greater than what number?

Answers

1. a) 0.8790; b) 0.5636; c) 0.0336; d) 0.0107; 2. a) 0.1210; b) 0.2743; c) 0.8708; d) 0.0606; 3. a) 0.4844; b) 0.7016; c) 0.3705; d) 0.6700; 4. a) 0.6628; b) 0.3372; c) 0.0455; d) 0.9545; e) 0.8670; f) 0.0966; g) 0.2524; h) 0.9495; i) 0.1283; 5. a) -0.94; b) 0.84; c) 1.28 and -1.28; d) 0.755; 6. a) 0.842; b) -0.386; c) 0.524; d) -0.675.

4.5. Standardizing a normal distribution

Fortunately, no new tables are required for probability calculations regarding the general normal distributions. Any normal distribution can be converted to the standard normal by the following relation:

Rule:

Let X be a normally distributed random variable with mean μ and variance σ^2 . Then random variable

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

It follows that if a and b are any numbers with $a < b$, then

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = F\left(\frac{b - \mu}{\sigma}\right) - F\left(\frac{a - \mu}{\sigma}\right),$$

where Z is the standard normal random variable and $F(z)$ denotes its cumulative distribution function.

Example:

If $X \sim N(12,4)$ find the probability that X is greater than 16.

Solution:

The probability that X assumes a value greater than 16 is given by the area under the area under the normal curve to the right of $x = 16$, as shown in Figure 4.19.

For $x = 16$:

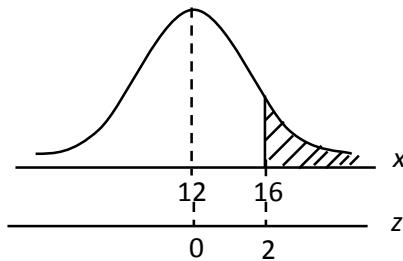


Fig.4.19. Finding $P(X > 16)$

$$z = \frac{X - \mu}{\sigma} = \frac{16 - 12}{2} = 2.00.$$

The required probability is given by the area to the right of $z = 2.00$.

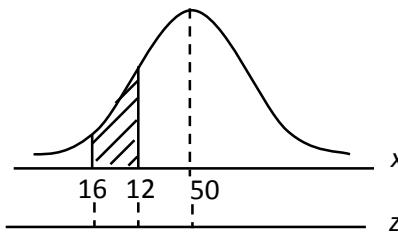
$$P(X > 16) = P(Z > \frac{16 - 12}{2}) = P(Z > 2) = 1 - F(2.00) = 1 - 0.9772 = 0.0228.$$

Example:

Let X be a continuous random variable that has a normal distribution with $\mu = 50$ and $\sigma = 8$. Find the probability $P(30 \leq X \leq 39)$.

Solution:

The probability $P(30 \leq X \leq 39)$ is given by the area from $x = 30$ to $x = 39$ under the normal curve.(Fig.4.20)



The z -values to $x = 30$ and $x = 39$

Fig.4.20

corresponding

are

$$\text{For } x = 30; z = \frac{30 - 50}{8} = -2.50;$$

$$\text{For } x = 39; z = \frac{39 - 50}{8} = -1.38.$$

We calculate:

$$P(30 \leq X \leq 39) = P(-2.50 \leq Z \leq -1.38) = 0.9938 - 0.9162 = 0.0776.$$

Example:

The number of calories in a salad on the lunch menu is normally distributed with mean $\mu = 200$ and standard deviation $\sigma = 5$. Find the probability that the salad you select will contain

- a) more than 208 calories;
- b) between 190 and 200 calories.

Solution:

Letting X denote the number of calories in the salad, we have the standardized variable

$$Z = \frac{X - 200}{5}$$

$$\begin{aligned} \text{a) } P(X > 208) &= P(Z > \frac{208 - 200}{5}) = P(Z > 1.6) = 1 - P(Z \leq 1.6) = \\ &= 1 - 0.9452 = 0.0548. \end{aligned}$$

$$\begin{aligned}
 b) P(190 \leq X \leq 200) &= P\left(\frac{190-200}{5} \leq Z \leq \frac{200-200}{5}\right) = \\
 &= P(-2.0 \leq Z \leq 0) = F(0) - F(-2.0) = 0.5 - (1 - F(2.0)) = \\
 &= 0.5 - 1 + F(2.0) = 0.5 - 1 + 0.9772 = 0.4772.
 \end{aligned}$$

Exercises

1. If $X \sim N(80, 16)$, find

- a) $P(X < 75)$;
- b) $P(X > 86)$;
- c) $P(73 \leq X \leq 89)$;
- d) $P(81 < X < 84)$

2. If X is normally distributed with a mean of 30 and a standard deviation of 5, find

- a) $P(X < 26)$;
- b) $P(X \geq 37)$;
- c) $P(20 \leq X \leq 40)$;
- d) $P(20 < X < 40)$

3. If X has a normal distribution with $\mu = 150$ and $\sigma = 5$, find b such that

- a) $P(X < b) = 0.9920$;
- b) $P(X > b) = 0.0197$;
- c) $P(X < b) = 0.063$

4. Scores on a university entrance examination follow a normal distribution with a mean of 500 and a standard deviation of 100. Find the probability that a student will score

- a) over 650;
- b) less than 250;
- c) between 325 and 675;

d) If the university admits students who score over 670, what proportion of the student pool would be eligible for admission?

e) What should be the limit if only the top 15% are to be eligible?

5. According to the children's growth chart that doctors use as a reference, the heights of two-year-old boys are nearly normally distributed with a mean of 85 cm inches and a standard deviation of 5 cm. If a two year-old boy is selected at random, what is the probability that he will be between 75 cm and 92 cm tall?

6. The weights of apples served at a restaurant are normally distributed with a mean of 125 grams and standard deviation of 8 grams. What is the

probability that the next person served will be given an apple that weights less than 120 grams?

7. The National bank is reviewing its service charge and interest-paying policies on checking accounts. The bank has found that the average daily balance on personal checking accounts is \$55000, with a standard deviation of \$15000. In addition, the average daily balances have been found to be normally distributed.

- a) What percentage of personal checking account customers carry average daily balances in excess of \$80000?
- b) What percentage of the bank's customers carry daily balances below \$20000?
- c) What percentage of the bank's customer carry average daily balances between \$30000 and \$70000?
- d) The bank is considering paying interest to customers carrying average daily balances in excess of a certain amount. If the bank does not want to pay interest to more than 5% of its customers, what is the minimum average daily balance it should be willing to pay interest on?

8. The sales of high-bright toothpaste are believed to be approximately normally distributed, with a mean of 10 000 tubes per week and a standard deviation of 1500 tubes per week.

- a) What is the probability that more than 12000 tubes will be sold in any given week?
- b) In order to have a 0.95 probability that the company will have sufficient stock to cover the weekly demand, how many tubes should be produced?

9. The attendance at football games at a certain stadium is normally distributed, with a mean of 45000 and a standard deviation of 3000.

- a) What percentage of the time should attendance be between 44000 and 48000?
- b) What is the probability of exceeding 50000?

c) Eighty percent of the time the attendance should be at least how many?

10. The lifetime of a color television picture tube is normally distributed, with a mean of 7.8 years and a standard deviation of 2 years.

- a) What is the probability that a picture tube will last more than 10 years?
- b) If the firm guarantees the picture tube for 2 years, what percentage of the television sets sold will have to be replaced because of picture tube failure?

c) If the firm is willing to replace the picture tubes in a maximum of 1% of the television set sold, what guarantee period can be offered for the television picture tubes?

11. It is estimated that the scores on the university entrance test are distributed normally with mean of 80 and standard deviation of 5.

a) For a randomly chosen participant taking this test, what is the probability of a score more than 72?

b) For a randomly chosen participant taking this test, what is the probability of a score between 73 and 85?

c) What is the minimum score needed in order to be in the top 5% of all participants taking this test?

d) What is the minimum score is needed to enter to the university if only the best 70% of all participants will pass this test?

e) Two participants are chosen at random. What is the probability that at least one of them scores more than 85?

Answers

- 1.** a) 0.1056; b) 0.0668; c) 0.9477; d) 0.2426; **2.** a) 0.2119; b) 0.0808;
c) 0.9544; d) 0.9544; **3.** a) 162.05; b) 160.3; c) 142.35; **4.** a) 0.0668;
b) 0.0062; c) 0.9198; d) 4.46%; e) 603.6; **5.** 0.8964; **6.** 0.266; **7.** a) 4.75% ;
b) 0.99%; c) 79.38%; d) \$79675; **8.** a) 0.0918; b) 12468; **9.** a) 0.4706;
b) 0.0475; c) 42450; **10.** a) 0.1357; b) 0.19%; c) 3.14 years or approximately
3 years; **11.** a) 0.9452; b) 0.7605; c) 88.225; d) 77.375; e) 0.2922.

4.6. The normal distribution approximation to the binomial distribution

Whenever the number of trials in a binomial experiment is small it is easy to find probabilities of the various values of x , the number of successes, by using formula

$$P(x) = C_x^n \cdot p^x \cdot q^{n-x}$$

As the number of trials increases, however, the effort involved in answering questions about probabilities associated with the experiment quickly becomes laborious.

For instance, suppose that we want to know the probability that in fifteen tosses of a fair coin we toss at least nine heads. You will undoubtedly agree that $n = 15$ is not a large number of trials. However, in order to find $P(x \geq 9)$ we say

$$P(x \geq 9) = P(x=9 \text{ or } 10 \text{ or } 11 \text{ or } 12 \text{ or } 13 \text{ or } 14 \text{ or } 15) =$$

$$= C_9^{15} \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^6 + \dots + C_{15}^{15} \left(\frac{1}{2}\right)^{15}$$

So we have seven probabilities to compute, after which we must perform the addition. This is not practically difficult, but it takes a fair amount of time.

To find only one of these probabilities, for example, we have

$$P(x=10) = C_{10}^{15} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^5 ;$$

$$C_{10}^{15} = \frac{15!}{10!5!} = \frac{11 \cdot 12 \cdot 13 \cdot 14 \cdot 15}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 3003$$

$$\left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^5 = \left(\frac{1}{2}\right)^{15} = \frac{1}{32768}.$$

Therefore

$$C_{10}^{15} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^5 = 3003 \cdot \frac{1}{32768} = 0.092.$$

Thus, you see that if we were to calculate other six such probabilities we would expend a considerable amount of time and energy.

In such cases, the normal distribution can be used to approximate the binomial probability. Note that, for a binomial problem, the exact probability is obtained by using the binomial formula. If we apply the normal distribution to solve a binomial problem, the probability that we obtain is an approximation to the exact probability.

Example:

According to an estimate, 50% of the people have at least one credit card. If a random sample of 30 persons is taken, what is the probability 19 of them will have at least one credit card?

Solution:

Let n be the total number of persons in the sample, x be the number of persons in the sample who have at least one credit card, and p be the probability that a person has at least one credit card. Then, this is a binomial problem with

$$\begin{aligned} n &= 30; & p &= 0.50; & q &= 1 - p = 0.50; \\ x &= 19; & \text{and} & & n - x &= 30 - 19 = 11. \end{aligned}$$

Using the binomial formula, the exact probability that 19 persons in a sample of 30 have at least one credit card is

$$P(x=19) = C_{19}^{30} \cdot 0.5^{19} \cdot 0.5^{11} = 0.0509$$

Now let us solve this problem using the normal distribution as an approximation to the binomial distribution. For this example,

$$n \cdot p = 30 \cdot 0.5 = 15 \text{ and}$$

$$n \cdot p \cdot q = 7.5.$$

Using the normal distribution as an approximation to the binomial involves the following steps:

Step1:

Compute μ and σ for the binomial distribution.

To use the normal distribution, we need to know the mean and standard deviation of the distribution. Hence, the first step in using the normal approximation to the binomial distribution is to compute the mean and standard deviation of the binomial distribution. As we know the mean and standard deviation of the binomial distribution are given by

$$\mu = n \cdot p \text{ and } \sigma = \sqrt{n \cdot p \cdot q}.$$

Using these formulas, we obtain

$$\mu = n \cdot p = 30 \cdot 0.50 = 15;$$

$$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{30 \cdot 0.5 \cdot 0.5} = 2.74.$$

Step2:

Convert the discrete random variable to a continuous random variable.

The normal distribution applies to a continuous random variable, whereas the binomial distribution applies to a discrete random variable. The second step is to convert the discrete random variable to a continuous random variable by making the **correction for continuity**.

To make the correction for continuity, we use the interval 18.5 to 19.5 for 19 persons.

Step3:

Compute the required probability using the normal distribution.

The area under the normal curve between $x = 18.5$ and $x = 19.5$ will give us the (approximate) probability that 19 persons possess at least one credit card. We calculate this probability as follows:

$$\text{For } x = 18.5; \quad z = \frac{18.5 - 15}{2.74} = 1.28;$$

$$\text{For } x = 19.5; \quad z = \frac{19.5 - 15}{2.74} = 1.64.$$

The required probability is given by the area under the standard normal curve between $z = 1.28$ and $z = 1.64$. (Fig 4.21).

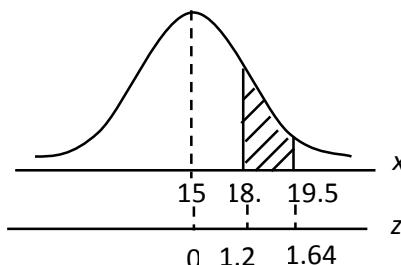


Fig.4.21

The required probability is

$$P(18.5 \leq x \leq 19.5) = P(1.28 \leq z \leq 1.64) = 0.9495 - 0.8997 = 0.0498.$$

Thus, based on the normal approximation, the probability that 19 persons in a sample of 30 will possess at least one credit card is approximately 0.0498. Earlier, using the binomial formula, we obtained the exact probability 0.0509. The error due to using the normal approximation is $0.0509 - 0.0498 = 0.0011$. Thus, the exact probability is underestimated by 0.0011 if the normal approximation is used.

Definition:

Let x be the number of successes from n independent trials, each with probability of success p . Then number of successes, x , is a binomial random variable and if $n \cdot p \cdot q > 9$ (where $q = 1 - p$) a good approximation is

$$P(a \leq x \leq b) = P\left(\frac{a - np}{\sqrt{n \cdot p \cdot q}} \leq Z \leq \frac{b - np}{\sqrt{n \cdot p \cdot q}}\right) \quad (4.1)$$

or if $5 < n \cdot p \cdot q < 9$ we can use the **continuity correction factor** to obtain

$$P(a \leq x \leq b) = P\left(\frac{a - 0.5 - np}{\sqrt{n \cdot p \cdot q}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{n \cdot p \cdot q}}\right) \quad (4.2)$$

where Z is a standard normal random variable.

Example:

Let X have a binomial distribution with $p = 0.6$ and $n = 150$. Approximate the probability that

- a) x lies between 82 and 101;
- b) x is greater than 97.

Solution:

Since $n \cdot p \cdot q = 150 \cdot 0.6 \cdot 0.4 = 36 > 9$, then we will use approximation without using the continuity correction.

Since $n \cdot p = 90$, $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{36} = 6$ we obtain:

$$\begin{aligned} \text{a) } P(82 \leq x \leq 101) &= P\left(\frac{82 - 90}{6} \leq Z \leq \frac{101 - 90}{6}\right) = P(-1.33 \leq Z \leq 1.83) = \\ &= F(1.83) - (1 - F(1.33)) = F(1.83) - 1 + F(1.33) = \\ &= 0.9664 - 1 + 0.9082 = 0.8746. \end{aligned}$$

$$\begin{aligned} \text{b) } P(x > 97) &= P\left(Z > \frac{97 - 90}{6}\right) = P(Z > 1.17) = \\ &= 1 - F(1.17) = 1 - 0.8790 = 0.1210. \end{aligned}$$

Example:

A large-scale survey conducted two years ago revealed that 30% of the adult population were regular users of alcoholic beverages. If this is still the current rate, what is the probability that in a random sample of 40 adults the number of users of alcoholic beverages will be

- a) less than 15
- b) 10 or more?

Solution:

For this example

$$n = 40, \quad p = 0.3; \quad q = 0.7.$$

Since $n \cdot p \cdot q = 40 \cdot 0.3 \cdot 0.7 = 8.4 < 9$ we must use continuity correction factor to obtain necessary probabilities.

$$\text{a) } P(x < 15) = P\left(Z < \frac{15 + 0.5 - 12}{\sqrt{40 \cdot 0.3 \cdot 0.7}}\right) = P\left(Z < \frac{3.5}{2.9}\right) = P(Z < 1.21) = 0.8869.$$

The probability that 15 out of 40 adults use alcoholic beverages regularly is 0.8869.

$$\begin{aligned} \text{b) } P(x > 10) &= P\left(Z > \frac{10 - 0.5 - 12}{\sqrt{40 \cdot 0.3 \cdot 0.7}}\right) = P\left(Z > -\frac{2.5}{2.9}\right) = \\ &= P(Z > -0.86) = F(0.86) = 0.8051. \end{aligned}$$

Exercises

1. For a binomial probability distribution, $n = 80$ and $p = 0.50$. Let x be the number of successes in 80 trials.

- a) Find the mean and standard deviation of the binomial distribution.
- b) Find $P(x \geq 37)$ using the normal approximation.
- c) Find $P(41 \leq x \leq 44)$ using the normal approximation

2. For a binomial probability distribution, $n = 120$ and $p = 0.6$. Let x be the number of successes in 120 trials.

- a) Find the mean and standard deviation of the binomial distribution.
- b) Find $P(x \leq 70)$ using the normal approximation.

c) Find $P(67 \leq x \leq 71)$ using the normal approximation

3. Find the following binomial probabilities using the normal approximation

a) $n = 70; p = 0.30; P(x = 18)$

b) $n = 200; p = 0.70; P(133 \leq x \leq 145)$

c) $n = 40; p = 0.25; P(x \geq 12)$

d) $n = 50; p = 0.10; P(x \leq 7)$

4. According to the statistics, 19% of cars in the city were at least 12 years old in 2004. Assume that this result holds true for the current population of all cars in the city. Random sample of 500 cars are selected at random. Find the probability that

a) exactly 92 cars are at least 12 years old.

b) 100 or more cars are at least 12 years old.

c) 90 to 98 cars are at least 12 years old?

5. According to a survey, 30% of credit card holders pay off their balances in full each month. Assume that this result holds true for the current population of credit card holders.

a) Find the probability that in a random sample of 400 credit card holders, exactly 125 pay off their balances in full each month.

b) Find the probability that in a random sample of 400 credit card holders, at least 110 pay off their balances in full each month.

c) What is the probability that in a random sample of 400 credit card holders, 115 to 130 pay off their balances in full each month?

6. A fast food chain store conducted a taste survey before marketing a new hamburger. The results of the survey showed that 70% of the people who tried this hamburger liked it. Encouraged by this result, the company decided to market the new hamburger. Assume that 70% of all people like this hamburger. On a certain day, 40 customers bought this hamburger.

a) Find the probability that exactly 32 of the 40 customers will like this hamburger.

b) What is the probability that 25 or less of the 40 customers will like this hamburger?

c) What is the probability that 31 to 34 of the 40 customers will like this hamburger?

7. According to a survey, 20.8% of the lawyers and judges are women.

- a) Find the probability that in a random sample of 200 lawyers and judges, exactly 35 are women.
- b) Find the probability that in a random sample of 200 lawyers and judges, at most 45 are women.
- c) What is the probability that in a random sample of 200 lawyers and judges, 43 to 50 are women?
8. Of the customers visiting the stereo section of a large electronic store, only 25% make a purchase. If 45 customers visit the stereo section tomorrow, find the probability that more than 10 will make a purchase.
9. The unemployment rate in a city is 7.9%. A sample of 100 persons is selected from the labor force. Approximate the probability that
- a) less than 11 unemployed persons are in the sample
- b) more than 9 unemployed persons are in the sample.
- c) between 8 and 12 unemployed persons are in the sample.

Answers

1. a) $\mu = 40$; $\sigma = 4.472$; b) 0.7486; c) 0.2262; 2. a) $\mu = 72$; $\sigma = 5.367$; b) 0.3557; c) 0.2485; 3. a) 0.0764; b) 0.6393; c) 0.2912; d) 0.8810; 4. a) 0.0413; b) 0.2843; c) 0.3488; 5. a) 0.0378; b) 0.8621; c) 0.5709; 6. a) 0.0525; b) 0.1949; c) 0.1824; 7. a) 0.0371; b) 0.7224; c) 0.3331; 8. 0.7257; 9. a) 0.9082; b) 0.4129; c) 0.5150.

4.7. The exponential probability distribution

The **exponential probability distribution** is another important probability density function. This probability distribution is closely related to the Poisson probability distribution.

The exponential probability distribution has only one parameter λ , which denotes the average number of occurrences per unit of time.

Remark:

The exponential distribution differs from the normal distribution in two important ways

1. it is restricted to random variables with positive values;
2. its distribution is not symmetric.

Definition:

The exponential random variable X ($x > 0$) has a probability density function

$$f(x) = \lambda \cdot e^{-\lambda x} \text{ for } x > 0$$

where λ is the mean number of occurrences per unit time, x is the number of time units until the next occurrence, and $e = 2.71828\dots$, then X is said to follow an **exponential probability distribution**. It can be shown that λ is the same parameter used for the Poisson distribution and that the mean time between occurrences is $1/\lambda$.

The cumulative distribution function is

$$F(x) = 1 - e^{-\lambda x} \text{ for } x > 0$$

The distribution has mean $1/\lambda$ and variance $1/\lambda^2$.

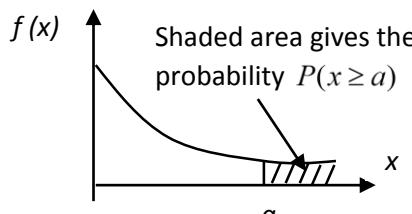


Fig.4.22

The probability $P(x \geq a)$ for the exponential probability distribution is given by the area in the tail of the exponential probability distribution curve beyond $x = a$, as shown in Figure 4.22.

As we know from earlier discussion, for a continuous random variable x , $P(x > a)$ is equal to $P(x \geq a)$. Hence for an exponential probability distribution,

$$P(x \geq a) = P(x > a) = e^{-\lambda a}$$

By using the complementary probability rule, we obtain:

$$P(x \leq a) = 1 - P(x > a) = 1 - e^{-\lambda a}$$

The probability that the x is between two successive occurrences is in the interval “ a ” to “ b ” is

$$\begin{aligned}P(a \leq x \leq b) &= P(x \leq b) - P(x \leq a) = (1 - e^{-\lambda b}) - (1 - e^{-\lambda a}) = \\&= 1 - e^{-\lambda b} - 1 + e^{-\lambda a} = e^{-\lambda a} - e^{-\lambda b}.\end{aligned}$$

Probabilities for the exponential probability distribution.

For the exponential probability distribution with the mean number of occurrences per unit of time equal to λ ,

$$\begin{aligned}P(x \geq a) &= e^{-\lambda a} \\P(x \leq a) &= 1 - e^{-\lambda a} \\P(a \leq x \leq b) &= e^{-\lambda a} - e^{-\lambda b}.\end{aligned}$$

Example:

A processing machine breaks down an average of once in four weeks. What is the probability that the next breakdown will not occur for at least six weeks after the previous breakdown? Assume that the time between breakdowns has an exponential distribution.

Solution:

Let x denote the lapse time between any two successive breakdowns of this machine. We are to find the probability

$$P(x \geq 6 \text{ weeks})$$

Because the unit of time for x is in the weeks, we must define the mean number of breakdowns λ per week. Since there is one breakdown in four weeks,

$$\lambda = \text{the mean number of breakdowns per week} = \frac{1}{4} = 0.25.$$

The required probability is calculated using the formula $P(x \geq a) = e^{-\lambda a}$.

In our example, $\lambda = 0.25$ and $a = 6$ weeks. The required probability is

$$P(x \geq 6 \text{ weeks}) = e^{-0.25 \cdot 6} = e^{-1.5} = 0.2231.$$

The value of $e^{-1.5}$ can be found from the Table 1 of the Appendix.

Example:

A teller at the bank serves, on average, 30 customers per hour. Assume that the service time for a customer has an exponential distribution.

- What is the probability that the next customer will take five minutes or more to be served?
- Find the probability that the next customer will take two minutes or less to be served.
- What is the probability the next customer will take two to four minutes to be served?

Solution:

Let x be the time taken by this teller to serve a customer. We must find the mean number of customers served per minute by this teller to define λ per unit of time (minute). The teller serves on average 30 customers per 60 minutes. Hence

$$\lambda = 30/60 = 0.5 \text{ customers served per minute.}$$

- We need to find the probability $P(x \geq 5)$. (Fig. 4.23.)

In this case $a = 5$.

$$P(x \geq 5) = e^{-\lambda a} =$$

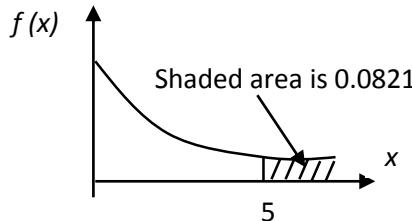


Fig.4.23. The probability for $P(x \geq 5)$

$$e^{-0.5 \cdot 5} = e^{-2.5} = 0.082085 = 0.0821$$

The probability is 0.0821 that a customer will take more than five minutes to be served.

- We are to find $P(x \leq 2)$. This probability will be calculated using the formula $P(x \leq 2) = 1 - e^{-\lambda a}$. In this case $a = 2$ minutes.

So

$$\begin{aligned} P(x \leq 2) &= 1 - P(x > 2) = 1 - e^{-\lambda \cdot a} = 1 - e^{-0.5 \cdot 2} = \\ &= 1 - e^{-1.0} = 1 - 0.367879 = 0.6321. \end{aligned}$$

Thus, the probability is 0.6321 that a customer will be served in two minutes or less.

$$\begin{aligned} \text{c) } P(2 \leq x \leq 4) &= P(x \geq 2) - P(x \geq 4) = e^{-0.5 \cdot 2} - e^{-0.5 \cdot 4} = \\ &= e^{-1} - e^{-2} = 0.367879 - 0.135335 = 0.2325. \end{aligned}$$

Thus, the probability that the teller will take two to four minutes to serve a customer is 0.2325.

Exercises

1. Let x be a continuous random variable that possesses an exponential probability distribution with $\lambda = 1.0$. Find the following probabilities

$$\text{a) } P(x \geq 3); \quad \text{b) } P(x \leq 4); \quad \text{c) } P(2 \leq x \leq 6)$$

2. The life of a pie is exponentially distributed with a mean of three days. What is the probability that a pie that is baked just now will still be good after four days?

3. At the supermarket, a customer has to wait an average of four minutes in line before being served. The time a customer has to wait is exponentially distributed.

- a) What is the probability that a customer will have to wait for more than eight minutes in line?
- b) What is the probability that a customer will have to wait for three to seven minutes in line?

4. On average, 20 telephone calls are received per hour at an office. The time between calls received at this office is exponentially distributed.

- a) What is the probability that no calls will come in during the next 10 minutes?
 - b) What is the probability that the next call will come in within 4 minutes?
5. Aysel works for a toy company and assembles five toys per hour on average. The assembly time for this toy follows an exponential distribution.
- a) Find the probability that the next toy will take more than 15 minutes to assemble.

- b) What is the probability that the next toy will take less than 8 minutes to assemble?
- c) What is the probability that the next toy will take 10 to 16 minutes to assemble?
- 6.** A student, working in a part time job, sells life insurance policies. The past data show that he (she) sells, on an average, 10 life insurance policies per 4-week period. Assume that the time between successive sales of life insurance policies by student has an exponential distribution.
- a) What is the probability that the next life insurance policy will not be sold for two weeks?
- b) What is the probability that the next life insurance policy will be sold within one week?
- c) What is the probability that the next life insurance policy will be sold in one to two weeks?

Answers

- 1.** a) 0.0498; b) 0.9817; c) 0.1329; **2.** 0.2725; **3.** a) 0.1353; b) 0.2986;
4. a) 0.0369; b) 0.7275; **5.** a) 0.2725; b) 0.5034; c) 0.1768; **6.** a) 0.0067;
b) 0.9179; c) 0.0754.

Chapter 5. Sampling distributions

5.1. Sampling and sampling distributions

Suppose that we want to select a sample of n objects from a population of N objects.

A **simple random sample** is selected such that every object has an equal probability of being selected and the objects are selected independently—the selection of one object does not change the probability of selecting any other objects.

It is important that a sample represents the population as a whole. If a marketing manager wants to assess reactions to a new food product, she would not sample only her friends and neighbors. People must be selected randomly and independently. Random selection is our insurance policy against allowing personal influence the selection.

We use sample information to make inferences about the parent population. The distribution of all values of interest in this population can be represented by a random variable. It would be too ambitious to attempt to describe the entire population distribution based on a small random sample of observations. However, we may well be able to make quite firm inferences about important characteristics of the population distribution, such as the population mean and variance.

Sampling distribution:

The probability distribution of \bar{x} is called its sampling distribution. It lists the various values that \bar{x} can assume and the probability for each value of \bar{x} . In general, the probability distribution of a sample statistic is called its **sampling distribution**.

Let us consider sampling distribution in example.

Example:

Suppose that there are only five employees working for a small company. The following data give the annual salaries (in thousands of dollars) of these employees:

17; 24; 35; 35; 43

Let X denote the annual salary of an employee. We can write the frequency distribution of annual salaries as in table 5.1

Table 5.1

Table 5.2

X	f
17	1
24	1
35	2
43	1

Population frequency distribution

X	$P(X)$
17	1/5=0.2
24	1/5=0.2
35	2/5=0.4
43	1/5=0.2

Population probability distribution

Dividing the frequencies of classes by the population size we obtain the relative frequencies, which can be used as probabilities of those classes. Table 5.2, which lists the probabilities of various X values, presents the probability distribution of the population.

Now, let us consider all possible samples of three salaries each, that can be selected, without replacement, from the population. The total number of possible samples, given by the combination

$$\text{Total number of samples} = C_3^5 = \frac{5!}{3!2!} = 10.$$

Suppose we assigns letters A, B, C, D, and E to the salaries of five employees so that

$$A=17; \quad B=24; \quad C=35; \quad D=35; \quad E=43.$$

Then 10 possible samples of three salaries are

ABC,	ABD,	ABE,	ACD,
ADE,	BCD,	BCE,	BDE,
			CDE.

These 10 samples and their respective means are listed in Table 5.3.

Note that the values of means of samples in Table 5.3 are rounded to two decimal places.

Table 5.3

Sample	Salaries in the sample	\bar{X}
ABC	17, 24, 35	25.33
ABD	17, 24, 35	25.33
ABE	17, 24, 43	28.00
ACD	17, 35, 35	29.00
ACE	17, 35, 43	31.67
ADE	17, 35, 43	31.67
BCD	24, 35, 35	31.33
BCE	24, 35, 43	34.00
BDE	24, 35, 43	34.00
CDE	35, 35, 43	37.67

All possible samples

their means when the sample size is 3.

and

By using the values of \bar{X} given in Table 5.3, we record the frequency distribution of \bar{X} in Table 5.4.

Table 5.4

\bar{X}	f
25.33	2
28.00	1
29.00	1
31.33	1
31.67	2
34.00	2
37.67	1

Frequency distribution of \bar{X}
when the sample size is 3.

By dividing the frequencies of various values of \bar{X} by the sum of all frequencies, we obtain the relative frequencies of classes, which can be used as probabilities of classes. These probabilities are listed in Table 5.5.

This table gives the sampling distribution of \bar{X} .

Table 5.5.

\bar{X}	$P(X)$
25.33	$2/10=0.20$
28.00	$1/10=0.10$
29.00	$1/10=0.10$
31.33	$1/10=0.10$
31.67	$2/10=0.20$
34.00	$2/10=0.20$
37.67	$1/10=0.10$
	$\sum P(x) = 1.0$

Sampling distribution of \bar{X}
when the sample size is 3.

If we draw just one sample of three salaries from the population of five salaries, we may draw any of 10 possible samples. Hence, the sample mean

\bar{X} can assume any of the values listed in Table 5.5 with the corresponding probability. This probability function is graphed in Figure 5.1. For example, the probability that the mean of a randomly drawn sample of three salaries is 31.67 is 0.20. This can be written as

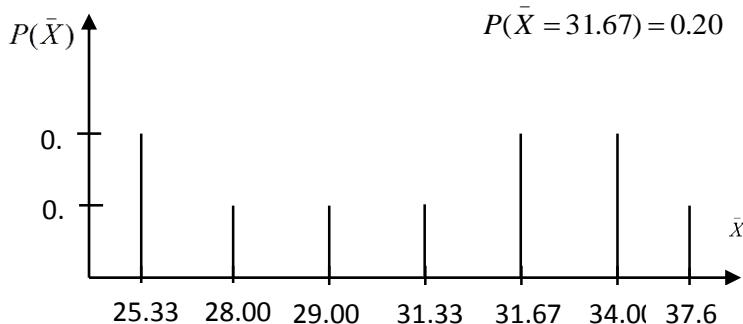


Fig 5.1. Probability function of sampling distribution for means of samples of three observations selected from the population

5.1.1. Mean and standard deviation of \bar{X}

The mean and standard deviation calculated for the sampling distribution of \bar{X} are called the **mean** and **standard deviation** of \bar{X} and denoted by $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ respectively.

Let us calculate the mean of the 10 values of \bar{X} listed in Table 5.3.

$$\begin{aligned}\mu_{\bar{X}} &= \frac{\sum \bar{X}}{10} = \\ &= \frac{25.33 + 25.33 + 28.00 + 29.00 + 31.67 + 31.67 + 31.33 + 34 + 34 + 37.67}{10} = \\ &= \frac{308}{10} = 30.80\end{aligned}$$

Alternatively, we can calculate the mean of the sampling distribution of \bar{X} listed in Table 5.5 as

$$\begin{aligned}\mu_{\bar{X}} &= \\ &= \sum \bar{X} \cdot P(\bar{X}) = 25.33 \cdot 0.2 + 28 \cdot 0.1 + 29 \cdot 0.1 + 31.33 \cdot 0.1 + 31.67 \cdot 0.2 + \\ &+ 34 \cdot 0.2 + 37.67 \cdot 0.1 = 30.80.\end{aligned}$$

Now let us calculate the mean of population: the annual salaries of all five employees:

$$\mu = \frac{17 + 24 + 35 + 35 + 43}{5} = \frac{154}{5} = 30.80$$

The mean of the sampling distribution of \bar{X} always equal to the mean of the population.

Mean of the sampling distribution:

The mean of the sampling distribution of \bar{X} is equal to the mean of the population.

Hence

$$\bar{X} = \mu.$$

Hence, if we take all possible samples (of the same size) from a population and calculate their means, the mean \bar{X} of all these sample means will be the same as the population mean.

The sample mean \bar{X} is called an estimator of the population mean μ . When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population mean, that sample statistic is said to be an **unbiased estimator**. For the sample mean \bar{X} , $\mu_{\bar{X}} = \mu$. Hence, \bar{X} is an unbiased estimator of μ .

Let us talk about population standard deviation and standard deviation of the sampling distribution of \bar{X} .

First, let us find standard deviation of salaries of five employees.

We will use

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

to obtain standard deviation of population.

$$\begin{aligned}\sigma &= \sqrt{\frac{(17-30.8)^2 + (24-30.8)^2 + (35-30.8)^2 + (35-30.8)^2 + (43-30.8)^2}{5}} = \\ &= \sqrt{\frac{190.44 + 46.24 + 17.64 + 17.64 + 144.84}{5}} = \sqrt{\frac{420.8}{5}} = \sqrt{84.16} = 9.174\end{aligned}$$

Now let us calculate the value of $\sigma_{\bar{X}}$.

We will use formula

$$\sigma_{\bar{X}} = \sqrt{\sum (\bar{X})^2 \cdot P(X) - (\mu_{\bar{X}})^2}.$$

We obtain

$$\begin{aligned}\sigma_{\bar{X}} &= \sqrt{\sum (25.33)^2 \cdot 0.2 + (28)^2 \cdot 0.1 + \dots + (37.67)^2 \cdot 0.1 - (30.80)^2} = \\ &= \sqrt{(128.32 + 78.4 + 84.1 + 98.16 + 200.6 + 231.2 + 141.90) - 948.64} = \\ &= \sqrt{962.68 - 948.64} = \sqrt{14.04} = 3.747.\end{aligned}$$

As we see, the standard deviation $\sigma_{\bar{X}}$ of \bar{X} is not equal to the standard deviation σ of the population distribution. The standard deviation of \bar{X} is equal to the standard deviation of the population divided by the square root of the sample size.

That is,

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

We also call $\sigma_{\bar{X}}$ as a standard error of \bar{X} .

We use the above formula for standard error if the sample size is a small in comparison to the population size. The sample size is considered to be small compared to the population size if the sample size is equal to or less than 5% of the population size, that is, if

$$n/N \leq 0.05$$

If this condition does not satisfied, we use the following formula to calculate $\sigma_{\bar{X}}$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

The term $\sqrt{\frac{N-n}{N-1}}$ is often called a **finite population correction factor**.

In most practical applications, the sample size is usually small compared to the population size.

Consequently, in most cases the formula used for calculating $\sigma_{\bar{X}}$ is

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

5.1.2. Central limit theorem

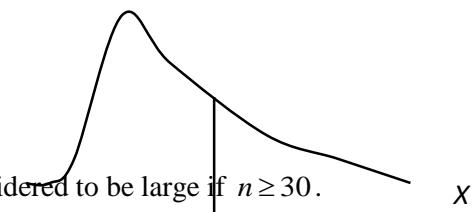
Theorem: Whatever the population, the distribution of \bar{X} is approximately normal when n is large. In random sampling from the population with mean μ and standard deviation σ , when n is large, the distribution of \bar{X} is approximately normal with mean μ and standard deviation $\sigma_{\bar{X}}$. Consequently,

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

is approximately $N \sim (0,1)$

- a) Population that is not normally distributed

The sample size is usually considered to be large if $n \geq 30$.



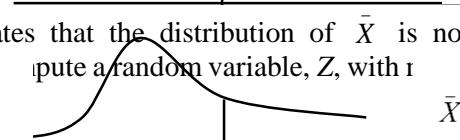
The central limit theorem states that the distribution of \bar{X} is normally distributed if the sample size is sufficiently large.

- b) Sampling distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

- c) Sampling distribution

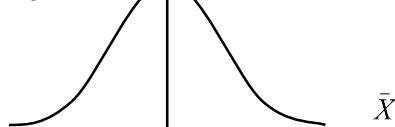
of \bar{V} for $n=20$



for all households in a large city have a mean income of \$30 and a standard deviation of

- d) Sampling distribution

204



deviation equal to \$7. Calculate the mean and standard deviation of \bar{X} and describe the shape of its sampling distribution when the sample size is

- a) $n = 35$; b) $n = 70$

Solution:

Although the population distribution is not normal, in each case the sample size is large ($n > 30$) .

Hence the central limit theorem can be applied.

a) Let \bar{X} be the mean value of telephone bills paid by a sample of 35 households. Then the sampling distribution of \bar{X} is approximately normal with

$$\mu_{\bar{X}} = \mu = \$30 \quad \text{and} \quad \sigma_{\bar{X}} = \sigma / \sqrt{n} = 7 / \sqrt{35} = 1.18.$$

Figure 5.3 shows the population distribution and sampling distribution of \bar{X} .

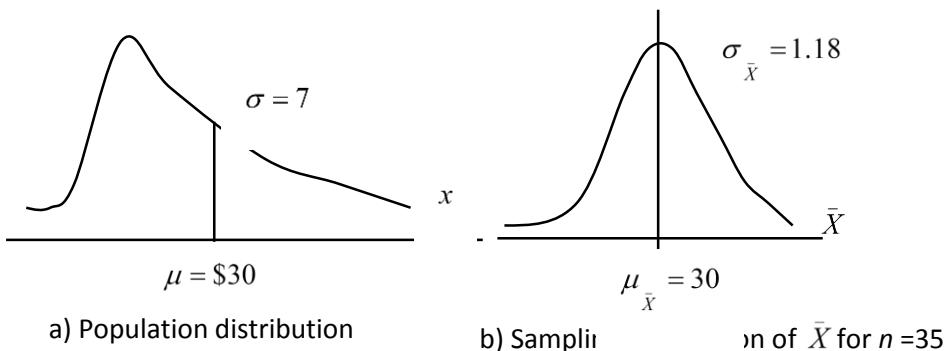


Fig. 5.3.

- b) Let \bar{X} be the mean value of telephone bills paid by a sample of 70 households. Then the sampling distribution of \bar{X} is approximately normal with

$$\mu_{\bar{X}} = \mu = \$30 \quad \text{and} \quad \sigma_{\bar{X}} = \sigma / \sqrt{n} = 7 / \sqrt{70} = 0.84.$$

Figure 5.4 shows the population distribution and sampling distribution of \bar{X} .

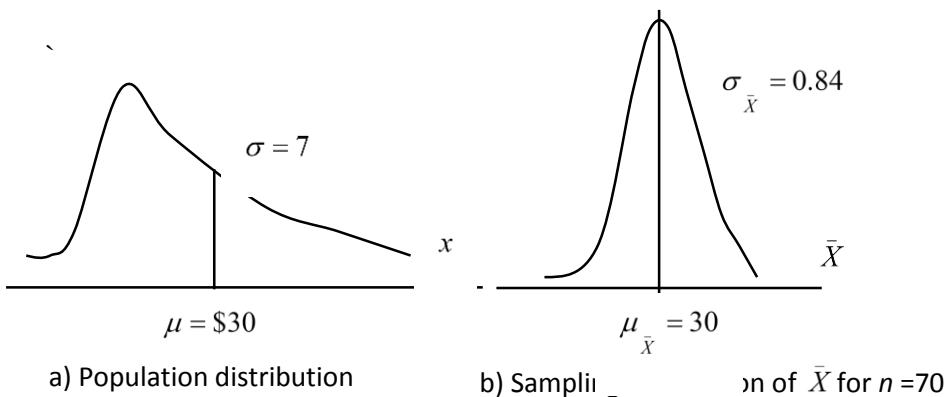


Fig. 5.4.

Example:

Consider a population with mean 75 and standard deviation of 11.

- a) If a random sample of size 64 is selected, what is the probability that the sample mean will be between 73 and 78?
- b) If a random sample of size 80 is selected, what is the probability that the sample mean will be between 68 and 83?

Solution:

We have $\mu = 75$ and $\sigma = 11$. Since $n = 64$ is large, the central limit theorem tells us that the distribution of \bar{X} is approximately normal.

a) To calculate $P(73 \leq \bar{X} \leq 78)$ we convert to the standardized variable

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

The Z-values corresponding to 73 and 78 are

$$\text{For } 73: \quad \frac{73 - 75}{11 / \sqrt{64}} = \frac{-2}{11/8} = -1.45$$

For 78: $\frac{76 - 75}{11/\sqrt{64}} = \frac{1}{11/8} = 0.73$

Consequently,

$$P(73 \leq \bar{X} \leq 78) = P(-1.45 < Z < 0.73) = F(0.73) - (1 - F(1.45)) = \\ = F(0.73) - 1 + F(1.45) = 0.7673 - 1 + 0.9265 = 0.6938$$

b) We now have $n=80$.

$$P(68 \leq \bar{X} \leq 83) = P\left(\frac{68 - 75}{11/\sqrt{80}} \leq Z \leq \frac{83 - 75}{11/\sqrt{80}}\right) = P(-5.69 \leq Z \leq 6.50) = \\ = F(6.50) - (1 - F(5.69)) = F(6.50) - 1 + F(5.69) \approx 1.$$

Example:

The prices of all houses in a large city have a probability distribution with a mean of \$80 000 and a standard deviation of \$15 000. Let \bar{X} be the mean price of a sample of 200 houses selected from this city.

- a) What is the probability that the mean price obtained from this sample will be within \$2 000 of the population mean?
- b) What is the probability that the mean price obtained from this sample will be more than the population mean by \$1 500 or more?

Solution:

The sampling distribution of \bar{X} is approximately normal because the sample size is large ($n>30$).

a) We need to find the probability

$$P(78000 \leq \bar{X} \leq 82000) = P\left(\frac{78000 - 80000}{15000/\sqrt{200}} \leq Z \leq \frac{82000 - 80000}{15000/\sqrt{200}}\right) = \\ = P(-1.89 \leq Z \leq 1.89) = F(1.89) - (1 - F(1.89)) = \\ = 0.9706 - 1 + 0.9706 = 0.9412.$$

b) The probability that the mean price obtained from the sample of 200 houses will be more than the population mean by \$1 500 or more is written as

$$P(\bar{X} \geq 81500) = P\left(Z \geq \frac{81500 - 80000}{15000/\sqrt{200}}\right) = P(Z \geq 1.42) = \\ = 1 - F(1.42) = 1 - 0.9222 = 0.0778.$$

Exercises

1. A large population has mean 90 and standard deviation 8. Calculate $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ for a random sample of size

- a) $n=9$ and b) $n=36$.

2. A large population has a standard deviation 10. What is the standard deviation of \bar{X} for a random sample of size

- a) $n=25$; b) $n=100$; and c) $n=400$

3. Consider a large population with $\mu=60$ and $\sigma=12$.

Assuming $n/N \leq 0.05$, find the mean and standard deviation of the sample mean \bar{X} for the sample size of

- a) 18; b) 90

4. A population of $N=5000$ has a $\sigma=20$. In each of the following cases which formula you use to calculate $\sigma_{\bar{X}}$ and why?

Using the appropriate formula, calculate $\sigma_{\bar{X}}$ for each of these cases.

- a) $n=300$; b) $n=200$; c) $n=500$; d) $n=100$.

5. For a population with $\mu=125$ and $\sigma=18$

a) For a sample selected from this population, $\mu_{\bar{X}}=125$ and $\sigma_{\bar{X}}=3.6$.

Find the sample size. Assume $n/N \leq 0.05$.

b) For a sample selected from this population, $\mu_{\bar{X}}=125$ and $\sigma_{\bar{X}}=2.25$. Find the sample size. Assume $n/N \leq 0.05$.

6. The following data show the number of automobiles owned in a population of five families:

2; 1; 0; 2; 3

a) List the ten possible samples of size 3 for this population.
(Use sampling without replacement).

b) Using the ten \bar{X} values, compute the mean and variance of \bar{X} .

c) Compute the mean and variance of the population. Compare your results to those in part b. Interpret your findings.

7. For a population, $N = 205000$, $\mu = 66$ and $\sigma = 7$, find the Z values for each of the following for $n = 49$.

a) $\bar{X} = 68.44$; b) $\bar{X} = 58.75$; c) $\bar{X} = 62.35$

8. A population has a mean of 58 and a standard deviation of 12. Assuming $n/N \leq 0.05$, find the following probabilities for a sample size of 50.

a) $P(53.7 \leq \bar{X} \leq 56.3)$; b) $P(\bar{X} \leq 59.2)$

9. Let X be a continuous random variable that has a distribution with $\mu = 90$ and $\sigma = 16$. Assuming $n/N \leq 0.05$, find the probability that the sample mean

\bar{X} for a random sample of 64 taken from this population will be

- a) less than 82.3;
- b) more than 86.7.

10. The amount of electric bills for all households in a city have a skewed probability distribution with a mean of \$65 and a standard deviation of \$25. Find the probability that the sample mean amount of electric bills for a random sample of 75 households selected from this city will be

- a) more than \$70;
- b) between \$58 and \$63;
- c) within \$6 of the population mean;
- d) more than the population mean by at least \$5?

11. The balances of all saving accounts at a local bank has a mean equal \$45 360 and standard deviation equal to \$5 900. Find the probability that the mean of a sample of 80 saving accounts selected from this bank will be

- a) less than \$46 500;
- b) between \$40 000 and \$49 200;
- c) within \$1 800 of the population mean;
- d) lower than the population mean by \$1 200 or more.

12. The mean and standard deviation of the population are 55 and 7, respectively. Sample of 40 observations is selected randomly from this population.

- a) What is the probability that sample mean \bar{X} will be between 54 and 56?
- b) Find the shortest interval centered at 55, where \bar{X} will lie with

probability 0.95.

13. Consider a random sample of size $n = 100$ from a population that has a standard deviation of $\sigma = 20$.

- Find the probability that the sample mean \bar{X} will lie within 2 units of the population mean—that is, $P(-2 \leq \bar{X} - \mu \leq 2)$.
- Find the number k so that $P(-k \leq \bar{X} - \mu \leq k) = 0.90$.
- What is the probability that \bar{X} will differ from μ by more than 4 units?

Answers

- 1.** a) $\mu_{\bar{x}} = 90$; $\sigma_{\bar{x}} = 2.67$; b) $\mu_{\bar{x}} = 90$; $\sigma_{\bar{x}} = 1.33$; **2.** a) 2; b) 1; c) 0.5;
3. a) $\mu_{\bar{x}} = 60$; $\sigma_{\bar{x}} = 2.828$; b) $\mu_{\bar{x}} = 60$; $\sigma_{\bar{x}} = 1.265$;
4. a) $\sigma_{\bar{x}} = 1.120$; b) $\sigma_{\bar{x}} = 1.414$; c) $\sigma_{\bar{x}} = 0.849$; d) $\sigma_{\bar{x}} = 2.000$; **5.** a) $n = 25$;
b) $n = 64$; **7.** a) $Z = 2.44$; b) $Z = -7.25$; c) $Z = -3.65$; **8.** a) 0.1530; b) 0.7611;
9. a) 0.0001; b) 0.9505; **10.** a) 0.0418; b) 0.2373; c) 0.9624; d) 0.0418;
11. a) 0.9582; b) 1.0; c) 0.9936; d) 0.0344; **12.** a) 0.6318; b) [52.82; 57.18];
13. a) 0.6826; b) $k = 3.29$; c) 0.0456.

5.2. Sampling distribution of a sample proportion

5.2.1. Population and sample proportions

The concept of proportion is the same as the concept of relative frequency discussed in Chapter 2 and the concept of probability of success in a binomial distribution. The relative frequency of a category or class gives the proportion of the sample or proportion that belongs to that category or class. Similarly, the probability of success in a binomial problem represents the proportion of the sample or population that possesses a given characteristic.

The **population proportion**, denoted by p , is obtained by taking the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population.

The **sample proportion**, denoted by \hat{p} (read as "p hat") gives a similar ratio for a sample.

Definition: The population and sample proportions, denoted by p and \hat{p} , respectively, are calculated as

$$p = \frac{x}{N} \quad \text{and} \quad \hat{p} = \frac{\hat{x}}{n}$$

where

N = total number of elements in the population;

n = total number of elements in the sample;

x = number of elements in the population or sample that possesses a specific characteristic.

Example:

Suppose a total of 393 217 families live in a city and 123 017 of them own at least one car. Then,

N = population size = 393 217

x = families in the population who own car = 123 017.

The proportion of families in this city who own car is

$$p = \frac{x}{N} = \frac{123017}{393217} = 0.31.$$

Now, suppose that a sample of 560 families is taken from this city and 215 of them have at least one car. Then

n = sample size = 560

x = families in the sample who own car = 215.

The sample proportion is

$$\hat{p} = \frac{\hat{x}}{n} = \frac{215}{560} = 0.38.$$

5.2.2. Sampling distribution of \hat{p} . Its mean and standard deviation

Just like the sample mean, \bar{X} , the sample proportion \hat{p} is also a random variable. Hence, it possesses a probability distribution, which is called its **sampling distribution**.

It can be shown by relying on the definition of the mean that the mean value of \hat{p} - that is, the mean of all possible values of \hat{p} is equal to the population proportion p just as the mean of the sampling distribution.

Definition:

The mean of the sample proportion \hat{p} is denoted by $\mu_{\hat{p}}$ and is equal to the population proportion p . Thus,

$$E(\hat{p}) = \mu_{\hat{p}} = p.$$

The mean of all possible \hat{p} values is equal to the population proportion p .

Since $E(\hat{p}) = p$, the sample proportion is an unbiased estimator of the population proportion.

Now we are interested in determining the standard deviation of the \hat{p} values.

Just as in the case of sample mean, \bar{X} , the standard deviation of \hat{p} depends on whether the sample size is a small proportion of the population or not.

Definition:

The standard deviation of the sample proportion \hat{p} is denoted by $\sigma_{\hat{p}}$ and defined as

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p \cdot q}{n}}$$

where

p – is the population proportion,

$q = 1 - p$, and n – is the sample size.

This formula is valid when $n/N \leq 0.05$, where N – is the population size.

If $n/N > 0.05$, then $\sigma_{\hat{p}}$ is calculated as follows:

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} \cdot \sqrt{\frac{N-n}{N-1}},$$

where $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor.

5.2.3. Form of the sampling distribution of \hat{p}

Now that we know the mean and standard deviation of \hat{p} , and we want to consider the form of the sampling distribution of \hat{p} . Applying the central limit theorem as it relates to the \hat{p} random variable, we have the following:

Definition:

According to the central limit theorem, the sampling distribution of \hat{p} is approximately normal for a sufficiently large sample size.

The random variable

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

is approximately distributed as a standard normal.

This approximation is good if $n \cdot p \cdot q > 9$.

Summary

Let \hat{p} be the sample proportion of success in a random sample from a population with proportion of success p .

Then

1. The sampling distribution of \hat{p} has mean p

$$E\left(\hat{p}\right) = \mu_{\hat{p}} = p.$$

2. The sampling distribution of \hat{p} has a standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p \cdot q}{n}} \quad \text{if } n/N \leq 0.05$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{if } n/N > 0.05.$$

3. The Z value for a value of \hat{p} is

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}.$$

Once again, the last approximation is good if $n \cdot p \cdot q > 9$.

Example:

The firm makes deliveries of a large number of products to its customers. It is known that 75% of all the orders it receives from its customers are delivered on time. Let \hat{p} be the proportion of orders in a random sample of 120 that are delivered on time. Find the probability that the value of \hat{p} will be

- a) between 0.73 and 0.80;
- b) less than 0.72.

Solution:

From the given information,

$$p = 0.75, \quad q = 1 - p = 1 - 0.75 = 0.25,$$

where p is the proportion of orders in the population.

The mean of the sample proportion \hat{p} is

$$\mu_{\hat{p}} = p = 0.75$$

The standard deviation of \hat{p} is

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0.75 \cdot 0.25}{210}} = 0.0299.$$

Let us find $n \cdot p \cdot q$.

$$n \cdot p \cdot q = 210 \cdot 0.75 \cdot 0.25 = 39.38.$$

Since $n \cdot p \cdot q = 39.38 > 9$, we can infer from the central limit theorem that the sampling distribution of \hat{p} is approximately normal.

Next, the two values of \hat{p} are converted to their respective Z values by

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}.$$

- a) For $\hat{p} = 0.73$; $Z = \frac{0.73 - 0.75}{0.0299} = -0.67$.
 For $\hat{p} = 0.80$; $Z = \frac{0.80 - 0.75}{0.0299} = 1.67$.

The required probability is (Figure 5.5).

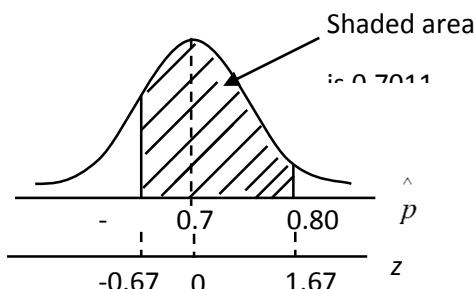


Fig.5.5 $P(0.73 < \hat{p} < 0.80)$

$$P(0.73 < \hat{p} < 0.80) = P(-0.67 < Z < 1.67) = F(1.67) - (1 - F(0.67)) = \\ = 0.9525 - 1 + 0.7486 = 0.7011.$$

Thus, the probability is 0.7011 that between 73% and 80% of orders of the sample of 210 orders will be delivered on time.

$$\text{b)} P(\hat{p} < 0.72) = P(Z < \frac{0.72 - 0.75}{0.0299}) = P(Z < -1.01) = \\ = 1 - F(1.01) = 1 - 0.8438 = 0.1567.$$

Thus, the probability that less than 72% of the sample of 210 orders will be delivered on time is 0.1567.

Exercises

1. For a population, $N = 40\,000$ and $p = 0.65$, find the Z value for each of the following for $n = 200$

$$\text{a)} \hat{p} = 0.59; \quad \text{b)} \hat{p} = 0.72; \quad \text{c)} \hat{p} = 0.43; \quad \text{d)} \hat{p} = 0.73$$

2. 83% of the households of a large city own VCRs. Let \hat{p} be the population of the households who own VCRs in a random sample of 400 households.

Find the probability that the value of \hat{p} will be

- a) between 0.85 and 0.88
- b) more than 0.80

3. A doctor believes that 80% of all patients having a particular disease will be fully recovered within 3 days after receiving a new drug. Assume that a random sample of 230 patients is selected.

- a) What is the mean of the sample proportion of patients?
- b) What is the variance of the sample proportion?
- c) What is the standard error (standard deviation) of the sample proportion?
- d) What is the probability that the sample proportion is less than 0.75?
- e) What is the probability that the sample proportion is between 0.78 and 0.85?

4. Sixty percent of adults favor some kind of government control on the prices of medicines.

- a) Find the probability that the proportion of adults in a random sample of 200 who favor some kind of government control on the prices of medicines is
- i) less than 0.55;
 - ii) between 0.57 and 0.68.
- b) What is the probability that the proportion of adults in a random sample of 200 who favor some kind of government control is within 0.04 of the population proportion?
- c) What is the probability that the sample proportion is greater than the population proportion by 0.06 or more?

5. Stress on the job is a major concern of a large number of people who go into managerial positions. Eighty percent of all managers of companies suffer

from stress. Let \hat{p} be the proportion in a sample of 100 managers of companies who suffer from stress.

- a) What is the probability that this sample proportion is within 0.08 of the population proportion?
- b) What is the probability that this sample proportion is not within 0.08 of the population proportion?
- c) What is the probability that this sample proportion is lower than the population proportion by 0.10 or more?
- d) What is the probability that this sample proportion is greater than the population proportion by 0.11 or more?

6. A private university has 1250 students. Of these, 357 concerned about the GPA. A random sample of 265 students was taken.

- a) What is the standard error (standard deviation) of the sample proportion of students who are concerned about the GPA?
- b) What is the probability that the sample proportion is less than 0.35?
- c) What is the probability that the sample proportion is between 0.25 and 0.33?

7. A plant has total of 736 employees. Of these, 342 are married. A random sample of 170 employees was taken.

- a) What is the mean of the sample proportion of married employees?
- b) What is the standard error of the sample proportion of married employees?
- c) What is the probability that the sample proportion is greater than 0.37?
- d) What is the probability that the sample proportion is between 0.43 and 0.53?

8. Suppose that 78% of all adults like sport.

- a) Find the probability that the proportion of adults who like sport in a random sample of 400 is
- i) more than 0.81;
 - ii) between 0.75 and 0.82
 - iii) less than 0.80;
 - iv) between 0.73 and 0.76
- b) What is the probability that the proportion of adults in a random sample of 400 who like sport is within 0.05 of the population proportion?
- c) What is the probability that the proportion of adults in a random sample of 400 who like sport is lower than the population proportion by 0.04 or more?

Answers

1. a) -1.78; b) 2.08; c) -6.53; d) 2.37; 2. a) 0.1426; b) 0.9429; 3. a) 0.80;
b) 0.000696; c) 0.0264; d) 0.0294; e) 0.7470; 4. a) i) 0.0735; ii) 0.7974;
b) 0.754; c) 0.0418; 5. a) 0.9544; b) 0.0456; c) 0.0062; d) 0.0030;
6. a) 0.0246; b) 0.9956; c) 0.8906; 7. a) 0.4647; b) 0.0336; c) 0.9976;
d) 0.8223; 8. a) i) 0.0735; ii) 0.8949; iii) 0.8289; iv) 0.8202; b) 0.9844;
c) 0.0268.

Chapter 6

Interval estimation

6.1. Introduction

The problem of statistical inference arises when we wish to make generalization about a population when only a sample will be available. Once a sample is observed, its main features can be determined by the methods of descriptive summary discussed in previous chapters. Our principal concern is with not just the particular data set, but what can be said about the population based on the information extracted from analyzing the sample data.

Statistical inference deals with drawing conclusions about population parameters from an analysis of the sample data.

The value(s) assigned to a population parameters based on the value of a sample statistic is called **an estimate** of the population parameters.

For example, suppose the manager selects a sample of 50 new employees and finds that the mean time \bar{x} taken to learn the job for these employees is 10 hours. If manager assigns this value to the population mean, then 10 hours will be called an estimate of μ . Thus, the sample mean \bar{x} is an estimator of the population mean μ , and the sample proportion \hat{p} is an estimator of the population proportion p .

An estimate may be a point estimate or an interval estimate.

Definition:

The value of a sample statistic that is used to estimate population parameters is called a **point estimate**.

Each sample taken from a population is expected to yield a different value of the sample statistics. Thus, the value assigned to a population parameter based on the point estimate depends on which of the sample is drawn. Consequently, the point estimate assigns a value to a population parameters almost always differs from the true value of the population parameters.

In the case of **interval estimation**, instead of assigning a simple value to a population parameter, an interval is constructed around the point estimate and then a probability statement that this interval contains the corresponding population parameter is made.

Definition:

In interval estimation, an interval is constructed around the point estimate, and it is stated that this interval likely to contain the corresponding population parameter.

6.2. Confidence interval and confidence level

Since interval estimators have been described as “likely” to contain the true, but unknown value of the population parameters, it is necessary to phrase such term as probability statement.

Suppose that a random sample is selected and based on the sample information, it is possible to find two random variables a and b . Then interval extending from a and b either includes the population parameter or it does not contain population parameter. However, suppose that the random samples are repeatedly selected from the population and similar intervals are found. In the long run, a certain percentage of this interval will contain the unknown value. According to the frequency concept of probability, an interpretation of such intervals follows:

If the population is repeatedly samples and intervals calculated, then in the long run 90% (or some other percentages) of the intervals would contain the true value of the unknown parameter. The interval from a and b is said to be 90% (or some other percentages) confidence interval estimator for population parameters.

Definition:

Let θ be unknown parameter. Suppose that based on sample information, random variables a and b are found such that

$$P(a < \theta < b) = 1 - \alpha ,$$

where α -is any number between 0 and 1.

The interval from a to b is called $100 \cdot (1 - \alpha)\%$ **confidence interval** for θ .

The quantity $(1 - \alpha)$ is called the **confidence level** of the interval.

If the population were repeatedly sampled a very large number of times, the true value of the parameter θ would be contained in $100 \cdot (1 - \alpha)\%$ of intervals calculated this way.

The confidence interval calculated in this way is written as $a < \theta < b$ with $100 \cdot (1 - \alpha)\%$ confidence.

Let us find the confidence intervals with any required confidence level $(1 - \alpha)$, where α is any number such that $0 < \alpha < 1$.

We will use the notation z_α for the number such that

$$P(Z > z_\alpha) = \alpha.$$

A notation z_α indicates the value in the standard normal table cuts off a right tail area of α . (Fig. 6.1).

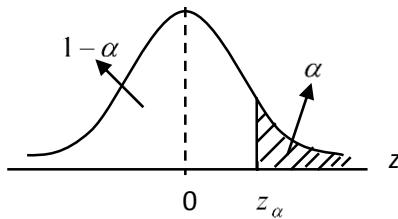


Fig. 6.1. $P(Z > z_\alpha), \dots$

For example, if $\alpha = 0.13$, then $1 - \alpha = 0.87$.

So,

$$F(z_\alpha) = F(z_{0.13}) = 0.87$$

and from the standard normal distribution table we obtain $z_{0.13} \approx 1.125$.

Therefore

$$P(Z > 1.125) = 0.13.$$

Now suppose that a $100 \cdot (1 - \alpha)\%$ confidence interval is required. (Fig. 6.2).

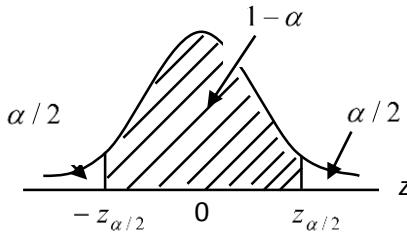


Fig. ... $-z_{\alpha/2} < Z < z_{\alpha/2}, = 1 - \alpha$

We have

$$P(Z > z_{\alpha/2}) = \alpha/2$$

By the symmetry about the mean

$$P(Z < -z_{\alpha/2}) = \alpha/2.$$

And it follows that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

where the random variable Z follows a standard normal distribution.

Example:

Find the value of $z_{\alpha/2}$ if $\alpha = 0.1$.

Solution:

$$\alpha = 0.1$$

$$\alpha/2 = 0.05.$$

$$P(Z > z_{\alpha/2}) = P(Z > z_{0.05}) = 0.05.$$

$$P(Z > z_{0.05}) = F(z_{0.05}) = 0.95 \text{ and}$$

$$z_{0.05} = 1.645 \text{ (Fig. 6.3).}$$

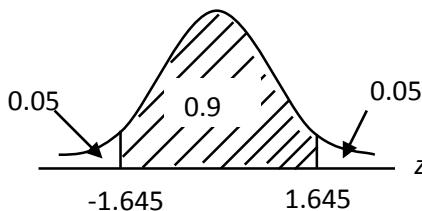


Fig.6.3

6.3. Confidence intervals for the mean of population that is normally distributed: population variance known

Let X_1, X_2, \dots, X_n be a random sample of n observations from a normal population with unknown μ and known variance σ^2 . Let \bar{x} be the sample mean. Then $100 \cdot (1 - \alpha)\%$ confidence interval for the population mean with known variances is given by

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the number for which $P(Z > z_{\alpha/2}) = \alpha/2$ and the random variable Z has a standard normal distribution.

Example:

Given a random sample of 36 observations from a normal population for which μ is unknown and $\sigma = 8$, the sample mean is found to be $\bar{x} = 45.3$. Construct a 95% confidence interval for μ .

Solution:

$$100 \cdot (1 - \alpha)\% = 95\%$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05 \quad \text{and} \quad \alpha/2 = 0.025$$

$$P(Z > z_{\alpha/2}) = P(Z > z_{0.025}) = 0.025$$

$$P(Z > z_{0.025}) = F(z_{0.025}) = 0.975$$

$$z_{\alpha/2} = z_{0.025} = 1.96$$

From

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

we obtain that

$$45.3 - 1.96 \cdot \frac{8}{\sqrt{36}} < \mu < 45.3 + 1.96 \cdot \frac{8}{\sqrt{36}}$$

$$42.69 < \mu < 47.91$$

So, $(42.68, 47.91)$ is a 95% confidence interval for μ .

It means, if sample of 36 observations are drawn repeatedly and independently from the population, then over a very large number of repeated trials, 95% of these intervals will contain the value of the true population mean.

Exercises

1. Find $z_{\alpha/2}$ for each of the following confidence levels

- a) 88%; b) 94%; c) 96%; d) 99%

2. The standard deviation for a population is $\sigma = 12.6$. A sample of 36 observations selected from this population gave a mean equal to 74.8.

- a) Make a 90 % confidence interval for μ .
- b) Construct a 95 % confidence interval for μ .
- c) Determine a 99 % confidence interval for μ .

3. The standard deviation for a population is $\sigma = 8.3$. A sample of 121 observations selected from this population gave a mean equal to 84.5.

- a) Make a 99 % confidence interval for μ
- b) Construct a 95 % confidence interval for μ
- c) Determine a 90 % confidence interval for μ

4. The standard deviation for a population is $\sigma = 6.30$. A random sample selected from this population gave a mean equal to 78.90.

- a) Make a 99 % confidence interval for μ assuming $n = 36$
- b) Make a 99 % confidence interval for μ assuming $n = 81$
- c) Make a 99 % confidence interval for μ assuming $n = 100$

Explain your results.

5. Given that a sample of size 16 from a normal distribution yielded $\bar{x} = 25$. The population variance is known to be 64. Find

- a) 90 % confidence interval for μ
- b) 95 % confidence interval for μ
- c) 99 % confidence interval for μ .

Answers

1. a) 1.555; b) 1.88; c) 2.055; d) 2.575; 2. a) (71.35; 78.26);
b) (70.68; 78.92); c) (69.39; 80.21); 3. a) 82.56 to 86.44; b) 83.03 to 85.97; c)
83.26 to 85.74; 4. a) 76.19 to 81.61; b) 77.09 to 80.71; c) 77.27 to 80.53; 5. a)
(21.71, 28.29); b) (21.08, 28.92); c) (19.85, 30.15).

6.4. Confidence intervals for the mean of population that is normally distributed: large sample size

Let X_1, X_2, \dots, X_n be a random sample of n observation from a normal population with unknown μ and unknown variance σ^2 . Let \bar{x} be the sample mean. If the sample size n is large ($n \geq 30$), then according to the central limit theorem, for a large sample the sampling distribution of the sample mean \bar{x} is (approximately) normal irrespective of the shape of the population from which the sample is drawn. Therefore, when the sample size is 30 or larger, we will use the normal distribution to construct a confidence interval for μ . If the variance of the population is unknown, it should be estimated by the sample variance, s_x^2 , and σ replaced by s_x in confidence interval formula for the case when population variance is known.

Remark:

For large sample sizes, s_x is usually close to the true value of σ .

An approximate $100 \cdot (1 - \alpha)\%$ confidence interval for the population mean with unknown variance is given by

$$\bar{x} - z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}}$$

where \bar{x} is based on a sample of at least thirty observations, $z_{\alpha/2}$ is the number for which $P(Z > z_{\alpha/2}) = \alpha/2$ and the random variable Z has a standard normal distribution.

Example:

A sample of 64 observations from a large population yielded the sample values, $\bar{x} = 172$ and $s_x^2 = 299$. Find an approximate 99 % confidence interval for μ .

Solution:

First we find the standard deviation of \bar{x} . Because σ is not known, we will use s_x as an estimator of σ .

$$s_x = \sqrt{s_x^2} = \sqrt{299} = 17.29.$$

Then

$$100 \cdot (1 - \alpha)\% = 99\%$$

$$1 - \alpha = 0.99$$

$$\alpha = 0.01 \quad \text{and} \quad \alpha/2 = 0.005$$

$$P(Z > z_{\alpha/2}) = P(Z > z_{0.005}) = 0.005$$

$$P(Z > z_{0.005}) = F(z_{0.005}) = 0.995$$

$$z_{\alpha/2} = z_{0.005} = 2.58$$

Substituting all the values in the formula, the 99 % confidence interval for μ is

$$\bar{x} - z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}}$$

$$172 - 2.58 \cdot \frac{17.29}{8} < \mu < 172 + 2.58 \cdot \frac{17.29}{8}$$

$$166.4 < \mu < 177.6$$

An approximate 99 % confidence interval for μ is (166.4, 177.6).

Example:

Radiation measurements on a sample of 69 microwave ovens produced

$\bar{x} = 0.13$ and $s_x = 0.04$. Determine a 94 % confidence interval for the mean radiation.

Solution:

Again since $n = 69 > 29$ we can use s_x as an estimator of σ .

Then $100 \cdot (1 - \alpha)\%$ confidence interval for the population mean with unknown variance is given by

$$\bar{x} - z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}}$$

$$1 - \alpha = 0.94$$

$$\alpha = 0.06 \quad \text{and} \quad \alpha/2 = 0.03$$

$$P(Z > z_{\alpha/2}) = P(Z > z_{0.03}) = 0.03$$

$$P(Z > z_{0.03}) = F(z_{0.03}) = 0.97$$

$$z_{\alpha/2} = z_{0.03} = 1.88$$

Substituting all the values in the formula we obtain

$$0.13 - 1.88 \cdot \frac{0.04}{\sqrt{69}} < \mu < 0.13 + 1.88 \cdot \frac{0.04}{\sqrt{69}}$$

$$0.121 < \mu < 0.139.$$

Thus, we can state with 94 % confidence that average radiation measure of microwave ovens is between (0.21, 0.139).

Exercises

1. Determine a 90 % confidence interval for μ if $n = 48$, $\bar{x} = 86.5$, and $s_x = 7.9$.

2. Determine a 98 % confidence interval for μ if $n = 150$, $\bar{x} = 0.865$, and $s_x = 0.057$.

3. A sample of size 50 from a population yielded the sample values $\bar{x} = 190$, and $s_x^2 = 800$. Find a 95 per cent confidence interval for μ .

4. For a sample data set, $\bar{x} = 16$, and $s_x = 5.3$

- a) Construct a 95 % confidence interval for μ assuming $n = 50$.
- b) Construct a 90 % confidence interval for μ assuming $n = 50$. Is the width of the 90 % confidence interval smaller than the width of the 95 % confidence interval calculated in part a? If yes, why it is so?
- c) Find a 95% confidence interval for μ assuming $n = 100$. Is the width of the 95 % confidence interval for μ with $n = 100$ smaller than the width of the 95 % confidence interval for μ with $n = 50$ calculated in part a?

If so, why?

5.

- a) A sample of 100 observations taken from a population produced a sample mean equal to 55.32 and a standard deviation equal to 8.4. Make a 90 % confidence interval for μ .
- b) Another sample of 100 observations taken from the same population produced a sample mean equal to 57.40 and a standard deviation equal to 7.5. Make a 90 % confidence interval for μ .
- c) A third sample of 100 observations taken from the same population produced a sample mean equal to 56.25 and a standard deviation equal to 7.9. Make a 90 % confidence interval for μ .
- d) The true population mean for this population is 55.80. How many of the confidence intervals constructed in a-c cover this population mean and how many do not?

6. The mean annual salaries of managers at the certain company is \$80 722 for males and \$65 258 for females. These mean salaries are based on samples of 400 male and 200 female managers. Assume that the standard deviation of the annual salaries of male managers is \$11 500 and standard deviation of the female managers is \$8 400.

- a) Construct a 95 % confidence interval for the mean annual salary of male managers.
- b) Construct a 95 % confidence interval for the mean annual salary of female managers.

7. From a random sample of 70 high school seniors, the sample mean and standard deviation of the math scores are found to be 96 and 17 respectively. Determine a 96% confidence interval for the mean math score of all seniors in the school.

8. With a random sample of size $n = 144$, someone proposes

$$\left(\bar{x} - 0.12 \cdot s_x; \bar{x} + 0.12 \cdot s_x \right)$$

to be a confidence interval for μ . What then is the level of confidence?

Answers

1. (84.63, 88.38); 2. (0.854, 0.876); 3. (182.2, 197.8); 4. a) (14.53, 17.47);
b) (14.76, 17.24); c) (14.96, 17.04); 5. a) 53.94 to 56.70; b) 56.17 to 58.63;
c) 54.95 to 57.55; 6. a) (\$79\ 595, \\$81\ 849); b) (\$64\ 093, \\$66\ 422);
7. 91.83 to 100.17; 8. 0.8502 or about 85 %.

6.5. Confidence intervals for the mean of a normal distribution: population variance unknown: small sample size

In previous topics we discussed inferences about a population mean when a large sample is available. Those methods are deeply rooted in the central

limit theorem, which guarantees that the distribution of \bar{X} is approximately normal.

Many investigations require statistical inferences to be drawn from small samples ($n < 30$). Since the sample mean \bar{x} will still be used for inferences about μ , we must address the question, “what is the sampling distribution of \bar{X} when n is not large?”. Unlike the large sample situations, here we do not have an unqualified answer, and central limit theorem is no longer applicable.

6.5.1. Student's t distribution

Consider a sampling situation where the population has a normal distribution with unknown σ . Because σ is unknown, an intuitive approach is to estimate σ by the sample standard S . Just as we did in the large sample situation, we consider the ratio

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

This random variable does not follow a standard normal distribution. Its distribution is known as **Student's t distribution**.

The graph of the t -distribution resembles the graph of the standard normal distribution: they both are symmetric, bell shaped curves with mean equal to zero. The graph of the Student's t distribution is lower at the center and higher at the extremities than the standard normal curve. (Fig. 6.4).

The new notation t is required in order to distinguish it from the standard normal variable Z . As the number of degrees of freedom increases, the difference between t distribution and the standard normal distribution becomes smaller and smaller.

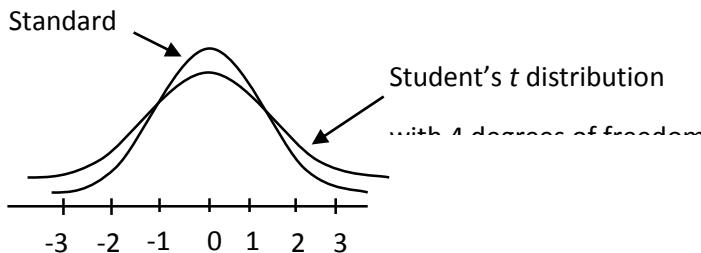


Fig.6.7.

The qualification “with $(n - 1)$ degrees of freedom” is necessary, because with each different sample size or value of $(n - 1)$, there is a different t distribution.

Definition:

The **number of degrees of freedom** is defined as the number of observations that can be chosen freely.

Example:

Suppose we know that the mean number of 5 values is 25. Consequently, the sum of these 5 values is $125 (5 \cdot 25 = 125)$. Now how many values out of 5 can be chosen freely so that the sum of these 5 values is 125? The answer is that we can freely choose $5 - 1 = 4$ values. Suppose we choose 15, 35, 45, and

10 as the 4 values. Given these 4 values and the information that the mean of the 5 values is 25, the 5th value is

$$125 - (15 + 35 + 45 + 10) = 15$$

Thus, once we have chosen 4 values, the fifth value is automatically determined. Consequently, the number of degrees of freedom for this example is

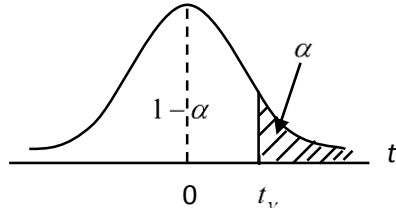
$$d.f. = n - 1 = 5 - 1 = 4$$

We subtract 1 from n because we lose one degree of freedom to calculate the mean.

The t -table in the Appendix (see table 4) is arranged to give the value t for several frequently used values of α and for a number of values ($n - 1$).

Definition:

A random variable having the standard distribution with v (Greek letter nu) Degrees of freedom will be denoted by t_v (Fig. 6.8). Then $t_{v,\alpha}$ is defined as the number for which



$$P(t_v > t_{v,\alpha}) = \alpha$$

Fig.6.8

Example: Find $t_{5,0.10}$

Solution:

In words it means we need to find a number that is exceeded with the probability 0.10 by a Student's t random variable with 5 degrees of freedom.

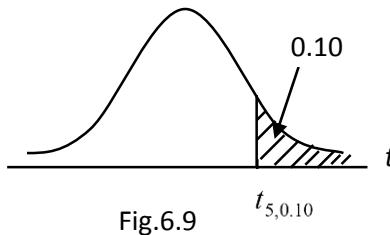


Fig.6.9

$$P(t_5 > t_{5,0.10}) = 0.10$$

From table 4 of the Appendix we read that $t_{5,0.10} = 1.476$. (Fig. 6.9).

Similarly, to $z_{\alpha/2}$ for Student's t distribution the value $t_{\nu,\alpha/2}$ is defined as
 $P(t > t_{\nu,\alpha/2}) = \alpha/2$.

6.5.2. Confidence interval for μ : small samples

Let us turn our attention to finding $100 \cdot (1 - \alpha)\%$ confidence interval for μ when sample size is small. Using

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

We can derive the formula for $100 \cdot (1 - \alpha)\%$ confidence interval for the case when a small sample is selected from a normally distributed population with mean μ and unknown variance. It is given by

$$\bar{x} - t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the number for which

$$P(t_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$$

The random variable t_{n-1} has a Student's t distribution with $\nu = (n - 1)$ degrees of freedom. (Fig. 6.10).

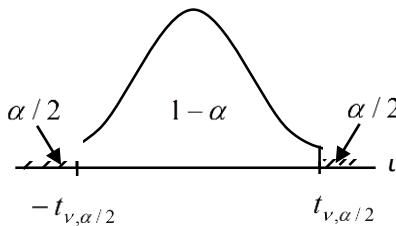


Fig.6.10

Remark:

If the sample is available, then standard deviation can be calculated as

$$S = \sqrt{S^2}, \text{ where } S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{or}$$

$$S^2 = \frac{1}{n-1} \left[\sum x_i^2 - n \cdot \left(\bar{x} \right)^2 \right]$$

Example:

For the t distribution with $n=10$, find the number b such that

$$P(-b < t < b) = 0.80$$

Solution:

The probability in the interval $(-b, b)$ is 0.80. (Fig. 6.11).

We must have a probability of 0.10 to the right of b and a probability of 0.10 to the left of $-b$.

So

$$P(t_9 > t_{9,0.10}) = 0.10$$

$$t_{9,0.10} = 1.383$$

$$b = 1.383 \text{ and } -b = -1.383.$$

Example:

A random sample of 25 busses shows a sample mean of 225 passengers carried per day per bus. The sample standard deviation is computed to be 60 passengers. Find a 90% confidence interval for the mean number of passengers carried per bus during a 1-day period.

Solution:

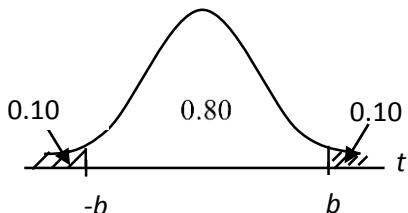


Fig.6.11

A 90 % confidence interval for the mean μ is given by

$$\bar{x} - t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

$$n = 25, \text{ so } v = n-1 = 25-1 = 24$$

$$100 \cdot (1-\alpha)\% = 90\%$$

$$1-\alpha = 0.90$$

$$\alpha = 0.10 \text{ and } \alpha/2 = 0.05$$

$$t_{n-1,\alpha/2} = t_{24,0.05}$$

$$P(t_{24} > t_{24,0.05}) = 0.05$$

$$t_{24,0.05} = 1.711$$

After substitution we obtain

$$225 - 1.711 \cdot \frac{60}{5} < \mu < 225 + 1.711 \cdot \frac{60}{5}$$

$$204.5 < \mu < 245.5 \text{ or } (204.5, 245.5).$$

We are 90 % confident the mean number of passengers carried per day by bus is between 204.5 and 245.5, because 90 % of the intervals calculated in this manner will contain the true mean number of passengers carried per day per bus.

Exercises

1. In each case, find the number b so that

- $P(t < b) = 0.95$ when $n = 7$
- $P(-b < t < b) = 0.95$ when $n = 16$
- $P(t > b) = 0.01$ when $n = 9$
- $P(t > b) = 0.99$ when $n = 12$

2. For each of the following, find the area in the appropriate tail of the t distribution

- $t = 2.060$ and $n = 26$
- $t = -3.686$ and $n = 17$
- $t = -2.650$ and $n = 15$
- $t = 2.845$ and $n = 22$

3. Find the value of t from t -distribution table for each of the following:

- a) Confidence level = 99 % and $d.f. = 18$
 - b) Confidence level = 95 % and $n = 26$
 - c) Confidence level = 90 % and $d.f. = 15$
- 4.** The mean number of the sample of 25 bolts produced on a specific machine per day was found to be 47 with a standard deviation of 2.4. Assume that the number of bolts produced per day on this machine has a normal distribution. Construct a 90 % confidence interval for the population mean μ .
- 5.** A random sample of 16 cars, which were tested for fuel consumption, gave a mean of 26.4 miles per gallon with a standard deviation of 2.3 miles per gallon. Assuming that the miles per gallon given by cars have a normal distribution, find a 99 % confidence interval for the population mean μ .
- 6.** A sample of eight adults was taken, and these adults were asked about the time they spend per week on sport activities. Their responses (in hours) are as follows:
- 45; 12; 31; 16; 28; 14; 18; 26
- Make a 95 % confidence interval for the mean of time spent per week by all adults on sport activities.
- 7.** A sample of 10 customers who visited a supermarket was taken. The following data give the money (in dollars) they spent during that visit:
74; 89; 121; 63; 146; 47; 91; 28; 84; 76
Assuming that the money spent by all customers at this supermarket has a normal distribution, construct a 90 % confidence interval for the population mean.
- 8.** The mean time taken to design a home plan by 20 designers was found to be 185 minutes with a standard deviation of 23 minutes. Assume that the time taken by all designers this home plan is normally distributed. Construct a 99 % confidence interval for the population mean μ .

Answers

- 1.** a) 1.943; b) 2.131; c) 2.896; d) -2.718; **2.** a) $\alpha = 0.025$; right tail;
b) $\alpha < 0.005$; left tail; c) less than $\alpha = 0.01$; left tail; d) less than $\alpha = 0.005$; right tail; **3.** a) 2.878; b) 2.060; c) 1.753; **4.** a) (46.18; 47.82);
5. (24.71; 28.09); **6.** (14.52; 32.98); **7.** 62.24 to 101.56; **8.** (170.29; 199.71).

6.6. Confidence intervals for population proportion: Large samples

The reason leading to estimation of a mean also applies to the problem of estimation of a population proportion.

Suppose that n elements are randomly selected from the large population. And let n consists of X elements possessing some characteristic.

Common sense suggests the sample proportion

$$\hat{p} = \frac{\hat{X}}{n}$$

as an estimator of p .

When the sample size n is only small fraction of the population size, the sample count X has the binomial distribution with mean $n \cdot p$ and standard deviation $\sqrt{n \cdot p \cdot q}$.

When n is large, the binomial variable X is well approximated by a normal distribution with mean $n \cdot p$ and standard deviation $\sqrt{n \cdot p \cdot q}$.

That is

$$Z = \frac{\hat{X} - n \cdot p}{\sqrt{n \cdot p \cdot q}}$$

is approximately standard normal.

If we divide both numerator and denominator by n we will get a statement about proportions:

$$Z = \frac{(\hat{X} - n \cdot p) / n}{\sqrt{n \cdot p \cdot q / n}} = \frac{\hat{p} - p}{\sqrt{p \cdot q / n}}$$

As we see the denominator of Z contains p and q . If sample size is large that $n \cdot p \cdot q > 9$, then a good approximation is obtained if p replaces the point estimator \hat{p} in the denominator.

Hence, for large sample size, $\left(n \cdot \hat{p} \cdot \hat{q} > 9 \right)$ the distribution of

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p} \cdot \hat{q} / n}}$$

is approximately standard normal.

We can use this result to obtain $100 \cdot (1 - \alpha)\%$ confidence interval for the population proportion. (Fig.6.12)

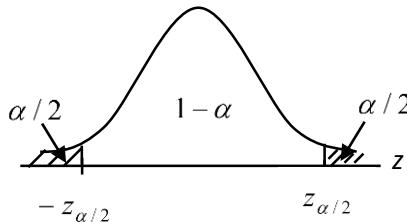


Fig.6.12

Using Fig.6.12 we

obtain

$$\begin{aligned}
 1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} < z_{\alpha/2}) = \\
 &= P\left(-z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < \hat{p} - p < z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right) = \\
 &= P\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < \hat{p} < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right)
 \end{aligned}$$

Definition:

If sample of n observations selected from the population is large enough that $n \cdot p \cdot q > 9$, then a $100 \cdot (1 - \alpha)\%$ confidence interval for the population proportion is given by

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where $\hat{p} = \frac{X}{n}$ is the sample proportion and $z_{\alpha/2}$ is the number for which

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}.$$

Example:

From a country labor force a random sample of 800 persons was selected and 75 people were found unemployed out of random sample of 800 persons. Compute 90 % confidence interval for the rate of unemployment in the country.

Solution:

The confidence interval for the population proportion is obtained from

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The observed $\hat{p} = \frac{75}{800} = 0.09375$ and $1 - \hat{p} = 0.90625$.

Since $n \cdot \hat{p} \cdot q = 800 \cdot 0.09375 \cdot 0.90625 = 67.97 > 9$, we say that sample size is large and a normal approximation to the sample proportion \hat{p} is justified.

Then

$$100 \cdot (1 - \alpha)\% = 90\%$$

$$1 - \alpha = 0.90$$

$$\alpha = 0.1 \quad \text{and} \quad \alpha/2 = 0.05$$

$$P(Z > z_{\alpha/2}) = P(Z > z_{0.05}) = 0.05$$

$$P(Z > z_{0.05}) = F(z_{0.05}) = 0.95$$

$$z_{\alpha/2} = z_{0.05} = 1.645$$

After substituting, we obtain

$$0.09375 - 1.645 \cdot \sqrt{\frac{0.09375 \cdot 0.90625}{800}} < p < 0.09375 + 1.645 \cdot \sqrt{\frac{0.09375 \cdot 0.90625}{800}}$$

$$0.0768 < p < 0.1107$$

Therefore, a 90 % confidence interval for the rate of unemployment in the country is (0.0768; 0.1107), or (7.68 %; 11.07 %).

Because our procedure will produce true statements 90 % of the time, we can be 90 % confident that the rate of unemployment is between 7.68 and 11.07.

Exercises

1. Check if the sample size is large enough to use the normal distribution to make a confidence interval for p for each of the following cases

- a) $n = 60$ and $\hat{p} = 0.30$
- b) $n = 180$ and $\hat{p} = 0.027$
- c) $n = 200$ and $\hat{p} = 0.73$
- d) $n = 65$ and $\hat{p} = 0.05$

2. A sample of 500 observations selected from a population gave a sample proportion equal to 0.72.

- a) make a 90 % confidence interval for p .
- b) construct a 95 % confidence interval for p .
- c) make a 99 % confidence interval for p .

Interpret your results.

3. A sample selected from a population gave a sample proportion equal to 0.73

- a) make a 98 % confidence interval for p assuming $n = 90$
- b) construct a 98 % confidence interval for p assuming $n = 500$
- c) construct a 98 % confidence interval for p assuming $n = 100$

Interpret your results.

4. A sample of 87 university students revealed that 53 carried their books and notes in a backpack. Obtain a 95 % confidence interval for the population of students who use backpacks.

5. The Beverage Company has been experiencing problems with the automatic machine that places labels on bottles. A sample of 300 bottles resulted in 27 bottles with improperly applied labels. Using these data, develop a 90 % confidence interval for the population proportion of bottles with improperly applied labels.

6. If 65 persons in a random sample of 180 required lawyer services, then find and interpret 96 % confidence interval for proportion of persons in the population who required a lawyer services.

7. Let sample proportion $\hat{p} = 0.7$. How large a sample should be taken to be 95 % sure that the error of estimation does not exceed 0.02 when estimating a proportion?

8. A sample of 20 managers was taken and they were asked whether or not they usually take work home. The responses are given below:

Yes	Yes	No	No	No	Yes	No	No
No	No	Yes	Yes	No	Yes	Yes	No
No	No	No	Yes				

Make a 99 % confidence interval for the percentage of all managers who take work home.

Answers

1. a) Yes, sample size is large; b) No, sample size is not large; c) Yes, sample size is large; d) No, sample size is not large; 2. a) (0.687; 0.753); b) (0.681; 0.759); c) (0.668; 0.772); 3. a) (0.621; 0.839); b) (0.684; 0.776); c) (0.627; 0.833); 4. (0.506; 0.712); 5. (0.063; 0.117); 6. (0.286; 0.434); 7. $n = 2017$; 8. (0.117; 0.683).

6.7. Confidence intervals for the difference between means of two normal populations

Let μ_X be the mean of the first population and μ_Y be the mean of the second population. Suppose we want to make a confidence interval for the difference between these two population means, that is, $\mu_X - \mu_Y$.

Let \bar{X} be the mean of a sample from the first population and \bar{Y} be the mean of a sample taken from the second population. Then $(\bar{X} - \bar{Y})$ is the sample statistic that is used to make an interval estimate. We will consider several cases.

6.7.1. Confidence intervals for the difference between means: paired samples

In the case of two dependent samples, two data values-one in each sample-are collected from the same source and these are called **paired** or **matched pairs**.

Suppose that n matched pairs of observations, denoted by $(x_1, y_1); (x_2, y_2) \dots \dots (x_n, y_n)$, are selected from two populations with means μ_X and μ_Y .

Our aim is to find $100 \cdot (1 - \alpha)\%$ confidence interval for $(\mu_X - \mu_Y)$.

To find interval estimation we apply following steps:

1. Find n differences $d_i = x_i - y_i$

2. Find \bar{d}

3. Calculate S_d

If the population distribution of differences is assumed to be normal, then $100 \cdot (1 - \alpha)\%$ confidence interval for the difference between means is given by

$$\bar{d} - t_{n-1, \alpha/2} \cdot \frac{S_d}{\sqrt{n}} < \mu_X - \mu_Y < \bar{d} + t_{n-1, \alpha/2} \cdot \frac{S_d}{\sqrt{n}}$$

where $t_{n-1, \alpha/2}$ is the number for which

$$P(t_{n-1} > t_{n-1, \alpha/2}) = \frac{\alpha}{2}$$

The random variable t_{n-1} has a Student's t distribution with ($n - 1$) degrees of freedom.

Example:

A company claims that its special exercise program significantly reduces weight. A random sample of seven persons were put on exercise program. The following table gives the weights (in kg) of those seven persons before and after the program

Before	68	81	98	86	110	92	80
After	62	76	86	79	103	87	82

Make a 95 % confidence interval for the mean of the population paired differences. Assume that the population of paired differences is (approximately) normally distributed.

Solution:

Let d be the difference between the weights before and after the program. The necessary calculations are shown in the following table

Before	After	Difference d	d^2
68	62	6	36
81	76	5	25
98	86	12	144
86	79	7	49
110	103	7	49
92	87	5	25
80	82	-2	4
		$\sum d = 40$	$\sum d^2 = 332$

The values of \bar{d} and S_d are calculated as follows:

$$\bar{d} = \frac{\sum d}{n} = \frac{40}{7} = 5.71$$

$$S_d = \sqrt{\frac{1}{n-1} \left[\sum d^2 - n \cdot (\bar{d})^2 \right]} = \sqrt{\frac{1}{6} (332 - 7 \cdot 5.71^2)} = \sqrt{17.30} = 4.16.$$

Then

$$100 \cdot (1 - \alpha)\% = 90\%$$

$$1 - \alpha = 0.90$$

$$\alpha = 0.1 \quad \text{and}$$

$$\alpha/2 = 0.05$$

$$t_{n-1, \alpha/2} = t_{6, 0.05} = 1.943.$$

In the end, 90 % confidence interval for $(\mu_x - \mu_y)$ is

Weekly sales Before	Weekly sales After
15	18
12	14
18	19
15	18
16	18

$$5.71 - 1.943 \cdot \frac{4.16}{\sqrt{7}} < \mu_X - \mu_Y < 5.71 + 1.943 \cdot \frac{4.16}{\sqrt{7}}$$

$$2.6 < \mu_X - \mu_Y < 8.82$$

Thus, we can state with 90 % confidence that the mean difference between the weights before and after exercise program is between 2.6 and 8.82 kg.

Exercises

1. Find the following confidence interval for the difference between two population means assuming that the populations of paired differences are normally distributed

a) $n = 10$ $\bar{d} = 23.6$; $S_d = 12.6$; confidence level = 99 %

b) $n = 26$ $\bar{d} = 13.2$; $S_d = 4.8$; confidence level = 95 %

c) $n = 14$ $\bar{d} = 46.2$; $S_d = 13.6$; confidence level = 90 %

2. A company attempts to evaluate the potential for a new bonus plan by selecting a random sample of 5 salespersons to use the bonus plan for a week period. The weekly sales volumes before and after the bonus plan implementation shown below

Construct a 90 % confidence estimate for the mean increase in weekly sales that can be expected if a new bonus plan is implemented.

3. A company claims that the course they offer significantly increases the writing speed of secretaries. The following table gives the scores of eight secretaries before and after they attended this course.

Before	81	75	89	91	65	70	90	69
After	97	72	93	110	78	69	115	75

Make a 90 % confidence interval for the mean ($\mu_x - \mu_y$) of the population paired differences, where a paired differences is equal to the score before attending the course minus the score after attending the course.

4. A company sent 7 of its employees to attend a course in building self-confidence. The following table gives the scores of these employees before and after attending the course

Before	8	5	4	9	6	8	5
After	10	7	5	11	6	7	9

Construct a 95 % confidence interval for the mean of population paired differences where a paired difference is equal to the score of an employee before attending the course minus after attending this course.

Answers

1. a) (10.65; 36.55); b) (11.26; 15.14); c) (39.76; 52.64); 2. 1.40 to 3.00;
3. (-16.54; -3.21); 4. (-2.93; 0.07).

6.7.2. Confidence intervals for the difference between means of two normal populations with known variances

Suppose that the random variable \bar{X} is based on a random sample of size n_x from a normal population with mean μ_x and known variance σ_x^2 .

Also suppose that the random variable \bar{Y} is based on a random sample of size n_Y from a normal population with mean μ_Y and known variance σ_Y^2 . The difference between population means has a mean $(\mu_X - \mu_Y)$ and

$$\text{variance} \left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} \right).$$

Therefore, the random variable

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

has a standard normal distribution.

We can use this fact to obtain $100 \cdot (1 - \alpha)\%$ confidence interval for the difference between the population means.

Definition:

When the variances σ_X^2 and σ_Y^2 of two normal are known, then $100 \cdot (1 - \alpha)\%$ confidence interval for $(\mu_X - \mu_Y)$ is given by

$$(\bar{X} - \bar{Y}) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} < \mu_X - \mu_Y < (\bar{X} - \bar{Y}) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}.$$

Example:

A sample of size 13 from a normal population with variance 100 yielded $\bar{X} = 31.4$. A sample of size 7 from a second normal population with variance 80 yielded $\bar{Y} = 38.1$. Find a 95 % confidence interval for $(\mu_X - \mu_Y)$.

Solution:

$$100 \cdot (1 - \alpha)\% = 95\%$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05 \quad \text{and} \quad \alpha/2 = 0.025$$

$$z_{\alpha/2} = z_{0.025} = 1.96$$

The 95 % confidence interval for $(\mu_X - \mu_Y)$ is

$$(31.4 - 38.1) - 1.96 \cdot \sqrt{\frac{100}{13} + \frac{80}{7}} < \mu_X - \mu_Y < (31.4 - 38.1) + 1.96 \cdot \sqrt{\frac{100}{13} + \frac{80}{7}}$$
$$-15.27 < \mu_X - \mu_Y < 1.87$$

Remark:

When n_X and n_Y are both large, the normal approximation remains valid if σ_X^2 and σ_Y^2 are replaced by their estimators S_X^2 and S_Y^2 . When n_X and n_Y are greater than 30, an approximate $100 \cdot (1 - \alpha)\%$ confidence interval for $(\mu_X - \mu_Y)$ is given by

$$(\bar{X} - \bar{Y}) - z_{\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} < \mu_X - \mu_Y < (\bar{X} - \bar{Y}) + z_{\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

where $z_{\alpha/2}$ is the number for which

$$P(z > z_{\alpha/2}) = \alpha/2$$

and Z follows standard normal distribution.

Example:

A sample of 50 yogurt cups produced by the company showed that they contain an average of 146 calories per cup with a standard deviation of 6.4 calories. A sample of 60 such yogurt cups produced by its competitor showed that they contained an average of 143 calories per cup with a standard deviation of 7.2 calories. Make a 97 % confidence interval for the difference between the mean number of calories in yogurt cups produced by the two companies.

Solution:

We can refer to the respective samples as sample 1 and sample 2.

Let μ_X and μ_Y be the means of populations 1 and 2 respectively, and let \bar{X} and \bar{Y} be the means of the respective samples.
From the given information:

$$n_X = 50; \quad \bar{X} = 146; \quad S_X = 6.4$$

$$n_Y = 60; \quad \bar{Y} = 143; \quad S_y = 7.2$$

Since both sample sizes are large ($n_X > 30, n_Y > 30$) we can replace σ_X^2 and σ_Y^2 by S_X^2 and S_Y^2 respectively.

Then $100 \cdot (1 - \alpha)\%$ confidence interval for $(\mu_X - \mu_Y)$ is given by

$$(\bar{X} - \bar{Y}) - z_{\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} < \mu_X - \mu_Y < (\bar{X} - \bar{Y}) + z_{\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

$$1 - \alpha = 0.97$$

$$\alpha = 0.03 \quad \text{and} \quad \alpha/2 = 0.015$$

$$z_{\alpha/2} = z_{0.015} = 2.17$$

Finally, substituting all the values in the confidence interval formula, we obtain 97 % confidence interval for $(\mu_X - \mu_Y)$ as

$$(146 - 143) - 2.17 \cdot \sqrt{\frac{6.4^2}{50} + \frac{7.2^2}{60}} < \mu_X - \mu_Y < (146 - 143) + 2.17 \cdot \sqrt{\frac{6.4^2}{50} + \frac{7.2^2}{60}}$$

$$0.18 < \mu_X - \mu_Y < 5.82.$$

Thus, with 97 % confidence we can state that the difference in the mean calories of the two population of yogurt cups produced by two different companies is between 0.18 and 5.82.

Exercises

- 1.** A random sample of size 10 from a normal population with variance 50 gave a mean 43.2. A second random sample of size 18 from a normal

population with variance 72 gave a mean 48.7. Find a 99 per cent confidence interval for the difference between two population means.

2. A random sample of size 100 yielded the sample values $\bar{X} = 509$, $S_x^2 = 950$. A random sample size 100 from another population yielded

$\bar{Y} = 447$ $S_y^2 = 875$. Find a 95 % confidence interval for $(\mu_x - \mu_y)$.

3. An urban planning group is interested in estimating the difference between mean household incomes for two cities. Independent samples of households in two cities provide the following results:

City 1	City 2
$n_1 = 32$	$n_2 = 36$
$\bar{x}_1 = \$500$	$\bar{x}_2 = \$375$
$S_1 = \$150$	$S_2 = \$130$

Develop an interval estimate of the difference between mean incomes in the two cities. Show the results for confidence coefficients of 0.90 and 0.95.

4. The management at the National Bank investigates mean waiting time for all customers at its two branches. They took a sample of 200 customers from the branch A and found that they waited an average of 4.60 minutes with a standard deviation of 1.2 minutes before being served. Another sample of 300 customers taken from the branch B showed that these customers waited an average of 4.85 minutes with a standard deviation of 1.5 minutes before being served. Make a 97 % confidence interval for the difference between the two population means.

5. Rural and urban students are to be compared on the basis of their scores on a nationwide university entrance test. Two random samples of sizes 80 and 95 are selected from rural and urban students. The summary statistics from the test scores are

	Rural	Urban
Sample size	80	95
Mean	78.6	85.7
Standard deviation	9.1	8.3

Establish a 96 % confidence interval for the difference in population mean scores between urban and rural students.

6. A business consultant wanted to investigate if providing day care facilities on premises by companies reduces the absentee rate of working mothers from companies that provide day care facilities on premises. Sample of 50 mothers selected from the companies that provide day care facilities was taken. These mothers missed an average of 6.4 days from work last year with a standard deviation of 1.20 days. Another sample of 50 such mothers taken from companies that do not provide day care facilities on premises showed that these mothers missed an average of 9.3 days last year with a standard deviation of 1.83 days. Construct a 98 % confidence interval for the difference between the two population means.

Answers

- 1.** (-13.24; 2.24); **2.** (53.63; 70.37); **3.** (68.68; 181.32); (57.89; 192.11);
4. (-0.51; 0.01); **5.** (4.37; 9.83); **6.** (-3.62 to -2.18 days).

6. 9. Confidence interval for the difference between the population proportions: (large samples)

As it was discussed earlier, for a large sample the sample proportion \hat{p} is approximately normally distributed with mean p and standard deviation $\sqrt{\frac{p \cdot (1-p)}{n}}$.

Suppose that a random sample of size n_X observations from a population with proportion of “success” p_X has a sample proportion of success \hat{p}_X , and that an independent random sample of size n_Y observations from a population with proportion of “success” p_Y yields sample proportion \hat{p}_Y .

Since n_X and n_Y both are large, their sample proportions \hat{p}_X and \hat{p}_Y are approximately normally distributed with means p_X and p_Y , and standard deviations

$$\sqrt{\frac{p_X(1-p_X)}{n_X}} \text{ and} \\ \sqrt{\frac{p_Y(1-p_Y)}{n_Y}} \text{ respectively.}$$

Then the random variable $\hat{p}_X - \hat{p}_Y$, and the variance

$$\sigma_{\hat{p}_X - \hat{p}_Y}^2 = \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}.$$

The standardized random variable

$$Z = \frac{\left(\hat{p}_X - \hat{p}_Y\right) - \left(p_X - p_Y\right)}{\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}}$$

is approximately standard normal.

In order to find confidence interval for $(\hat{p}_X - \hat{p}_Y)$, we must either know or estimate the quantity of

$$\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}.$$

We can estimate the population proportion p_X by the sample proportion \hat{p}_X ; and we can estimate the population proportion p_Y by the sample proportion \hat{p}_Y . Then

$$\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y} \approx \frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}$$

Then an approximate $100(1-\alpha)\%$ confidence interval is given by

$$\left(\hat{p}_X - \hat{p}_Y \right) - z_{\alpha/2} \cdot A < p_X - p_Y < \left(\hat{p}_X - \hat{p}_Y \right) + z_{\alpha/2} \cdot A$$

where

$$A = \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}.$$

Example:

Mike and Tom like to throw darts. Mike throws 100 times and hits the target 54 times; Tom throws 100 times and hits the target 49 times. Find a 95 % confidence interval for $(p_X - p_Y)$, where p_X represents the true proportion of hits in Mike's tosses, and p_Y represents the true proportion of hits in Tom's tosses.

Solution:

$$100(1-\alpha)\% = 95\% \quad \text{and} \quad z_{0.025} = 1.96$$

$$\hat{p}_X = \frac{54}{100} = 0.54$$

$$\hat{p}_Y = \frac{49}{100} = 0.49$$

$$A = \sqrt{\frac{\hat{p}_X \cdot (1 - \hat{p}_X)}{n_X} + \frac{\hat{p}_Y \cdot (1 - \hat{p}_Y)}{n_Y}} = \sqrt{\frac{0.54 \cdot 0.46}{100} + \frac{0.49 \cdot 0.51}{100}} = 0.0706$$

$$(0.54 - 0.49) - 1.96 \cdot 0.0706 < p_X - p_Y < (0.54 - 0.49) + 1.96 \cdot 0.0706$$

$$-0.088 < p_X - p_Y < 0.188$$

Thus, with 95 % confidence we can state that the difference between the proportions of Mike's and Tom's tosses is between -0.088 and 0.188.

Exercises

1. Find a 90 % confidence interval for $p_X - p_Y$, if a sample of size 200 yielded $\hat{p}_X = 0.70$ and a sample of size 300 yielded $\hat{p}_Y = 0.65$.

2. Construct a 99 % confidence interval for $p_X - p_Y$ if

$$n_X = 300; \quad \hat{p}_X = 0.53$$

$$n_Y = 200; \quad \hat{p}_Y = 0.59$$

3. A sample of 400 observations taken from the first population gave $x_1 = 150$. Another sample of 700 observations taken from the second population gave $x_2 = 225$. Make a 96 % confidence interval for $p_1 - p_2$.

4. A sample of 500 items produced by a supplier A possessed 270 defective items. A random sample of 360 items produced by supplier B possessed 162 defective items. Compute a 95 % confidence interval estimate for the difference in proportion defective from the two suppliers.

5. Assume that 66 % of single women and 81.9 % of single men own cars. Also assume that these estimates are based on random samples of 1640 single women and 1800 single men. Develop a 99 % confidence interval for the difference between the two population proportions.

6. The management of a market wanted to investigate if the percentage of men and women who prefer to buy national brand products over the store

brand products are different. A sample of 500 men shoppers at supermarkets showed that 175 of them prefer to buy national brand products over the store brand products. Another sample of 800 women shoppers showed that 360 of them prefer to buy national brand products over the store brand products. Construct a 95 % confidence interval for the difference between the proportions of all men and women shoppers at supermarket who prefer to buy national brand products over the store brand products.

7. A sample of 600 females was selected from ethnic group A and a sample of 700 from ethnic group B. Each female was asked “Did you get married before you were 22?”. 246 of females from group A and 266 of females from group B answered “yes”. Find a 95 % confidence interval for the two population proportions.

8. According to a survey, 1010 adults conducted and 74.2 % of male and 88.8 % of women said that they are concerned about living near a nuclear power plant. Assume that there were 520 men and 490 women in this sample.

Construct a 99 % confidence interval for the difference between the proportions of all men and all women who are concerned about living near power plant.

Answers

1. (-0.02; 0.12); **2.** (-0.18; 0.06); **3.** (-0.006; 0.116); **4.** (0.023; 0.157);
5. (-0.20; -0.12); **6.** (-0.154; 0.046); **7.** (-0.02; 0.08); **8.** (-0.21; -0.08).

6.11. Sample size determination

The reason why we always conduct a sample observations and not a census is that almost always we have limited resources at our disposal. In our calculations, if a smaller sample can serve our purpose, then we will be wasting our resources by taking a larger sample. For example, suppose we want to estimate the mean life of certain type of lights bulbs. If a sample of 50 light bulbs can give us the type of confidence interval that we are looking for, then we will be wasting money and time if we take a sample of much larger size, say 800 light bulbs. In such cases if we know the confidence

interval that we want, then we can find the (approximate) size of the sample that will produce the required result.

6.11.1. Sample size determination for the estimation of mean

Suppose that sample of n observations is taken from a normally distributed population with mean μ and known variance σ^2 . We know that

100·(1− α)% confidence interval for the population mean μ is given by

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where \bar{x} is the sample mean and $z_{\alpha/2}$ is the appropriate cutoff point of the standard normal distribution. This confidence interval is centered on the sample mean and extends a distance of L , the margin of error (also called the sample error, the bound, or the interval half width) is given by

$$L = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Suppose that we predetermine the size of L and want to find the size of the sample that will yield this margin error. From the above expression, the following formula is obtained that determines the required sample size n .

Definition:

Given the confidence level and standard deviation of the population (or population variance), the sample size that will produce a predetermined margin error L of the confidence interval estimate of μ is

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{L^2}$$

Remark 1:

If we do not know σ , we can take a sample and find sample standard deviation. Then we can use S for σ in the formula.

Remark 2:

n must be rounded to the next higher integer, because a sample size can not be fractional.

Example:

Suppose that we want to estimate the mean family size for all country families at 99 % confidence level. It is known that the standard deviation σ for the sizes of all families in the country is 0.45.

How large a sample should we select if we want its estimate to be within 0.02 of the population mean?

Solution:

We want the 99 % confidence interval for the mean family size to be

$$\bar{x} \pm 0.02.$$

Hence, the margin of errors is to be 0.02, that is

$$L = 0.02$$

The value of $z_{\alpha/2}$ for a 99 % confidence level is 2.58.

The value of σ is given to be 0.45. Therefore, substituting all values in the formula and simplifying, we obtain

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{L^2} = \frac{(2.58)^2 \cdot (0.45)^2}{(0.02)^2} = 3369.8 \approx 3370$$

Thus, the required sample size is 3370. If we will take a sample of 3370 families, compute the mean family size for this sample, and then margin of a 99 % confidence interval around this sample, the margin of error of the estimate will be approximately 0.02.

6.11.2. Sample size determination for the estimation of proportion

Just as we did with the mean, we can also determine the sample size for estimating the population proportion p .

We know that $100 \cdot (1 - \alpha)\%$ confidence interval for p is given by

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where \hat{p} - is the sample proportion.

This interval is centered on the sample proportion and extends a distance L :

$$L = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

This result can not be used directly to determine the sample size n necessary to obtain a confidence interval of some specific width, since it involves \hat{p} , which is not known. But whatever the outcome, $\hat{p}(1 - \hat{p})$ can not be bigger than 0.25, its value when the sample proportion is 0.5. Thus, the largest possible value for L is given by

$$L = z_{\alpha/2} \cdot \sqrt{\frac{0.25}{n}} = \frac{0.5 \cdot z_{\alpha/2}}{\sqrt{n}}$$

Using basic algebra, we obtain

$$\sqrt{n} = \frac{0.5 \cdot z_{\alpha/2}}{L}$$

and squaring yields

$$n = \frac{0.25 \cdot (z_{\alpha/2})^2}{L^2}$$

Definition:

Let a random sample be selected from a normal population. $100 \cdot (1 - \alpha)\%$ confidence interval for the population proportion, extending a distance of at most L on each side of the sample proportion, can be guaranteed if the sample size is

$$n = \frac{0.25 \cdot (z_{\alpha/2})^2}{L^2}$$

Example:

A public health survey is to be designed to estimate the proportion p of a population having defective vision. How many persons should be examined if the public health doctor wishes to be 98 % certain that error of estimation is below 0.05?

Solution:

The public health doctor wants the 98 % confidence interval to be $\hat{p} \pm 0.05$

Therefore $L = 0.05$. The value of $z_{\alpha/2}$ for a 98 % confidence level is 2.33.

After substituting we obtain that the required sample size is

$$n = \frac{0.25 \cdot (z_{\alpha/2})^2}{L^2} = \frac{0.25 \cdot (2.33)^2}{(0.05)^2} = 542.89 \approx 543.$$

Thus, if the doctor takes a sample of 543 persons, the estimate of p will be within 0.05 of the population proportion.

Exercises

1. Determine the sample size for the estimate of μ for the following:

- a) $L = 0.17$; $\sigma = 0.90$; confidence level = 99%
- b) $L = 4.1$; $\sigma = 23.45$; confidence level = 95%
- c) $L = 25.9$; $\sigma = 122.25$; confidence level = 90%

2. Determine the most conservative sample size for estimation of the population proportion for the following:

- a) $L = 0.025$; confidence level = 99 %
- b) $L = 0.045$; confidence level = 96 %
- c) $L = 0.015$; confidence level = 90 %

3. A sample of 50 workers' average weekly earnings gave $\sigma = \$35$.

Determine the sample size that is needed for estimating the population mean weekly earnings with a 98 % error margin of \$ 3.50.

4. How large a sample should be taken to be 95 % sure that the error of estimation does not exceed 0.02 when estimating a population proportion?

5. A food service manager wants to be 95 % confident that the error in the estimate of the mean number of sandwiches dispensed over the lunch hour is 10 or less. What sample size should be selected if $\sigma = 40$?

6. One department manager wants to estimate at 90 % confidence level the mean amount spent by all customers at this store. He knows that the standard deviation of amounts spent by customers at this store is \$ 27. What sample size he chooses so that the estimate is within \$ 3 of the population mean?

7. A teacher wants to estimate the proportion of all students who own mobile telephones. How large should the sample size be so that the 99 % confidence interval for the population proportion has a maximum error of 0.03?

8. A private university wants to determine a 99 % confidence interval for the mean number of hours that students spend per week doing homework. How large a sample should be selected so that the estimate is within 1 hour of the population mean? Assume that the standard deviation for the time spent per week doing homework by students is 3 hours.

Answers

1. a) 186; b) 126; c) 61; **2.** a) 2653; b) 437; c) 3007; **3.** 543; **4.** 2401; **5.** 62;
6. 220; **7.** 1842; **8.** 60.

Chapter 7

Hypothesis testing

7.1. Introduction

Inferential statistics consists of methods that use sample results to help make decisions or predictions about a population. The point and interval estimation procedures are forms of statistical inference. Another type of statistical inference is hypothesis testing. In hypothesis testing we begin by stating a hypothesis about a population characteristic. This hypothesis, called the **null hypothesis**, is assumed to be true unless sufficient evidence can be found in a sample to reject it. The situation is quite similar to that in a criminal trial. The defendant is assumed to be innocent; if sufficient evidence to the contrary is presented, however, the jury will reject this hypothesis and conclude that the defendant is guilty.

In statistical hypothesis testing, often the null hypothesis is an assumption about the value of a population parameter. A sample is selected from the population, and a point estimate is computed. By comparing the value of the point estimate to the hypothesized value of the parameter we draw a conclusion with respect to whether or not there is a sufficient evidence to reject the null hypothesis. A decision is made and often a specific action is taken depending upon whether or not the null hypothesis about the population parameter is accepted or rejected.

7.1.2. Concepts of hypothesis testing

Let us consider example about coffee cans. A company may claim that, on average, its cans contain 100 grams of coffee. A government agency may want to test whether or not such cans contain, on average, 100 grams of coffee.

Suppose we take a sample of 50 cans of the coffee under investigation. We then find out that the mean amount of coffee in these 50 cans is 97 grams.

Based on these results, can we state that on average, all such cans contain less than 100 grams of coffee and that the company is lying to the public?

Not until we perform a test of hypothesis. The reason is that the mean $\bar{x} = 97$ grams is obtained from the sample. The difference between 100 grams (the required amount for the population) and 97 grams (the observed average amount for the sample) may have occurred only because of the sampling error. Another sample of 100 cans may give us a mean of 105 grams. Therefore, we make a test of hypothesis to find out how large the difference between 100 grams and 97 grams is and to investigate whether or not this difference has occurred as a result of chance alone. If 97 grams is the mean of all cans and not for only 100 cans, then we do not need to make a test of hypothesis. Instead, we can immediately state that the mean amount of coffee in all such cans is less than 100 grams. We perform a test of hypothesis only when we are making a decision about a population parameter based on the value of a sample statistic.

7.1.3. The null and alternative hypothesis

We will begin our general discussion by using θ to denote a population probability distribution parameter of interest, such as the mean, variance, or proportion. Our discussion begins with a hypothesis about the parameter that will be maintained unless there is strong contrary evidence. In statistical language it is called the **null hypothesis**.

For example, we might initially accept company's claim that on average, the contests of the cans weight at least 100 grams. Then after

collecting sample data this hypothesis can be tested. If the null hypothesis is not true, then some alternative must be true. In carrying out a hypothesis test the investigator defines an **alternative hypothesis** against which the null hypothesis is tested.

For this coffee cans example a likely alternative is that on average can's weights are less than 100 grams. These hypotheses are chosen such that one or the other must be true. The null hypothesis will be denoted as H_0 and the alternative hypothesis as H_1 .

Definition: A **null hypothesis** is a claim (or statement) about a population parameter that is assumed to be true until it is declared false.

Definition: An **alternative hypothesis** is a claim about population parameter that will be true if the null hypothesis is false.

Our analysis will be designed with the objective of seeking strong evidence to reject the null hypothesis and accept the alternative hypothesis. We will only reject the null hypothesis when there is a small probability that the null hypothesis is true. Thus rejection will provide strong evidence against H_0 and in favor of the alternative hypothesis, H_1 . If we fail to reject H_0 then either H_0 is true or our evidence is not sufficient to reject H_0 and hence accept H_0 . Thus we will be more comfortable with our decision if we reject H_0 and accept H_1 .

A hypothesis, whether null or alternative, might specify a single value, say θ_0 , for the population parameter θ . In that case, the hypothesis is said to be a simple hypothesis designated as

$$H_0 : \theta = \theta_0$$

That is read as, "The null hypothesis is that the population parameter θ is equal to the specific value θ_0 ".

Alternatively, a range of values might be specified for unknown parameter. We define such hypothesis as a composite hypothesis, and it will hold true for more than one value of the population parameter. In many applications, a simple null hypothesis, say

$$H_0 : \theta = \theta_0$$

is tested against a composite alternative. One possibility would be to test the null hypothesis against the general two-sided hypothesis

$$H_1 : \theta \neq \theta_0$$

In other cases, only alternatives on one side of the null hypothesis are of interest. For example, a government agency would be perfectly happy if the mean weight of coffee cans greater than 100 grams. Then we could write the null hypothesis as

$$H_0 : \theta \geq \theta_0$$

and the alternative hypothesis of interest might be

$$H_1 : \theta < \theta_0$$

We call these hypothesis one- sided composite alternatives.

Example:

A company intends to accept the product unless it has evidence to suspect that more than 10% of products are defective. Let θ denote the population proportion of defectives. The null hypothesis is that the proportion is less than 0.1, that is

$$H_0 : \theta \leq 0.1$$

and the alternative hypothesis is

$$H_1 : \theta > 0.1$$

The null hypothesis is that the product is of adequate quality overall, while the alternative is that the product is not adequate quality. In this case the product would only be rejected if there is strong evidence that there are more than 10% defectives.

Once we have specified a null hypothesis and alternative hypothesis and collected sample data, a decision concerning the null hypothesis must be made. We can either accept the null hypothesis or reject it in favor of the alternative. For good reasons many statisticians prefer not to use the term “accept the null hypothesis” and instead say “fail to reject”. When we accept or fail to reject the null hypothesis, then either the hypothesis is true or our test procedure was not strong enough to reject and we have committed an error. When we use the term **accept a null hypothesis** that statement can be considered shorthand for failure to reject.

From our discussion of sampling distributions, we know that the sample mean is different from the population mean. With only a sample mean we can not be certain of the value of the population mean. Thus the decision rule we adopt will have some chance of reaching an erroneous conclusion. One error we call Type I error. **Type I error** is defined as the rejection of the null hypothesis when the null hypothesis is true. We will see that our decision rules will be defined so that the probability of rejecting a true null hypothesis, denoted as α , is “small”. The probability, α , is defined as the **significance level** of the test. Since the null hypothesis is either accepted or rejected, it follows that the probability of accepting the null hypothesis when it is true is $(1 - \alpha)$. The other possible error, called **Type II error**, arises when false null hypothesis is accepted. We say that for a particular decision rule, the probability of making such an error when the null hypothesis is false is denoted β . Then, the probability of rejecting a false null hypothesis is $(1 - \beta)$ which is called the power of test.

Type I error

A type I error occurs when a true null hypothesis is rejected. The value α represents the probability of committing this type of error, that is

$$\alpha = P(H_0 \text{ is rejected} / H_0 \text{ is true})$$

The value α represents the significance level of the test.

Type II error

A Type II error occurs when a false null hypothesis is not rejected. The value β represents the probability of committing a Type II error,

that is

$$\beta = P(H_0 \text{ is not rejected} / H_0 \text{ is false})$$

The value $(1 - \beta)$ is called the power of the test. It represents the probability of not making a Type II error.

7.1.4. Tails of the test

In statistics, the rejection region for a hypothesis testing problem can be on both sides with non rejection region in the middle, or it can be on the left side or in the right side of the non rejection region. A test with two rejection regions is called a **two tailed test**, and a test with one rejection region is called

a **one tailed test**. The one tailed test is called a **left tailed test** if the rejection region is in the left tail of the distribution curve, and a **right tailed test** if the rejection region is in the right tail of the distribution curve.

a) A two tailed test

Example:

The mean family size in a particular country was 3.75 in 1990. We want to check whether or not this mean has changed since 1990. The mean family size has changed if it has either increased or decreased during this period. This is an example of two tailed test. Let μ be the current mean family size for all families. We write the null and alternative hypothesis for this test as

$$H_0 : \mu = 3.75 \text{ (The mean family size has not changed)}$$

$$H_1 : \mu \neq 3.75 \text{ (The mean family size has changed)}$$

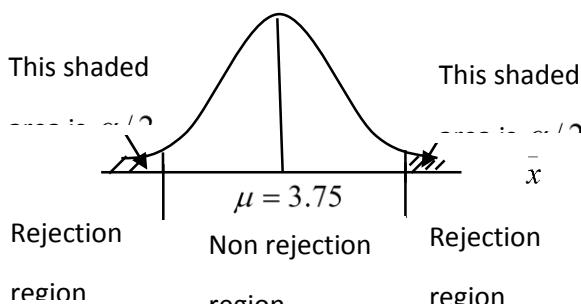


Fig.1.1

As shown in Figure 1.1, a two tailed test has two rejection regions, one in each tail of the distribution curve.

b) A left tailed test

Reconsider the example of the mean amount of coffee can produced by company. The company claims that these cans, on average, contain 100 grams of coffee. However, if these cans contain less than the claimed amount of coffee, then the company can be accused of cheating. Suppose that the government agency wants to test whether the amount of coffee can is less than 100 grams. Note that the key phrase this time is *less than*, which indicates a left tailed test. Let μ be the mean amount of coffee in all cans. The null and alternative hypothesis for this test are written as

$$H_0 : \mu = 120 \text{ grams} \quad (\text{The mean is not less than 120 grams})$$

$$H_1 : \mu < 120 \text{ grams} \quad (\text{The mean is less than 120 grams})$$

In this case, we can also write the null hypothesis as $H_0 : \mu \geq 120$ grams . This will not affect the result of the test as long as the sign in H_1 is *less than*.

When the alternative hypothesis has a *less than* ($<$) sign, as in this case, the test is always left tailed. In a left tailed test the rejection region is always in the left tail of the distribution curve, as shown in Figure 1.2.

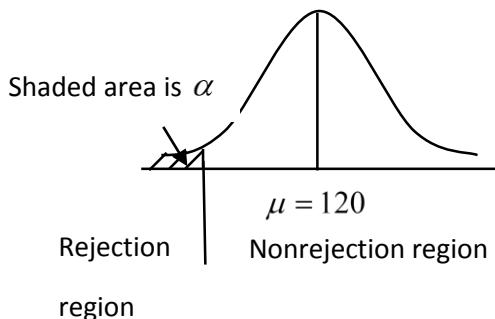


Fig.1.2

c) A right tailed test

Suppose that mean monthly income of all households was 45 500 tg in 2001. We want to test if current income of all households is higher than 45 500 tg. The key phrase in this case is *higher than*, which indicates a right tailed test.

Let μ be the mean income of all households.

We write the null and alternative hypothesis for this test as

$$H_0 : \mu = 45500 \text{ (The current income is not higher than 45 500 tg)}$$

$$H_1 : \mu > 45500 \text{ (The current income is higher than 45 500 tg)}$$

In this case, we can also write the null hypothesis as $H_0 : \mu \leq 45500$, which states that current mean income is either equal to or less than 45 500 tg. Again, the result of the test will not be affected whether we use an *equal to* ($=$) or a *less or equal to* (\leq) sign in H_0 as long as the alternative hypothesis has a *greater than* ($>$) sign.

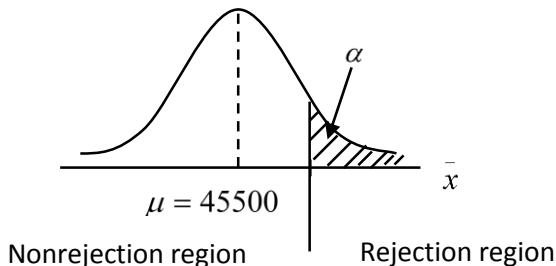


Fig.1.3

When an alternative hypothesis has a *greater than* ($>$) sign, the test is always right tailed. As shown in the Fig. 1.3, in a right tailed test, the rejection region is in the right tail of the distribution curve. The area of this rejection region is equal to α , the significance level. We will reject H_0 if the value of \bar{x} obtained from the sample falls in the rejection region. Otherwise, we will not reject H_0 .

Remark: Note that the null hypothesis always has an *equal to* ($=$) or a *less or equal to* (\leq) or a *greater than or equal to* (\geq) sign and the alternative hypothesis always has a *not equal to* (\neq) or a *greater than* ($>$) or a *less than* ($<$) sign.

Exercises

1. Explain which of the following is a two tailed test, a left tailed test, or a right tailed test.

a) $H_0 : \mu = 25,$ $H_1 : \mu < 25$

b) $H_0 : \mu \leq 134,$ $H_1 : \mu > 134$

$$\text{c) } H_0 : \mu = 16, \quad H_1 : \mu \neq 16$$

Show the rejection and nonrejection regions for each of these cases by drawing a sampling distribution curve for the sample mean, assuming that sample size is large in each case.

2. Consider $H_0 : \mu = 35$, against $H_1 : \mu < 35$.

- a) What type of error would you make if the null hypothesis is actually false and you fail to reject it?
- b) What type of error would you make if the null hypothesis is actually true and you reject it?

3. For each of the following rejection regions, sketch the sampling distribution for z and indicate the location of rejection region.

- a) $z > 2.05$;
- b) $z > 2.75$;
- c) $z < -1.28$;
- d) $z < -2.13$;
- e) $z < -2.575$ or $z > 2.575$;
- g) $z < -1.82$ or $z > 1.82$

4. Write the null hypothesis and alternative hypothesis for each of the following examples. Determine if each is a case of a two tailed, a left tailed, or a right tailed test.

- a) To test whether or not the mean price of houses in a certain city is greater than \$ 45 000.
- b) To test if the mean number of hours spent working per week by students who hold jobs is different from 18 hours.
- c) To test whether the mean life of a particular brand of auto batteries is less than 28 days.
- d) To test if the mean amount of time taken by all workers to do a certain job is more than 45 minutes.

- f) To test the mean age of all managers of companies is different from 40 years.
 - g) To test the mean time for an airline passenger to obtain his or her luggage, once luggage starts coming out the conveyer belt, is less than 180 seconds.

7.2. Tests of the mean of a normal distribution:

Population variance known

In this and following sections we will present specific procedures for developing and implementing hypothesis test procedures with applications to business and economic problems.

We are given a random sample of n observations from a normal population with mean μ and known variance σ^2 . If the observed sample mean is \bar{x} , then the test statistic is

$$T.S. = z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

and we can use the following tests with significance level α .

1. To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \leq \mu_0$ against the

$$H_1: \mu > \mu_0$$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha}$

2. To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \geq \mu_0$ against the alternative

$$H_1 : \mu < \mu_0$$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$H_0 : \mu = \mu_0$ against the two sided alternative

$$H_1 : \mu \neq \mu_0$$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$,

where $z_{\alpha/2}$ is the number for which

$$P(Z > z_{\alpha/2}) = \alpha/2$$

and Z is the standard normal distribution.

A statistical test of hypothesis procedure contains the following five steps:

1. State the null and alternative hypothesis
2. Select the distribution to use
3. Determine the rejection and nonrejection regions
4. Calculate the value of the test statistic
5. Make a decision.

Example:

A manufacturer of detergent claims that the content of boxes sold weigh on average at least 160 grams. The distribution of weights is known to be normal, with standard deviation of 14 grams. A random sample of 16 boxes yielded a sample mean weight of 158.9 grams. Test at the 10% significance level the null hypothesis that the population mean is at least 160 grams.

Solution:

Let μ be the mean average of all boxes and \bar{x} be the corresponding mean for the sample.

$$n = 16; \quad \sigma = 14; \quad \bar{x} = 158.9$$

The significance level is α is 0.1. That is, the probability of rejecting the null hypothesis when it is actually true should not exceed 0.1. This is the probability of making a Type I error. We perform the test of hypothesis using the five steps as follows.

Step 1. State the null and alternative hypothesis

We write the null and alternative hypothesis as

$$H_0 : \mu \geq 160 \text{ grams}$$

$$H_1 : \mu < 160 \text{ grams}$$

Step 2. Select the distribution to use

Since population standard deviation is known we will use $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$.

Step 3. Determine the rejection and nonrejection regions

The significance level is 0.1. The $<$ sign indicates that the test is left tailed. We look for 0.9 from in the standard normal distribution table, (Table 1 of Appendix). The value of z is -1.28 . (Fig. 1.4).

Step 4. Calculate the value of the test statistic

The decision to reject or not to reject the null hypothesis will depend on whether the evidence from the sample falls in the rejection or nonrejection region. If the value of the sample mean \bar{x} falls in rejection region, we reject H_0 . Otherwise we do not reject the null hypothesis. To locate the position of $\bar{x} = 158.9$ on the sampling distribution curve of \bar{x} in Figure 1.4 we first calculate z value for $\bar{x} = 158.9$. This is called the *value of the test statistic*.

$$T.S. = z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{158.9 - 160}{14 / \sqrt{16}} = -0.31$$

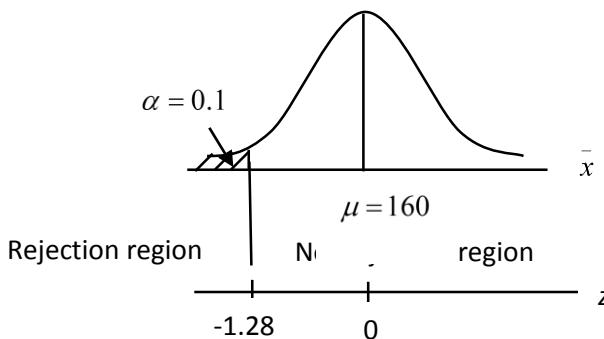


Fig.1.4

Step 5. Make a decision

In the final step we make a decision based on the value of the test statistic $T.S. = z$ for \bar{x} in previous step. This value of $z = -0.31$ is not less than the critical value of $z = -1.28$, and it falls in the nonrejection region. Hence we accept H_0 and conclude that based on sample information, it appears that the mean weight of all boxes is greater than 160 grams.

By accepting the null hypothesis we are stating that the difference between the sample mean $\bar{x} = 158.9$ and the hypothesized value of the population mean $\mu = 160$ is not too large and may occurred because of the chance or sampling error. There is a possibility that the mean weight is less than 160 grams, by the luck of the draw, we selected a sample with a mean that is not too far from required mean of 160 grams.

7.3. Tests of the mean of a normal distribution:

Population variance unknown (Large sample size)

When the population standard deviation is unknown, we simply estimate σ with the value of the sample standard deviation s . We must consider separately the large sample ($n \geq 30$) and small sample size ($n < 30$) cases.

If the sample size n is large, the test procedure developed for the case when population variance is known can be employed when it is unknown, replacing σ^2 by the observed sample variance s^2 . All the hypotheses and decision rules are stated in the same way as before (i.e. when σ^2 is known).

Example:

When a machine that is used to make bolts is working properly, the mean length of these bolts 2.5 cm. However, from time to time this machine falls out of alignment and produces bolt that have a mean length of either less than or 2.5 cm or more than 2.5 cm. When this happens, the process is stopped and the machine is adjusted. To check whether or not the machine is producing bolts with a mean length of 2.5 cm, the quality control department at the company takes a sample of bolts each week and makes a test of hypothesis. One such a sample of 49 bolts produced a mean length of

2.49 cm and a standard deviation of 0.021 cm. Using the 5% significance level, can we conclude that the machine needs to be adjusted?

Solution:

Let μ be the mean length bolts made on this machine and \bar{x} be the corresponding mean for the sample.

$$n = 49; \quad \bar{x} = 2.49 \text{ cm}; \quad s = 0.021 \text{ cm}$$

The mean length of all bolts is supposed to be 2.5 cm. The significance level is α is 0.05. That is, the probability of rejecting the null hypothesis when it is actually is true should not exceed 0.05.

Step 1. State the null and alternative hypothesis

We are testing to find whether or not the machine needs to be adjusted. The machine will need an adjustment if the mean length of these bolts is either less than 2.5 cm or more than 2.5.

We write the null and alternative hypothesis as

$$H_0 : \mu = 2.5 \text{ cm} \text{ (The machine does not need adjustment)}$$

$$H_1 : \mu \neq 2.5 \text{ cm} \text{ (The machine needs an adjustment)}$$

Step 2. Select the distribution to use

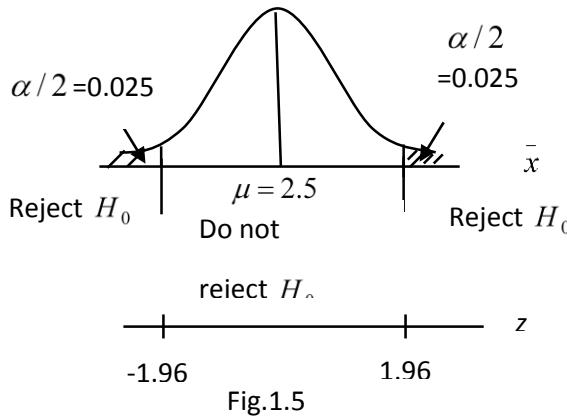
Because the sample size is large ($n > 30$), the sampling distribution of \bar{x} is (approximately) normal. Consequently we will use $z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ to make the test.

Step 3. Determine the rejection and nonrejection regions

The significance level is 0.05. The \neq sign indicates that the test is two tailed with two rejection regions, one in each tail of the normal distribution curve of \bar{x} . Because the total area of both rejection regions is 0.05 (the significance level), the area of the rejection region in each tail is 0.025.

These areas are shown in Fig.1.5. To find the z values for these critical points, we look for 0.975 in the standard normal distribution table.

Hence, the z values of the two critical points as shown in Fig.1.5, are -1.96 and 1.96 .



Step 4. Calculate the value of the test statistic

The value of \bar{x} from the sample is 2.49. As σ is not known, we calculate the z value as follows

$$T.S. = z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{2.49 - 2.5}{0.003} = -3.33$$

$z = -3.33$ is the value of the test statistic.

Step 5. Make a decision

The value of $z = -3.33$ is less than the critical value of $z = -1.96$, and it falls in the rejection region in the left tail. Hence we reject H_0 and conclude that based on sample information, it appears that the mean length of all bolts produced on this machine is not equal to 2.5 cm. Therefore, the machine needs to be adjusted

By rejecting the null hypothesis we are stating that the difference between the sample mean $\bar{x} = 2.49$ and the hypothesized value of the population mean $\mu = 2.5$ is too large and may not have occurred because of the chance or sampling error. This difference seems to be real and, hence the mean length of bolts is different from 2.5 cm. Note that the rejection of the null hypothesis does not necessarily indicate that the mean length of bolts is definitely different from 2.5 cm. It simply indicates that there is strong evidence (from sample) that the mean length of bolts is not equal to 2.5 cm.

There is a possibility that the mean length of bolts equal to 2.5 cm. If so, we have wrongfully rejected the null hypothesis H_0 . This is Type I error and probability of making such an error in this case is 0.05.

Exercises

1. Make the following tests of hypotheses.

a) $H_0 : \mu = 25$; $H_1 : \mu \neq 25$; $n = 81$; $\bar{x} = 28$; $s = 3$; $\alpha = 0.01$

b) $H_0 : \mu = 12$; $H_1 : \mu < 12$; $n = 45$; $\bar{x} = 11$; $\sigma = 4.5$; $\alpha = 0.05$

c) $H_0 : \mu = 40$; $H_1 : \mu > 40$; $n = 100$; $\bar{x} = 46$; $s = 7$; $\alpha = 0.1$

2. Consider $H_0 : \mu = 100$; against the two sided alternative $H_1 : \mu \neq 100$.

a) A random sample of 64 observations produced a sample mean of 98 and a standard deviation of 12. Using $\alpha = 0.01$, would you reject the null hypothesis?

b) Another random sample of 64 observations taken from the same population produced a sample mean of 104 and a standard deviation of 10. Using $\alpha = 0.01$, would you reject the null hypothesis?

Comment on the results of parts a) and b).

3. A survey showed that people with a bachelor's degree earned average of \$2116 a year in 2001. A sample of 900 persons with a bachelor's degree taken recently by a researcher showed that the persons in this sample earned on average of \$2345 a year with a standard deviation of \$210. Test at 5% significance level whether people with a bachelor's degree currently earn an average of \$2116 against the alternative that it is more than \$2116 in a year.

4. The manufacturer of a certain brand of auto batteries claims that the mean life of these batteries is 45 month. A consumer protection agency that wants to check this claim took a random sample of 36 such batteries and found the mean life for this sample is 43.75 month with a standard deviation of 4 month. Using the 2.5% significance level, test the manufacturer claim against the alternative that the mean life of batteries is less than 45 month.

5. A random sample of 100 observations from a population with standard deviation 60 yielded a sample mean of 110.

a) Test the null hypothesis that $\mu = 100$ against the alternative hypothesis that $\mu > 100$ using $\alpha = 0.05$. Interpret the results of the test.

b) Test the null hypothesis that $\mu = 100$ against the alternative hypothesis that $\mu \neq 100$ using $\alpha = 0.05$. Interpret the results of the test.

c) Compare the results of the two tests you conducted. Explain why the results differ.

6. In a random sample of 250 observations, the mean and standard deviation are found to be 169.8 and 31.6, respectively. Is the claim that μ larger than 169 substantiated by these data at the 10% level of significance?

7. From records, it is known that the duration of treating a disease by a standard therapy has a mean of 15 days. It is claimed that a new therapy can reduce the treatment time. To test this claim, the new therapy is tried on 70 patients, and from the data of their times to recovery, the sample mean and standard deviation are found to be 14.6 and 3.0 days, respectively.

Perform the hypothesis test using a 2.5% level of significance.

8. Suppose that you are to verify the claim that $\mu > 20$ on the basis of a random sample of size 70, and you know that $\sigma = 5.6$.

a) If you set the rejection region to be $\bar{x} > 21.31$, what is the level of significance of your test?

b) Find the numerical value of c so that the test $\bar{x} \geq c$ has a 5% level of significance.

Answers

1. a) $T.S.=9.00$; reject H_0 ; b) $T.S.=-1.49$; do not reject H_0 ; c) $T.S.=8.57$; reject H_0 ; 2. a) $T.S.=-1.33$; do not reject H_0 ; b) $T.S.=3.20$; reject H_0 ; 3. $T.S.=32.71$; reject H_0 ; 4. $T.S.=-1.87$; accept H_0 ; 5. a) $z=1.67$; reject H_0 ; b) $z=1.67$; accept H_0 ; 6. $T.S.=0.4$; accept H_0 ; 7. $z=-1.12$; H_0 is not rejected at $\alpha=0.025$; 8. a) $\alpha=0.025$; b) $c=21.10$.

7.4. Hypothesis testing using the p –value approaches

In previous section, the value of the significance level α was selected before the test performed. Sometimes we may prefer not to predetermine α . Instead, we may want to find a value such that a given null hypothesis will be rejected for any α greater than this value and it will not be rejected for any α smaller than this value. In this approach, we calculate the p -value for the test, which is defined as the smallest level of significance at which the given null hypothesis is rejected.

Definition:

The p -value is the smallest significance level at which the null hypothesis is rejected.

Using the p -value approach, we reject the null hypothesis if

$$p\text{- value} < \alpha$$

and we do not reject the null hypothesis if

$$p\text{- value} \geq \alpha$$

Steps necessary for calculating the p -value for a test of hypothesis

1. Determine the value of the test statistic $T.S. = z$ corresponding to the result of the sampling experiment.

2.

a) If the test is one-tailed, the p -value is equal to the tail area beyond z in the same direction as the alternative hypothesis. Thus, if the alternative hypothesis is of the form $>$, the p -value is the area to the right of, or above, the observed z value. Conversely, if the alternative is of the form $<$, the

p -value is the area to the left of, or below, the observed z value. (Fig.1.6;1.7)

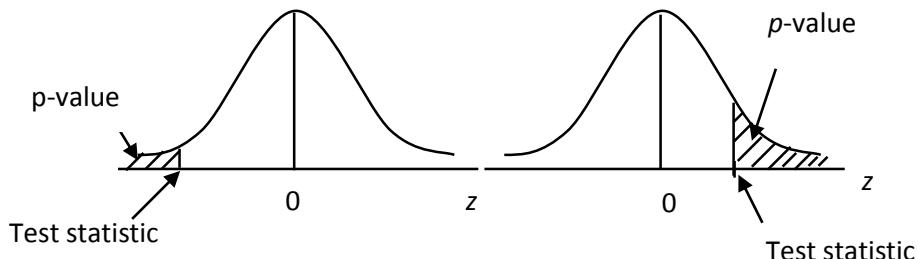


Fig.1.6. Left tailed test

Fig.1.7. Right tailed test

b) If the test is two tailed, the p -value is equal to twice the area beyond the observed z -value in the direction of the sign of z . That is, if z is positive, the p -value is twice the area to the right of, or above, the observed z -value. Conversely, if z is negative, the p -value is twice the area to the left of, or below, the observed z -value. (See Fig.1.8)

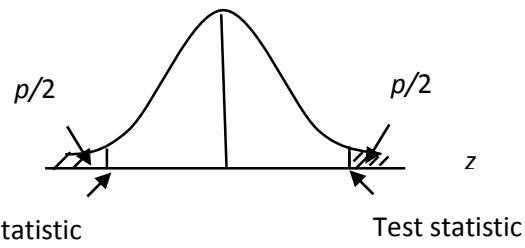


Fig.1.8. Finding the p -value for a two tailed test

Example:

The management of Health club claims that its members lose an average of 10kg or more within the first month after joining the club. A random sample of 36 members of this health club was taken and found that they lost an average of 9.2 kg within the first month of membership with standard deviation of 2.4kg. Find the p - value for this test.

Solution:

Let μ be the mean weight lost during the first month of membership by all members and \bar{x} be corresponding mean for the sample.

Step 1. State the null and alternative hypothesis

$$H_0 : \mu \geq 10 \text{ (The mean weight lost is 10kg or more)}$$

$$H_1 : \mu < 10 \text{ (The mean weight lost is less than 10kg)}$$

Step 2. Select the distribution to use

Because the sample size is large we use the normal distribution to make the test and calculate p -value.

Step 3. Calculate the p -value.

The $<$ sign in the alternative hypothesis indicates that test is left tailed. The p -value is given by the area in the left tail of the sampling distribution curve of \bar{x} where \bar{x} is less than 9.2. To find this area, we first find the z value for $\bar{x} = 9.2$ as follows

$$T.S. = z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{9.2 - 10}{2.4 / \sqrt{36}} = -2.00$$

The area to the left of $\bar{x} = 9.2$ under the sampling distribution of \bar{x} is equal to the area under the standard normal curve to the left of $z = -2.00$. The area to the left of $z = -2.00$ is 0.0228. Consequently,

$$p\text{-value} = 0.0228$$

Thus, based on the p -value of 0.0228 we can state that for any α (significance level) greater than 0.0228 we will reject the null hypothesis and for any α less than 0.0228 we will accept the null hypothesis.

Suppose we make the test for this example at $\alpha = 0.01$. Because $\alpha = 0.01$ is less than p -value of 0.0228, we will not reject the null hypothesis. Now suppose we make the test at $\alpha = 0.05$. Because $\alpha = 0.05$ is greater than the p -value of 0.0228, we will reject the null hypothesis.

Exercises

1. Find the p -value for each of the following hypothesis tests

a) $H_0 : \mu = 18$; $H_1 : \mu \neq 18$; $n = 50$; $\bar{x} = 20$; $s = 5$;

b) $H_0 : \mu = 15$; $H_1 : \mu < 15$; $n = 80$; $\bar{x} = 13.2$; $s = 5.5$;

c) $H_0 : \mu = 38$; $H_1 : \mu > 38$; $n = 35$; $\bar{x} = 40.6$; $s = 7.2$

2. Consider $H_0 : \mu = 29$; against the alternative $H_1 : \mu \neq 29$.

A random sample of 60 observations taken from this population produced a sample mean of 31.4 and a standard deviation of 8.

- Calculate the p -value.
- Considering the p -value of part a), would you reject the null hypothesis if the test were made at the significance level of 0.05?
- Considering the p -value of part a), would you reject the null hypothesis if the test were made at the significance level of 0.01?

3. In a given situation, suppose H_0 was rejected at $\alpha = 0.05$. Answer the following questions as “yes”, “no”, or “can’t tell” as the case may be.

- Would H_0 also be rejected at $\alpha = 0.02$?
- Would H_0 also be rejected at $\alpha = 0.10$?
- Is the p -value smaller than 0.05?

4. In a problem of testing $H_0 : \mu = 75$ against $H_1 : \mu > 75$, the following sample quantities are recorded.

$$n = 56; \quad \bar{x} = 77.04; \quad s = 6.80$$

- State the test statistic and find the rejection region with $\alpha = 0.05$.
- Calculate the test statistic and draw a conclusion with $\alpha = 0.05$.
- Find the p -value and interpret the results.

Answers

1. a) 0.0046; b) 0.0017 ;c) 0.0162; 2. a) 0.0204; b) yes, reject H_0 ; c) no, do not reject H_0 ; 3. a) can't tell; b) yes; c) no; 4. a) $T.S. = Z = \frac{\bar{x} - 75}{s / \sqrt{n}}$; $Z \geq 1.645$; b) $T.S. = 2.24$, H_0 is rejected at $\alpha = 0.05$; c) 0.0125;

7.5. Tests of the mean of a normal distribution:

Population variance unknown. Small samples

Many times the size of a sample that is used to make test of hypothesis about μ is small, that is, $n < 30$. If the population is (approximately) normally distributed, the population standard deviation σ is not known and the sample size is small ($n < 30$), then the normal distribution is replaced by the Student's t distribution to make a test of hypothesis about μ . In such a case the random variable

$$t_{n-1} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

has a Student's t distribution with $(n - 1)$ degrees of freedom.

The value of test statistic t for the sample mean \bar{x} is computed as

$$T.S. = t_{n-1} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

and we can use the following tests with significance level α .

1. To test either null hypothesis

$H_0 : \mu = \mu_0$ or $H_0 : \mu \leq \mu_0$ against the alternative

$$H_1 : \mu > \mu_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha}$

2. To test either null hypothesis

$$H_0 : \mu = \mu_0 \quad \text{or} \quad H_0 : \mu \geq \mu_0$$

against the alternative

$$H_1 : \mu < \mu_0$$

the decision rule is

Reject H_0 if $T.S. < -t_{n-1,\alpha}$

3. To test the null hypothesis

$$H_0 : \mu = \mu_0 \quad \text{against the two sided alternative}$$

$$H_1 : \mu \neq \mu_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha/2}$ or $T.S. < -t_{n-1,\alpha/2}$,

Here, $t_{n-1,\alpha}$ is the number for which

$$P(t_{n-1} > t_{n-1,\alpha}) = \alpha$$

where the random variable t_{n-1} follows a Student's t distribution with $(n - 1)$ degrees of freedom.

Example:

The company that produces auto batteries claims that its batteries are good, for an average, for at least 64 days. A consumer protection agency tested 15 such batteries to check this claim. It found the mean life of these 15 batteries to be 62 days with a standard deviation of 3 days. At the 5% significance level, can you conclude that the claim of the company is true? Assume that the life of such a battery has an approximate normal distribution.

Solution:

Let μ be the mean life of all batteries and \bar{x} be the corresponding mean for the sample. Then from the given information,

$$n = 15; \quad \bar{x} = 62 \text{ days}; \quad s = 3 \text{ days}$$

The mean life of all batteries is supposed to be at least 64 days. The significance level is α is 0.05. That is, the probability of rejecting the null hypothesis when it is actually true should not exceed 0.05.

Step 1. State the null and alternative hypothesis

We write the null and alternative hypothesis as

$$H_0 : \mu \geq 64 \text{ days} \text{ (The mean life is at least 64 days)}$$

$$H_1 : \mu < 64 \text{ days} \text{ (The mean life is less than 64 days)}$$

Step 2. Select the distribution to use

The sample size is small ($n=15$), and the life of a battery is approximately normally distributed. Since population standard deviation is unknown, we use the Student's t distribution to make the test.

Step 3. Determine the rejection and nonrejection regions

The significance level is 0.05. The $<$ sign in the alternative test indicates that the test is left tailed with the rejection region in the left tail of the t distribution curve.

$$\text{Area in the left tail} = \alpha = 0.05$$

$$\text{Degree of freedom} = n - 1 = 15 - 1 = 14$$

From the Student's t distribution table (Table 2 of Appendix), the critical value of t for 14 degrees of freedom and an area 0.05 in the left tail is -1.761 . (Fig.1.9).

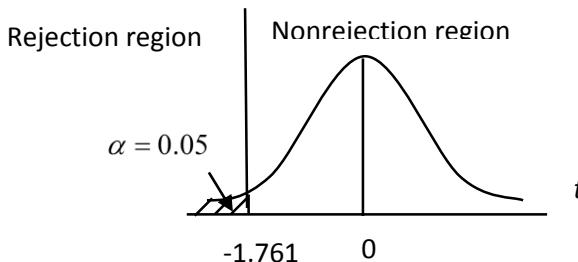


Fig.1.9

Step 4. Calculate the value of the test statistic

As σ is not known, and sample size is small, we calculate the t value as follows

$$T.S. = t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{62 - 64}{3 / \sqrt{15}} = -2.50$$

Step 5. Make a decision

The value of $T.S. = t = -2.50$ is less than the critical value of $t = -1.761$, and it falls in the rejection region. Therefore, we reject H_0 and conclude that the sample mean is too small compared to 62 days (company's claimed value of μ) and the difference between the two may not be attributed to chance alone. We can conclude that the mean life of company's batteries is less than 62 days.

Remark: The conclusion of a t -test can also be strengthened by reporting the significance probability (p - value) of the observed statistic. Since the t table provides only a few selected percentage points, we can get an idea about the p -value but not its exact determination. For instance, the data in example above gave an observed value $T.S. = t = -2.50$ with degree of freedom=14. Scanning the t table for $(n - 1) = 14$, we notice that that 2.50 lies between $t_{0.025}$ and $t_{0.010}$. Therefore, the p -value of $t = -2.50$ is higher than 0.025 but not as great as 0.010.

Exercises

1. For each of the following examples of tests of hypothesis about μ , show the rejection and nonrejection regions on the t distribution curve.

- A two tailed test with $\alpha = 0.2$ and $n = 14$
- A left tailed test with $\alpha = 0.005$ and $n = 23$
- A right tailed test with $\alpha = 0.025$ and $n = 14$

2. Consider the null hypothesis $H_0 : \mu = 45$ about the mean of a population that is normally distributed. Suppose a random sample of 20 observations is

taken from this population to make this test. Using $\alpha = 0.05$ show the rejection and nonrejection regions and find critical value(s) for t for

- a) left tailed test; b) two tailed test; c) right tailed test

3. Consider $H_0 : \mu = 40$ versus $H_1 : \mu > 40$ for a population that is normally distributed.

a) A random sample of 16 observations taken from this population produced a sample mean of 45 and a standard deviation of 5. Using $\alpha = 0.025$, would you reject the null hypothesis?

b) Another random sample of 16 observations taken from the same population produced a sample mean of 41.9 and a standard deviation of 7. Using $\alpha = 0.025$, would you reject the null hypothesis?

Comment on the result of parts a) and b).

4. Assuming that respective populations are normally distributed, make the following hypothesis tests.

a) $H_0 : \mu = 60$; $H_1 : \mu \neq 60$; $n = 14$; $\bar{x} = 56$; $s = 9$; $\alpha = 0.05$

b) $H_0 : \mu = 35$; $H_1 : \mu < 35$; $n = 24$; $\bar{x} = 29$; $s = 5.4$; $\alpha = 0.005$

c) $H_0 : \mu = 47$; $H_1 : \mu > 47$; $n = 18$; $\bar{x} = 51$; $s = 6$; $\alpha = 0.001$

5. A business school claims that students who complete a three month course of typing course can type on average, at least 1200 words an hour.

A random sample of 25 students who completed this course typed, on average, 1130 words an hour with a standard deviation of 85 words. Assume that the typing speeds for all students who complete this course have an approximate normal distribution.

Using the 5% significance level, can you conclude that the claim of the business school is true?

- 6.** The supplier of home heating furnaces of a new model claims that the average efficiency of the new model is at least 60. Before buying these heating furnaces, a distributor wants to verify the supplier's claim is valid. To this end, the distributor chooses a random sample of 9 heating furnaces of a new model and measures their efficiency. The data are

63; 72; 64; 69; 59; 65; 66; 64; 65

Determine the rejection region of the test with $\alpha = 0.05$. Apply the test and state your conclusion.

- 7.** A past study claims that adults spend an average of 18 hours a week on leisure activities. A researcher wanted to test this claim. He took a sample of 10 adults and asked them about the time they spend per week on leisure activities. Their responses (in hours) were as follows

14; 25; 22; 38; 16; 26; 19; 23; 41; 33

Assume that the time spent on leisure activities by all adults is normally distributed. Using the 5% significance level, can you conclude that the claim of earlier study is true?

- 8.** According to the department of Labor, private sector workers earned, on average \$354.32 a week in 2001. A recently taken random sample of 400 private sector worker showed that they earn, on average, \$362.50 a week with a standard deviation of \$72. Find p -value for the test with an alternative hypothesis that the current mean weekly salary of private sector workers is different from \$354.32.

- 9.** A manufacturer of a light bulbs claims that the mean life of these bulbs is at least 2500 hours. A consumer agency wanted to check whether or not this claim is true. The agency took a random sample of 36 such bulbs and tested them. The mean life for the sample was found to be 2447 hours with a standard deviation of 180 hours.

a) Do you think that the sample information supports the company's claim?

Use $\alpha = 2.5\%$.

b) What is the Type I error in this case? Explain. What is the probability of making this error?

c) Will your conclusion of part a) change if the probability of making a Type I error is zero?

10. Given the eight sample observations 31, 29, 26, 33, 40, 28, 30, and 25, test the null hypothesis that the mean equals 35 versus the alternative that it does not. Let $\alpha = 0.01$.

Answers

- 1.a) reject H_0 if $t < -1.350$ or $t > 1.350$; b) reject H_0 if $t < -2.819$; c) reject H_0 if $t > 2.160$; 2. a) reject H_0 if $t < -1.729$; b) reject H_0 if either $t > 2.093$ or $t < -2.093$; c) reject H_0 if $t > 1.729$; 3. a) $T.S. = t = 4.00$; reject H_0 ; b) $T.S. = t = 1.086$; accept H_0 ; 4. a) $T.S. = -1.663$; accept H_0 ; b) $T.S. = -5.443$; reject H_0 ; c) $T.S. = 2.828$; accept H_0 ; 5. $T.S. = -4.118$; reject H_0 ; 6. $t \geq 1.860$; $T.S. = 4.26$; H_0 is rejected at $\alpha = 0.05$; 7. $T.S. = 2.692$; reject H_0 ; 8. $T.S. = 2.27$; p -value = 0.0232; 9. a) $T.S. = z = -1.77$; accept H_0 ; b) 0.025; c) no; 10. $T.S. = -2.85$; H_0 is not rejected.

1.6. Tests of the population proportion (Large sample)

Often we want to conduct test of hypothesis about a population proportion.

This section presents the procedure to perform tests of hypothesis about the population proportion, p for large samples ($n \geq 40$). The procedure to make such tests is similar in many respects to the one for the population mean μ .

The value of the test statistic $T.S. = z$ for the sample proportion \hat{p} computed as

$$T.S. = z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

where \hat{p} is the sample proportion, and the value of p_0 used in this formula is the one used in the null hypothesis.

Then, if the number of sample observations is large and observed proportion is \hat{p} , the following tests have significance level α :

1. To test either null hypothesis

$$H_0 : p = p_0 \quad \text{or} \quad H_0 : p \leq p_0 \quad \text{against the alternative}$$

$$H_1 : p > p_0$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > z_\alpha$$

2. To test either null hypothesis

$H_0 : p = p_0$ or $H_0 : p \geq p_0$ against the alternative

$H_1 : p < p_0$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$H_0 : p = p_0$ against the two sided alternative

$H_1 : p \neq p_0$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

Once again, z_α is the number for which

$$P(Z > z_\alpha) = \alpha$$

and Z is the standard normal distribution.

Example:

Mr. A and Mr. B are running for local public office in a large city. Mr. A says that only 30% of the voters are in favor of a certain issue, a law to sell liquor on Sundays. Mr. B doubts A's statement and believes that more than 30% favor such legislation. Mr. B pays for an independent organization to make a study of this situation. In a random sample 400 voters, 160 favored the legislation. What conclusions should the polling organization report to Mr. B?

Solution:

Let p_0 be proportion of all people who favor such legislation and \hat{p} the corresponding sample proportion. Then from given information,

$$n = 400; \quad p_0 = 0.30; \quad \hat{p} = \frac{160}{400} = 0.40. \text{ Let } \alpha = 0.05.$$

The null and alternative hypotheses are as follows

$$H_0 : p = p_0 = 0.30$$

$$H_1 : p > 0.30$$

The decision rule is to reject the null hypothesis in favor of alternative if

$$T.S. > z_\alpha$$

$$\alpha = 0.05; \quad \alpha/2 = 0.025.$$

$$P(Z > z_{\alpha/2}) = P(Z > z_{0.025}) = 0.025.$$

$$P(Z > z_{0.025}) = F(z_{0.025}) = 0.975 \text{ and}$$

$$z_{\alpha/2} = z_{0.025} = 1.645$$

From the given information we calculate the value of test statistic as

$$T.S. = z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.40 - 0.30}{\sqrt{0.30 \cdot 0.70 / 400}} = 4.36$$

Since $4.36 > 1.645$ we reject H_0 . We make conclusion that more than 30% of voters are in favor of a law to sell liquor on Sundays.

Exercises

1. Make the following hypothesis tests about p .

a) $H_0 : p = 0.45$; $H_1 : p \neq 0.45$; $n = 100$; $\hat{p} = 0.48$; $\alpha = 0.10$

b) $H_0 : p = 0.72$; $H_1 : p < 0.72$; $n = 700$; $\hat{p} = 0.65$; $\alpha = 0.05$

c) $H_0 : p = 0.30$; $H_1 : p > 0.30$; $n = 200$; $\hat{p} = 0.34$; $\alpha = 0.01$

2. Consider $H_0 : p = 0.70$ versus $H_1 : p \neq 0.70$.

a) A random sample of 600 observations produced a sample proportion equal to 0.67. Using $\alpha = 0.01$, would you reject the null hypothesis?

b) Another random sample of 600 observations taken from the same population produced a sample proportion of 0.76. Using $\alpha = 0.01$, would you reject the null hypothesis?

Comment on the result of parts a) and b).

3. A food company is planning to market a new type of ice cream. Before marketing this ice cream, the company wants to find what percentage of the people like it. The company's management has decided that it will market this ice cream only if at least 35% of people like it. The company's research department selected a random sample of 400 persons and asked them to test this ice cream. Of these 400 persons, 128 said they liked it.

a) Testing at 2.5% significance level, can you conclude that the company should market this yogurt?

b) What will your decision be in part a) if the probability of making a Type I error is zero?

4. A mail order company claims that at least 60% of all orders are mailed within 48 hours. The quality control department took a sample of 500 orders and found that 310 of them were mailed within 48 hours of the placement of the orders. Testing at 1% significance level, can you conclude that the company's claim is true?

5. Let p =proportion of adults in a city who required a lawyer in the past year.

a) Determine the rejection region for $\alpha = 0.05$ level test of $H_0 : p = 0.25$

against $H_1 : p > 0.25$.

b) If 65 persons in a random sample of 200 required lawyer services, what does the test conclude?

6. A magazine claims that 25% of its readers are university students. A random sample of 200 readers is taken and 42 of these readers are university students. Use $\alpha = 0.10$ level of significance to test the validity of the magazine's claim.

7. Suppose that in order to test the hypothesis that $p = 0.6$ against the alternative that $p < 0.6$, we decide to obtain a sample of size 100 and reject H_0 if we obtain fewer than 48 successes.

a) What is the approximate size of the Type I error?

b) If the value of p is really 0.5, what is the size of Type II error?

8. An educator wishes to test $H_0 : p = 0.3$ against $H_1 : p > 0.3$, where

p -proportion of football players who graduate university in four years.

a) State the test statistic and the rejection region having $\alpha = 0.05$.

b) If 19 out of a random sample of 48 players graduated in four years, what does the test conclude? Also evaluate p -value.

9. The president of a company that produces national brand coffee claims that 40% of the people prefer to buy national brand coffee. A random sample of 700 people who buy coffee showed that 252 of them buy national brand coffee.

- Using $\alpha = 0.01$, can you conclude that the percentage of people who buy national brand coffee is different from 40%?
- Find the p -value for the test. Using this p -value, would you reject the null hypothesis at $\alpha = 0.05$? What if $\alpha = 0.02$?

Answers

- 1.** a) $T.S. = z = 0.60$; do not reject H_0 ; b) $T.S. = -4.12$; reject H_0 ;
- c) $T.S. = 1.23$; do not reject H_0 ;
- 2.** a) $T.S. = -1.60$; do not reject H_0 ;
- b) $T.S. = 3.21$; reject H_0 ;
- 3.** a) $T.S. = -1.26$; do not reject H_0 ; b) do not reject H_0 ;
- 4.** $T.S. = 0.91$; accept H_0 ;
- 5.** a) $z \geq 1.645$; b) $T.S. = 2.45$; reject H_0 ;
- 6.** $T.S. = -1.31$; accept H_0 ;
- 7.** a) about 0.0071 b) approximately 0.6554;
- 8.** a) $z \geq 1.645$; b) $T.S. = 1.45$; accept H_0 for $\alpha = 0.05$; p -value = 0.0735;
- 9.** a) $T.S. = -2.16$; do not reject H_0 ; b) p -value = 0.0308; reject H_0 at $\alpha = 0.05$; do not reject H_0 at $\alpha = 0.02$

7.7. Tests for the difference between two population means

7.7.1. Tests based on paired samples

Suppose that a random sample of n matched pairs of observations is obtained from populations with means μ_x and μ_y . The observations will be denoted by $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Let

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

$$\text{and } s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n d_i^2 - n(\bar{d})^2}{n-1}}$$

denote the observed sample mean and standard deviation for the n differences $d_i = x_i - y_i$. Let us denote difference between two population means by $D_0 = \mu_x - \mu_y$. In this case test statistic will be calculated as

$$T.S. = \frac{\bar{d} - D_0}{s_{\bar{d}} / \sqrt{n}}$$

If the population differences is a normal distribution, then the following tests have significance level α

1. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{or} \quad H_0 : \mu_x - \mu_y \leq D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y > D_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha}$

2. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{or} \quad H_0 : \mu_x - \mu_y \geq D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y < D_0$$

the decision rule is

Reject H_0 if $T.S. < -t_{n-1,\alpha}$

3. To test the null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{against the two sided alternative}$$

$$H_1 : \mu_x - \mu_y \neq D_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-1,\alpha/2}$ or $T.S. < -t_{n-1,\alpha/2}$

Here, $t_{n-1,\alpha}$ is the number for which

$$P(t_{n-1} > t_{n-1,\alpha}) = \alpha$$

where the random variable t_{n-1} follows a Student's t distribution with $(n-1)$ degrees of freedom.

Remark: When we want to test the null hypothesis that the two population means are equal, we set $D_0 = 0$.

Example:

A medical researcher wishes to determine if a pill has the undesirable side effect of reducing the blood pressure of the user. The study involves recording the initial blood pressures of 7 college age adults. After they use the pill regularly for three month, their blood pressures are again recorded. The researcher wishes to draw inferences about the effect of the pill on blood pressure from the information given in table

Before x_i	64	71	68	66	73	62	70
After y_i	60	66	66	69	63	57	62

Do the data substantiate the claim that use of the pill reduces the blood pressure? Use $\alpha = 0.01$. Assume that the population of paired differences has a normal distribution.

Solution:

Let d be the difference between the pressures before and after using pills.

$$d = \text{before} - \text{after} = x_i - y_i$$

The necessary calculations are shown in the following table

Before	After	Difference d	d^2
64	60	4	16
71	66	5	25
68	66	2	4
66	69	-3	9

73	63	10	100
62	57	5	25
70	62	8	64
		$\sum d = 31$	$\sum d^2 = 243$

The values of \bar{d} and S_d are calculated as follows:

$$\bar{d} = \frac{\sum d}{n} = \frac{31}{7} = 4.43$$

$$S_d = \sqrt{\frac{1}{n-1} \left[\sum d^2 - n \cdot (\bar{d})^2 \right]} = \sqrt{\frac{1}{6} (243 - 7 \cdot 4.43^2)} = 4.198.$$

Let μ_x be the mean blood pressure for all adults before and μ_y -after using the pill.

The null and alternative hypotheses are

$$H_0 : \mu_x - \mu_y = 0 \text{ (no difference)}$$

against

$$H_1 : \mu_x - \mu_y > 0 \text{ (mean decreases)}$$

The decision rule is that

Reject H_0 if $T.S. > t_{n-1,\alpha}$

$$T.S. = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}} = \frac{4.43 - 0}{4.198/\sqrt{7}} = 2.792$$

$$t_{n-1,\alpha} = t_{6,0.01} = 3.14$$

Since $2.792 < 3.14$, we accept H_0 and make conclusion at the level 0.01 that using pills does not affect blood pressure.

Exercises

1. Perform the following tests of hypothesis assuming that the population of paired differences are normally distributed.

a) $H_0 : \mu_d = D_0 = 0 ; H_1 : \mu_d = D_0 \neq 0 ; n = 9 ; \bar{d} = 6.7 ; s_d = 2.5 ; \alpha = 0.10$

b) $H_0 : \mu_d = D_0 = 0 ; H_1 : \mu_d = D_0 > 0 ; n = 22 ; \bar{d} = 14.3 ; s_d = 6.4 ; \alpha = 0.05$

c) $H_0 : \mu_d = D_0 = 0 ; H_1 : \mu_d = D_0 < 0 ; n = 17 ; \bar{d} = -9.3 ; s_d = 4.8 ; \alpha = 0.01$

2. It is claimed that an industrial safety program is effective in reducing the loss of working hours due to factory accidents. The following data are collected concerning the weekly hours due to accidents in nine plants both before and after the safety program is installed

Before x_i	90	86	72	65	44	52	46	38	43
After y_i	85	87	70	62	44	53	42	35	46

Do the data substantiate the claim? Use $\alpha = 0.05$.

Assume that the population of paired differences is (approximately) normally distributed.

- 3.** A company claims that the course it offers significantly increases the writing speed of secretaries. The following table gives the scores of 8 secretaries before and after they attended this course

Before x_i	81	75	89	91	65	70	90	69
After y_i	97	72	93	110	78	69	115	75

Using the 5% significance level, can you conclude that attending this course increases the writing speed of secretaries?

Assume that the population of paired differences is (approximately) normally distributed.

- 4.** A random sample of nine employees was selected to test for the effectiveness of hypnosis on their job performance. The following table gives the job performance ratings (on a scale of 1 to 4, with 1 being the lowest and 4 being the highest) before and after these employees tried hypnosis.

Before x_i	2.3	2.8	3.1	2.7	3.4	2.6	2.8	2.5
After y_i	2.6	3.2	3.0	3.5	3.7	2.4	2.9	2.9

Test at the 5% significance level if there is an improvement in the job performances of employees due to hypnosis.

Assume that the population of paired differences is (approximately) normally distributed.

Answers

- 1.** a) $T.S. = 8.040$; reject H_0 ; b) $T.S. = 10.847$; reject H_0 ; c) $T.S. = -7.989$; reject H_0 ;
2. $T.S. = 1.48$; do not reject H_0 ;
3. $T.S. = -2.807$; reject H_0 ;
4. $T.S. = -2.236$; accept H_0 .

7.7.2. Tests based on independent samples

(Known variance or large sample size)

Let us consider the case where we have independent random samples from two normally distributed populations. The first population has mean μ_x and variance σ_x^2 and we obtain a random sample of size n_x . The second population has mean μ_y and variance σ_y^2 and we obtain a random sample of size n_y .

We know that if the sample means are denoted \bar{x} and \bar{y} , then the random variable

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

has a standard normal distribution. If the population variances are known, tests for the difference between the population means can be based on this result. The value of the test statistic z for $(\bar{x} - \bar{y})$ is computed as

$$T.S. = z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

and the following tests have a significance level α

1. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{or} \quad H_0 : \mu_x - \mu_y \leq D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y > D_0$$

the decision rule is

Reject H_0 if $T.S. > z_\alpha$

2. To test either null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{or} \quad H_0 : \mu_x - \mu_y \geq D_0$$

against the alternative

$$H_1 : \mu_x - \mu_y < D_0$$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{against the two sided alternative}$$

$$H_1 : \mu_x - \mu_y \neq D_0$$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

Remark: If the sample sizes are large ($n_x > 30; n_y > 30$) then a good approximation at significance level α can be made if the population variances σ_x^2 and σ_y^2 are replaced by the sample variances s_x^2 and s_y^2 .

In addition the central limit theorem leads to good approximations even if the populations are not normally distributed.

Example:

According to the Bureau of Labor Statistics, last year university instructors earned an average \$440 per month and college instructors earned an average of \$420 per month. Assume that these mean earnings have been calculated for samples of 400 and 600 instructors taken from the two populations, respectively. Further assume that the standard deviations of monthly earnings of the two populations are \$50 and \$63, respectively. Test at 1% significance level if the mean monthly earnings of the two groups of the instructors are different.

Solution:

From the information given above,

$$n_x = 400; \quad \bar{x} = 440; \quad \sigma_x = 50;$$

$$n_y = 600; \quad \bar{y} = 420; \quad \sigma_y = 63;$$

where the subscript x refers to university instructors and y -to college instructors. Let

μ_x = mean monthly earnings of all university instructors

μ_y = mean monthly earnings of all college instructors.

We are to test if the two population means are different. The null and alternative hypotheses are

$$H_0 : \mu_x - \mu_y = 0 \text{ (the monthly earnings are not different)}$$

$H_1: \mu_x - \mu_y \neq 0$ (the monthly earnings are different).

The decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > z_{\alpha/2} \quad \text{or} \quad T.S. < -z_{\alpha/2}$$

First of all we find the value of $z_{\alpha/2}$. Since $\alpha/2 = 0.005$, the value of $z_{\alpha/2}$ is (approximately) 2.58 and $-z_{\alpha/2} = -2.58$.

The value of the test statistic $T.S. = z$ is computed as follows:

$$T.S. = z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = \frac{(440 - 420) - (0)}{\sqrt{\frac{50^2}{400} + \frac{63^2}{600}}} = 5.57.$$

5.57 > 2.58 and the value of test statistic $T.S. = z = 5.57$ falls in the rejection region, we reject the null hypothesis H_0 . Therefore, we conclude that the mean monthly earnings of the two groups of instructors are different.

Note that we can not say for sure that two means are different. All we can say is that the evidence from the two samples is very strong that the corresponding population means are different.

Exercises

- 1.** The following information is obtained from two independent samples selected from two populations

$$n_1 = 155; \bar{x} = 5.58; s_x = 1.62$$

$$n_2 = 190; \bar{y} = 4.80; s_y = 1.52$$

Test at the 1% significance level if the two population means are the same against the alternative that they are different.

2. Daily wage is \$13.62 for transportation workers and \$11.61 for factory workers. Assume that these two estimates are based on random samples of 1000 and 1200 workers taken, respectively, from the two populations. Also assume that the standard deviations of the two populations are \$1.85 and \$1.40, respectively.

- a) Test at the 5% significance level if the mean daily wage of transportation workers and factory workers are the same against the alternative that it is higher for transportation workers.
- b) What will your decision be in part a) if the probability of making a Type I error is zero. Explain.

3. A consulting firm was asked by a large insurance company to investigate if business majors were better salespersons. A sample of 40 salespersons with a business degree showed that they sold an average of 10 insurance policies per week with a standard deviation of 1.80. Another sample of 45 salespersons with a degree other than business showed that they sold an average of 8.5 insurance policies per week with a standard deviation of 1.35. Using the 1% significance level, can you conclude that person with a business degree are better salespersons than those who have a degree in another area?

4. The management at the bank A claims that the mean waiting time for all customers at its branches is less than that at the bank B, which is main competitor. They took a sample of 200 customers from the bank A and found that they waited an average of 4.60 minutes with a standard deviation of 1.2 minutes before being served. Another sample of 300 customers taken from the bank B showed that these customers waited an average of 4.85 minutes with a standard deviation of 1.5 minutes before being served.

- a) Test at the 2.5% significance level if the claim of the management of the bank A is true.
- b) Calculate the p -value. Based on this p -value, would you reject the null hypothesis if $\alpha = 0.01$? What if $\alpha = 0.05$?

5. A production line is designed on the assumption that the difference in mean assembly times for two operations is 5 minutes. Independent tests for the two assembly operations show the following results:

Operation A

$$n_1 = 100$$

$$\bar{x} = 14.8 \text{ minutes}$$

$$s_x = 0.8 \text{ minutes}$$

Operation B

$$n_2 = 50$$

$$\bar{y} = 10.4 \text{ minutes}$$

$$s_y = 0.6 \text{ minutes}$$

For $\alpha = 0.02$, test the hypothesis that the difference between the mean assembly times is 5 minutes.

6. An investigation was carried out to determine if women employees are as well paid as their male counterparts. Random samples of 75 males and 64 females are selected. Their mean salaries were 45 530 and 44 620, standard deviations were 780 and 750, correspondingly. If you were to test the null hypothesis that the mean salaries are equal against the two sided alternative, what would be the conclusion of your test with $\alpha = 0.05$?

7. For a random sample of 125 state companies, the mean number of job changes was 1.91 and the standard deviation was 1.32. For a random sample of 86 private companies, the mean number of job changes was 0.21 and the standard deviation was 0.53. Test the null hypothesis that the population means are equal against the alternative that the mean number of job changes is higher in state companies than for private companies.

Answers

1. T.S. = $z = 4.56$; reject H_0 ; 2. a) T.S. = 28.27 ; reject H_0 ; b) do not reject H_0 ;
3. T.S. = 4.30 ; reject H_0 ; 4. a) T.S. = -2.06 ; reject H_0 ; b) p -value = 0.0197 ;
do not reject H_0 at $\alpha = 0.01$?; reject H_0 at $\alpha = 0.05$; 5. T.S. = -5.15 ; reject H_0 ; 6. T.S. = 7 ; reject H_0 ; 7. T.S. = 13 ; reject H_0 at any level.

7.8. Tests for the difference between two population proportions (Large samples)

Next we will develop procedures for comparing two population proportions. We will consider standard model with a random sample of n_x observations with proportion \hat{p}_x "successes" and an independent random sample of n_y observations from population with proportion \hat{p}_y "successes".

We know that for large samples, proportions can be approximated as normally distributed random variables and as a result

$$Z = \frac{\left(\hat{p}_x - \hat{p}_y \right) - (p_x - p_y)}{\sqrt{\frac{p_x(1-p_y)}{n_x} + \frac{p_y(1-p_y)}{n_y}}}$$

has a standard normal distribution.

We want to test the hypothesis that the population proportions p_x and p_y are equal. Denote their common value by p_0 , then the value under this hypothesis

$$Z = \frac{\hat{(p_x - p_y)}}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}}$$

follows a good approximation a standard normal distribution.

Finally, unknown common proportion p_0 can be estimated by a pooled estimator defined as

$$p_0 = \frac{n_x \cdot \hat{p}_x + n_y \cdot \hat{p}_y}{n_x + n_y} \text{ or } p_0 = \frac{x_1 + y_1}{n_x + n_y}$$

where x_1 and x_2 are number of “successes” in n_x and n_y , respectively.

Which of these formulas is used to calculate p_0 depends on whether the values of x_1 and y_1 or the values of \hat{p}_x and \hat{p}_y are known.

Testing equality of two population proportions:

We are given independent samples of size n_x and n_y with proportion of successes \hat{p}_x and \hat{p}_y . When we assume that the population proportions are equal, an estimate of the common proportion is

$$p_0 = \frac{\hat{n}_x \cdot \hat{p}_x + \hat{n}_y \cdot \hat{p}_y}{\hat{n}_x + \hat{n}_y} \text{ or } p_0 = \frac{x_1 + y_1}{n_x + n_y}.$$

For large sample sizes $\left(n \cdot \hat{p} \cdot \hat{q} > 9 \right)$ the value of the test statistic z for is computed as

$$T.S. = z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}}$$

Then the following tests have significance level α :

1. To test either null hypothesis

$$H_0 : p_x - p_y = 0 \quad \text{or} \quad H_0 : p_x - p_y \leq 0$$

against the alternative

$$H_1 : p_x - p_y > 0$$

the decision rule is

Reject H_0 if $T.S. > z_\alpha$

2. To test either null hypothesis

$$H_0 : p_x - p_y = 0 \quad \text{or} \quad H_0 : p_x - p_y \geq 0$$

against the alternative

$$H_1 : p_x - p_y < 0$$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$H_0 : p_x - p_y = 0$ against the two sided alternative

$H_1 : p_x - p_y \neq 0$

the decision rule is

Reject H_0 if $T.S. > z_{\alpha/2}$ or $T.S. < -z_{\alpha/2}$

Example:

A company is planning to buy a few machines. Company is considering two types of machines, but will buy all of the same type. The company selects one machine from each type and uses for a few days. A sample of 900 items produced on machine A showed that 55 of them were defective. A sample of 700 items produced on machine B showed that 41 of them were defective. Testing at 1% significance level, can we conclude based on the information from these samples that the proportions of the defective items produced on the two machines are different?

Solution:

Let p_x be the proportion of all items produced on machine A, and p_y be the proportion of all items produced on machine B.

Let \hat{p}_x and \hat{p}_y be the corresponding sample proportions. Let x_1 and x_2 be the number of defective items in two samples respectively.

$$\text{Machine A: } n_x = 900; \quad x_1 = 55$$

$$\text{Machine B: } n_y = 700; \quad y_1 = 41$$

The two sample proportions are calculated as follows:

$$\hat{p}_x = \frac{x_1}{n_x} = \frac{55}{900} = 0.0611;$$

$$\hat{p}_y = \frac{y_1}{n_y} = \frac{41}{700} = 0.0586$$

The null and alternative hypotheses are

$$H_0: p_x - p_y = 0 \quad (\text{the two proportions are equal})$$

$$H_1: p_x - p_y \neq 0 \quad (\text{the two proportions are different})$$

The decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > z_{\alpha/2} \quad \text{or} \quad T.S. < -z_{\alpha/2}$$

Let us check if the sample sizes are large:

$$\left(n_x \cdot \hat{p}_x \cdot \hat{q}_x > 9 \right) = 900 \cdot 0.0611 \cdot 0.9389 = 51.63 > 9$$

$$\left(n_y \cdot \hat{p}_y \cdot \hat{q}_y > 9 \right) = 700 \cdot 0.0586 \cdot 0.9414 = 38.62 > 9$$

Since the samples are large and independent we apply the normal distribution to make a test.

The pooled sample proportion is

$$p_0 = \frac{x_1 + y_1}{n_x + n_y} = \frac{55 + 41}{900 + 700} = 0.06$$

The value of the test statistics is

$$T.S. = z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}} = \frac{(0.0611 - 0.0586)}{\sqrt{\frac{0.06 \cdot 0.94}{900} + \frac{0.06 \cdot 0.94}{700}}} = 0.2089.$$

Let us find the value of $z_{\alpha/2}$.

$$\alpha = 0.01; \alpha/2 = 0.005$$

$$F_z(z_{\alpha/2}) = F_z(z_{0.005}) = 0.995$$

$$z_{0.005} = 2.58 \text{ and } -z_{0.005} = -2.58$$

The value of the test statistic $T.S. = z = 0.2089$ falls in the nonrejection region. Consequently, we fail to reject the null hypothesis. As a result, we can conclude that proportions of defective items produced by two machines are not different.

Exercises

1. The following information is obtained from two independent samples selected from two populations

$$n_x = 750; \quad \hat{p}_x = 0.56; \quad n_y = 600; \quad \hat{p}_y = 0.61;$$

Test at the 1% significance level if p_x is equal p_y , against the alternative that it is less.

2. A sample of 600 observations taken from the first population gave $x_1 = 320$. Another sample of 700 observations taken from the second population gave $y_1 = 370$. Show the rejection and nonrejection regions on the sampling distribution of $\hat{p}_x - \hat{p}_y$ for $H_0 : p_x = p_y$ against the alternative $H_1 : p_x > p_y$ using significance level of 2.5%. Will you reject the null hypothesis?

3. According to the statistics, 65% of single women and 80% of single men own cars. Assume that these estimates are based on random samples of 1700 singles women and 1900 single men. At the 1% significance level, can you conclude that the proportion of single women who own cars is the same as the proportion of men who own cars against the alternative that it is less than the proportion of men who own cars? Also find p -value.

4. The management of a supermarket wanted to investigate if the percentage of men and women who prefer to buy national brand products over the store brand products are different. A sample of 600 men shoppers at the company's supermarkets showed that 246 of them prefer to buy national brand products over the store brand products. Another sample of 700 women shoppers showed that 266 of them prefer to buy national brand products over the store brand products. Testing at the 5% significance level, can you conclude that the proportion of all men and all women shoppers at these supermarkets who prefer to buy national brand products over the store brand products are equal?

5. A sample of 500 male registered voters showed that 57% of them are in a favor of higher taxes on wealthy people. Another sample of 400 female registered voters showed that 54% of them are in favor of higher taxes on wealthy people. Test at the 1% significance level if the proportion of all male voters who are in favor of higher taxes on wealthy people is not different from that of female voters against two-sided alternative.

6. A medical researcher investigates if the smoking results in wrinkled skin around the eyes. By observing 150 smokers and 250 nonsmokers, the researcher finds that 95 of the smokers and 103 of the nonsmokers have prominent wrinkles around their eyes. Do these data substantiate the belief that prominent wrinkles around eyes are more prevalent among smokers than nonsmokers? Answer by calculating p -value.

7. In a comparative study of two new drugs, A and B, 120 patients treated with drug A and 150 patients with drug B, and the following results were obtained

	Drug A	Drug B
Cured:	52	88
Not cured:	68	62
Total:	120	150

Do these results demonstrate statement that these two drugs have the same effect against the alternative that higher cure rate with drug A? Test at $\alpha = 0.05$.

8. According to a 2001 survey, 48% of managers “would choose the same career if they were starting over again”. In a similar survey conducted 10 years ago, 60% of managers said that they “would choose the same career if they were starting over again”. Assume that the 2001 survey is based on a sample of 800 managers and the one done 10 years ago included 600 managers. Test at the 5% significance level if the proportion of all managers who “would

choose the same career if they were starting over again" has not changed against the alternative that it decreased during the past 10 years.

Answers

- 1.** $T.S. = z = -1.85$; accept H_0 ;
2. $T.S. = 0.17$; accept H_0 ;
3. $T.S. = -9.80$; reject H_0 ; can reject H_0 at virtually any level;
4. $T.S. = z = 1.10$; do not reject H_0 ;
5. $T.S. = 0.90$; accept H_0 ;
6. $T.S. = 4.28$; $p\text{-value} = 0.0002$; claim is strongly supported;
7. $H_0 : p_A = p_B$; $H_1 : p_A < p_B$; $T.S. = -2.52$; reject H_0 ;
8. $T.S. = -4.45$; reject H_0 .

Chapter 9

Some nonparametric tests

9.1. Introduction

When the methods of statistical inference are based upon the assumption that the population has a certain probability distribution, such as the normal, the resulting collection of statistical tests and procedures is referred to as *parametric methods*. In this chapter we will consider several statistical procedures that do not require knowledge of the form of the probability distribution from which the measurements come. The methods of statistical inference we will study here are called *nonparametric methods*. Since nonparametric methods do not require assumptions about the form of the population distribution they are often referred to as *distribution free methods*.

From this discussion we see that one reason for using nonparametric methods is that in some situations there is insufficient knowledge about the form of the population distribution. Thus assumptions necessary for use of parametric tests can not be made.

A second reason for using nonparametric methods concerns data measurement. Nonparametric methods are often applied to rank order or preference data. Preference data are the type of data generated when people express preference for one product over another, one service over another, etc. Parametric procedures can not be applied with these data, but nonparametric ones can.

This chapter presents an introduction to some of the commonly used statistical procedures that can be classified as nonparametric or distribution free methods. The emphasis will be on the type of problems that can be solved, how the statistical calculations are made, and how appropriate conclusions can be developed to assist management in the decision-making process.

9.2. 1. The Sign test for paired or matched samples

In Chapter 1 we considered z and t statistics for testing hypothesis about a population mean. For both of them, the sample was selected at random from a *normal* distribution. The question is: How can we conduct a test of hypothesis when we have a small sample from a nonnormal distribution?

The **Sign test** is a relatively simple and most frequently employed nonparametric procedure for testing hypothesis about the central tendency of a nonnormal probability distribution. The sign test is used in studies to identify if consumer preference exists for one of two products.

Suppose that paired or matched samples are taken from a population, and the differences equal to 0 are discharged, leaving n observations. The Sign test can be used to test the null hypothesis that the population median of the differences is 0. Let “+” indicates a positive difference, and “-“ indicates a negative difference. If the null hypothesis were true, our sequence of “+” and “-“ differences could be regarded as a random sample from a population in which the probabilities for “+” and “-“ were each 0.5. In that case, the observations would constitute a random sample from a binomial population in which the probability of “+” was 0.5. Thus, if p denotes the true population proportion of “+”’s in the population (that is, the true proportion of positive differences), the null hypothesis is simply

The Sign test is then based on the fact that the number of positive observations, S , in the sample has a binomial distribution (with $p = 0.5$ under the null hypothesis).

$$H_0 : p = 0.5$$

Sign test for paired samples

Suppose that paired random samples are taken from a population and the differences equal to 0 are ignored. Calculate the difference for each pair and record the sign of this difference. The Sign test is used to test:

$$H_0 : p = 0.5$$

where p -is the proportion of nonzero observations in the population that are positive. The test statistic S for the Sign test for paired samples is simply

S = the number of pairs with **positive** difference

and S has a binomial distribution with $p = 0.5$ and n = the number of nonzero differences.

After determining the null and alternative hypotheses and finding a test statistic, the next step is to determine the p -value and to draw conclusions based on a decision rule.

Determining p - value for a Sign test

The **p -value** for a Sign test is found using the binomial distribution with n = the number of nonzero differences, S = the number of pairs with **positive** differences and $p = 0.5$.

1. For right tailed test,

$$H_1 : p > 0.5, \quad \text{p-value} = P(x \geq S)$$

2. For left tailed test,

$$H_1 : p < 0.5, \quad \text{p-value} = P(x \leq S)$$

3. For two tailed test,

$$H_1 : p \neq 0.5, \quad 2 \cdot (p - \text{value})$$

Example:

In the study 8 individuals were asked to rate on a scale from 1 to 10 the test of products of two brands: Brand A and Brand B. The scores of the test comparison are shown in the following table

<u>N</u>	<u>Brand A</u>	<u>Brand B</u>
1	5	7
2	3	10
3	4	8
4	9	6
5	8	8
6	5	7
7	6	5
8	9	6

Do the data indicate an overall tendency to prefer the Brand B to the Brand A?

Solution:

First of all, let us calculate differences

<u>N</u>	<u>Brand A</u>	<u>Brand B</u>	<u>Difference (A-B)</u>	<u>Sign of difference</u>
1	5	7	-2	-
2	3	10	-7	-
3	4	8	-4	-
4	9	6	3	+
5	8	8	0	0
6	5	7	-2	-
7	6	5	1	+
8	9	6	3	+

We are discarding those who rated the brands equally. In this example the values for fifth person is omitted in future analysis and the effective sample size is reduced to $n = 7$. The only sample information on which our test is based is that **three** of the seven tasters preferred the brand A. Hence, the value of the Sign test is $S = 3$.

Let p -denotes the true proportion of “+”s in the population. Then the null hypothesis is

$$H_0 : p = 0.5 \text{ There is no overall tendency to prefer one Brand to the other}$$

A one tailed test is used to determine if there is an overall tendency to prefer the Brand B to the Brand A. The alternative hypothesis is that in the population, the majority of preferences are for Brand B. The alternative hypothesis is expressed as

$$H_0 : p < 0.5 \quad \text{Majority prefer the Brand B}$$

The next step is the finding the p -value. If we denote by $P(x)$ the probability of observing x “successes” (“+”s) in $n = 7$ binomial trials, each with probability of success 0.5, then the cumulative binomial probability of observing three or fewer “+”s can be obtained using binomial formula

$$\begin{aligned} p\text{-value} &= P(x \leq 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) = \\ &= C_0^7 (0.5)^0 (0.5)^7 + C_1^7 (0.5)^1 (0.5)^6 + C_2^7 (0.5)^2 (0.5)^5 + C_3^7 (0.5)^3 (0.5)^4 = \\ &= 0.0078 + 0.0547 + 0.1641 + 0.2734 = 0.5000 \end{aligned}$$

For this example p - value is 50%. We are unable to reject the null hypothesis and conclude that data is not sufficient to suggest that population have a preference for Brand B. Since the p -value is the smallest significance level at which the null hypothesis can be rejected, for this example, the null hypothesis can be rejected at 50% or higher. It is unlikely that one would be willing to accept such a high significance level. Again, we conclude that the data is not statistically significant to recommend that Brand B is preferred by majority.

2.2.2. The sign test: Normal approximation (Large samples)

As a consequence of the central limit theorem, the normal distribution can be used to approximate the binomial distribution if the sample size is large. Experts differ on the exact definition of “large”. We suggest that the normal approximation is acceptable if the sample size exceeds 20. With large n , the binomial distribution with $p = 0.5$ is close to the normal distribution with mean $n/2$ and standard deviation $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{0.25 \cdot n} = \sqrt{n/4}$.

The test statistic is

$$T.S. = z = \frac{S - n/2}{\sigma} = \frac{S - n/2}{\sqrt{n/4}}$$

where

S -is the number of positive signs,

n -is the number of nonzero sample observations.

The null hypothesis to be tested is that the proportion p - of nonzero observations in the population that are positive is 0.5; that is

1. To test either null hypothesis

$$H_0 : p = 0.5 \quad \text{against the alternative}$$

$$H_1 : p > 0.5$$

the decision rule is

Reject H_0 if $T.S. > z_\alpha$

2. To test either null hypothesis

$$H_0 : p = 0.5 \quad \text{against the alternative}$$

$$H_1 : p < 0.5$$

the decision rule is

Reject H_0 if $T.S. < -z_\alpha$

3. To test the null hypothesis

$$H_0 : p = 0.5 \quad \text{against the two sided alternative}$$

$$H_1 : p \neq 0.5$$

the decision rule is

$$\text{Reject } H_0 \text{ if } T.S. > z_{\alpha/2} \quad \text{or} \quad T.S. < -z_{\alpha/2}$$

Example:

In a TV commercial, filmed live, 100 persons tested two brands of coffee: say brand A and brand B and each selected their favorite. 56 persons preferred coffee of brand A, 40 preferred coffee of brand B, and four expressed no preference. Test at the 5% significance level the null hypothesis that for the population, there is no overall preference for coffee of brand A over the brand B.

Solution:

From the information given above we obtain that

$$S = 56; \quad n = 96$$

To test if there is no overall preference in this population for one brand of coffee over the other, the hypotheses are

$$H_0 : p = 0.5 \text{ (People have no preference for either brand of coffee)}$$

$$H_1 : p \neq 0.5 \text{ (People have preference for one brand of coffee)}$$

The decision rule is

$$\text{Reject } H_0 \text{ if } \frac{S - n/2}{\sqrt{n/4}} < -z_{\alpha/2} \quad \text{or} \quad \frac{S - n/2}{\sqrt{n/4}} > z_{\alpha/2}$$

The value of the test statistic is

$$\frac{S - n/2}{\sqrt{n/4}} = \frac{56 - 96/2}{\sqrt{96/4}} = -1.633$$

$$\alpha = 0.05; z_{\alpha/2} = 1.96 \text{ and } -z_{\alpha/2} = -1.96.$$

Since -1.633 is not less than -1.96 we fail to reject the null hypothesis. And at 5% significance level we accept that there is no preference for either brand of coffee.

From the standard normal distribution it follows that the approximate

$$p\text{-value} = 0.1032 \text{ or } 10.32\%.$$

Hence the null hypothesis can be rejected at all significance levels greater than 10.32%.

Exercises

1. Two computer specialists estimated the amount of computer memory (in gigabytes) required by five different offices.

Office	Specialist A	Specialist B
1	4.2	6.1
2	6.3	6.7
3	3.1	3.0
4	2.2	2.9
5	7.2	10.9

Use the Sign test to test the null hypothesis that two specialists estimations are the same against the alternative that specialist B estimates higher than specialist A.

2. In a test of two chocolate chip cookie recipes, 13 out of 18 subjects favored recipe A. Using the sign test, find the significance probability when H_1 states that recipe A is preferable.

3. A firm attempting to determine if a difference exists in two manufacturing methods. A sample of 10 workers was selected, and each worker completed the production task using each of the two production methods.

Worker	Method 1(minutes)	Method 2(minutes)
1	10.2	9.5
2	9.6	9.8
3	9.2	8.8
4	10.6	10.1
5	9.9	10.3
6	10.2	9.3
7	10.6	10.5
8	10.0	10.0
9	10.7	10.2
10	10.9	10.2

Use the sign test and perform the null hypothesis that there is no overall preference for one method over the other.

4. A social researcher interviews 25 newly married couples. Each husband and wife are independently asked the question: "How many children would you like to have?" The following data are obtained

<u>Answer of</u>			<u>Answer of</u>		
<u>Couple</u>	<u>Husband</u>	<u>Wife</u>	<u>Couple</u>	<u>Husband</u>	<u>Wife</u>

1	3	2	14	2	1
2	2	2	15	3	2
3	2	1	16	2	2
4	2	3	17	0	0
5	5	1	18	1	2
6	0	1	19	2	1
7	0	2	20	3	2
8	1	3	21	4	3
9	2	2	22	3	1
10	3	1	23	0	0
11	4	2	24	2	3
12	1	2	25	2	2
13	3	3			

Use the Sign test with $\alpha = 0.05$ to test against two sided alternative the null hypothesis that, for the population of families no difference in opinions between husbands and wives.

5. A random sample of 80 sale managers was asked to predict whether next year's sale would be higher than, lower than, or about the same as in the current year. The results are shown below. Test the null hypothesis that the opinion of managers is evenly divided on the question against a two sided alternative.

<u>Prediction</u>	<u>Number</u>
Higher	37
Lower	28

About the same 15

6. Of a random sample of 120 university students, 67 expected to achieve a better GPA than last year, 48 expected a lower GPA than last year, and

5 expected about the same GPA. Do these data present strong evidence that, for population of students they are divided evenly on the expectations, against the alternative that more expect a lower GPA compared with last year?

7. Of a random sample of 150 university instructors, 62 believed that student's skills in solving problems increased over the last decade, 54 believed these skills had deteriorated and 4 saw no change. Evaluate the strength of the sample evidence suggesting that, for all university instructors, teachers are divided evenly on the issue against the alternative that more teachers believe that student's skills in solving problems have improved.

8. In a coffee taste test 48 individuals stated a preference for one of two well-known brands. Results showed 28 favoring brand A, 16 favoring brand B, and 4 undecided. Use the sign test with $\alpha = 0.10$ to test the null hypothesis that there is no difference in the preferences for the two brands of coffee against a two sided alternative.

Answers

1. p - value = 0.1874 or 18.74%; **2.** p -value = 0.0482 or 4.82%;

3. p - value = 0.1798 or 17.98%; **4.** $T.S. = 0.94$; accept H_0 ; **5.** $T.S. = 3.39$;

p -value = 0.06%; **6.** $T.S. = 1.77$; p -value = 3.84%; **7.** $T.S. = 0.74$;

p -value = 22.96%; **8.** $T.S. = 1.81$; reject H_0 .

9.3. The Wilcoxon signed test

9.3.1. The Wilcoxon signed test for paired samples (small sample size)

One disadvantage of the sign test is that it takes into account only a very limited amount information-namely, the signs of the differences. The **Wilcoxon signed rank test** provides a method to use information about the magnitude of the differences between matched pairs. It is still a distribution free test. Like many nonparametric test, it is based on ranks.

Table 2.1

Worker	Method I	Method II	Difference	Absolute value of difference	Rank (+)	Rank (-)
1	10.2	9.5	0.7	0.7	8	
2	9.6	9.8	-0.2	0.2		2
3	9.2	8.8	0.4	0.4		3.5
4	10.6	10.1	0.5	0.5		5.5
5	9.9	10.3	-0.4	0.4		
6	10.2	9.3	0.9	0.9	10	
7	10.6	10.5	0.1	0.1	1	
8	10.0	10.0	0	0	--	
9	11.2	10.6	0.6	0.6	7	
10	10.7	10.2	0.5	0.5	5.5	
11	10.6	9.8	0.8	0.8	9	
					49.5	5.5

To demonstrate the use of the Wilcoxon signed ranked test let us consider a manufacturing firm that is attempting to determine if a difference exists in

two production methods. A sample of 11 workers was selected, and each worker completed the production task using each of the two production methods. Each worker in the sample provides a pair of observations, as shown in Table 2.1. Table 2.1 also provides the difference in the completion times. A positive value indicates that Method I require more time, and a negative value indicates that Method II require more time. The statistical question is whether or not the data indicate that the methods are significantly different in terms of completion times. Thus the null and alternative hypothesis can be written as

H_0 :The two populations of task completion times are identical

H_1 :The two populations of task completion times are not identical

As with the sign test, we ignore any difference of “0”, so sample size in example above is reduced to $n = 10$. The nonzero absolute differences are then ranked in ascending order of magnitude. That is, the smallest absolute value 0.1 is given a rank of “1”. If two or more values are equal, they are assigned the average of the next available ranks. In example above, absolute value of difference-0.4 occurs twice. The rank assigned to them is therefore the average of ranks 3 and 4—that is 3.5. The next absolute value-0.5 occurs twice. The rank assigned to them is therefore the average of ranks 5 and 6—that is 5.5. The next absolute value is assigned rank 7, and so on.

The ranks for positive and negative differences are summed separately. The smaller of these sums is the Wilcoxon Signed Rank Statistic $T.S.$.

Hence $T.S.=5.5$.

We will now suppose that the population distribution of the paired differences is symmetric. The null hypothesis to be tested is that the center of this distribution is 0. In example above, we are assuming that differences in the task completion times have a symmetric distribution, and we want to test whether that distribution is centered on 0—that is no difference between task completion times.

Cutoff points for the distribution of this random variable are given in Appendix (Table 4) for tests against a one sided alternative that the population distribution of the paired differences is specified either to be centered on some number bigger than 0 or to be centered on some number less than 0. For sample size, n , the table shows, for selected probabilities α , the number T_α such that $P(T < T_\alpha) = \alpha$. In other words, the null hypothesis is rejected if $T.S.$ is less than or equal to the corresponding number in the Table4.

In example above, $T.S. = 5.5$. For $n = 10$ we find that the null hypothesis will be rejected for any significance level greater than $\alpha = 0.005$.

Steps in the Wilcoxon Signed Rank test for paired samples

- 1. Calculate the differences**
- 2. Discard (ignore) any difference of “0”**
- 3. Find absolute value of differences**
- 4. Rank the absolute value of differences in ascending order of magnitude. Rank positive and negative differences in two different columns**
- 4. Assign tied absolute differences (if any ties) the average of the ranks they would receive if they were unequal but occurred in successive order**
- 5. Find separately sum of ranks of positive and negative differences.**
- 6. The smaller of the two sums is the Wilcoxon Signed Rank Statistic.**
- 7. Reject the null hypothesis if the value of the test statistic is less than or equal to the value in Appendix table 3.**

9.3.2. The Wilcoxon signed test for paired samples

(large sample size)

When the number of n nonzero differences in the sample is large ($n > 20$), the normal distribution provides a good approximation to the distribution of the Wilcoxon statistic T under the null hypothesis that the population differences are centered on 0.

Let T denote the smaller of the rank sums.

With increasing sample size of n ($n > 20$) nonzero differences, the null hypothesis is that the population differences are centered on 0, Wilcoxon Signed Rank test has mean and variance given by

$$E(T) = \mu_T = \frac{n(n+1)}{4}$$

and

$$Var(T) = \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

For large n , the distribution of the random variable, Z , is approximately standard normal where

$$Z = \frac{T - \mu_T}{\sigma_T}$$

If the number of nonzero differences is large and T is the observed value of the Wilcoxon Signed test statistic, then the following tests have significance level α ,

1. If the alternative hypothesis is one sided, reject the null hypothesis if

$$\frac{T - \mu_T}{\sigma_T} < -z_\alpha$$

2. If the alternative hypothesis is two sided, reject the null hypothesis if

$$\frac{T - \mu_T}{\sigma_T} < -z_{\alpha/2}$$

Example:

A random sample of 38 students who had just completed courses in statistics and accounting was asked to rate each in terms of level of interest, on a scale from one (very uninteresting) to ten (very interesting). The 38 differences in

the pairs of ratings were calculated and the absolute differences ranked. The smaller of the rank sums, which was for those finding accounting the more interesting, was 278. Test at 5 % significance level the null hypothesis that the population of students would rate these courses equally against the alternative that the statistics course is viewed as the more interesting.

Also find the p -value.

Solution:

From the given information

$$n = 38; \quad T = 278$$

The mean and variance of the Wilcoxon statistic are

$$\mu_T = \frac{n(n+1)}{4} = \frac{38 \cdot (38+1)}{4} = 370.5$$

$$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24} = \frac{38 \cdot 39 \cdot 77}{24} = 4754.75$$

So the standard deviation is

$$\sigma_T = 68.95$$

According to the condition, the null and alternative hypothesis can be written as

$$H_0 : \text{both courses rated equally interesting}$$

$$H_1 : \text{statistics course rated more interesting}$$

If T is the observed value of the test statistic, the null hypothesis is rejected against one sided alternative if

$$\frac{T - \mu_T}{\sigma_T} < -z_\alpha$$

Here, the value of T is $T = 278$ and the value of test statistic is

$$\frac{T - \mu_T}{\sigma_T} = \frac{278 - 370.5}{68.95} = -1.34$$

$$\alpha = 0.05; \quad F_z(z_{0.05}) = 0.95; \quad z_{0.05} = 1.65; \quad -z_{0.05} = -1.65$$

Since -1.34 is not less than -1.65 we fail to reject H_0 , and accept it.

The value of α corresponding to $z_\alpha = -1.34$ is, from Table 1 of the Appendix, $(1 - 0.9099) = 0.0901$. Then the null hypothesis can be rejected at all significance levels greater than 9.01%. The data contain modest evidence suggesting that statistics course is more interesting.

Exercises

1. Two critics rate the service at six award winning restaurants on a continuous 0 to 10 scale. Apply Wilcoxon signed rank test with $\alpha = 0.05$ if there is no difference between the critics' ratings?

<u>Restaurant</u>	<u>Critic 1</u>	<u>Critic 2</u>
1	6.2	8.4
2	5.3	5.8
3	7.5	7.1
4	7.4	7.0
5	4.3	5.1
6	9.8	9.9

2. Two computer specialists estimated the amount of computer memory (in gigabytes) required by five different offices

<u>Office</u>	<u>Specialist A</u>	<u>Specialist B</u>
1	5.7	6.1
2	6.4	6.8
3	3.2	3.1
4	2.0	2.9
5	8.1	12.3

Apply Wilcoxon signed rank test with $\alpha = 0.05$ to test the null hypothesis that there is no difference between estimations against a two sided alternative.

3. Twelve customers were asked to estimate the selling price of two models of refrigerators. The estimates of selling price provided by the customers are shown below:

<u>Customer</u>	<u>Model A</u>	<u>Model B</u>
1	\$650	\$900
2	760	720
3	740	690
4	700	850
5	590	920
6	620	800
7	700	890
8	690	920
9	900	1000
10	500	690
11	610	700
12	720	700

Use these data and test at the 0.05 level of significance to determine if there is no difference in the customers' perception of selling price of the two models.

- 4.** A certain brand of microwave oven was priced at 12 stores in two different cities.

These data are presented below:

<u>District A</u>	<u>District B</u>
18 500	16 700
16 000	20 500
12 000	23 000
20 000	17 500
19 000	22000
17 000	21 000
16 500	21 500
19 000	19 500
15 500	17 000
16 000	23 000
17 500	21 000
18 000	22 000

Use a 0.05 level of significance and apply the Wicoxon signed rank test to test whether or not prices for the microwave oven are the same in the two cities.

- 5.** The company is interested in the impact of the newly introduced quality management program on job satisfaction of workers. A random sample of 34 workers was asked to assess level of satisfaction on a scale from 1 to 10 two

month before the program. These same sample members were asked to make this assessment again two month after the introduction of the program. The 34 differences in the pairs of ratings were calculated and absolute differences ranked. The smaller of the rank sums, which was for those more satisfied before the introduction of the program, was 178. What can be concluded from these findings?

6. A random sample of 90 members was taken. Each sample member was asked to assess the amounts of time in a month spent watching TV and the amounts of time in a month spent reading. The 90 differences in times spent were then calculated and their absolute differences ranked. The smaller of the rank sums, which was for watching TV, was 1680. Test the null hypothesis that the population amounts of time spent on watching TV and reading divides equally against the alternative that watching TV takes more amounts of time.

7. Suppose you wish to test hypothesis that two treatments, A and B, are equivalent against the alternative that the responses for A tend to be larger than those of B. If the number of pairs equals 25, and smaller of the rank of the absolute differences is 273, then what would you decide? Use $\alpha = 5\%$ then find p -value for the test and interpret it.

8. An experiment was conducted to compare two print types, A and B, to determine whether type A is easier to read. A sample of 22 persons was given the same material to read. First they read the material printed with type A, then read the same material printed with type B. The times necessary for each person to read the materials (in seconds) were

Type A: 95;122;101;99;108;122;135;127;119;127;99;98;97;96;112;97;100;
116; 111;117;102;103

Type B: 110;102;115;112;120;117;119;127;137;119;99;100;102;103;118;
99;89;97;112;116; 178; 94.

Do the data provide sufficient evidence to indicate that print type A and print type B are the same for reading against the alternative that print type A is easier to read? Test using $\alpha = 0.05$.

Answers

1. T.S. = 7; accept H_0 ; 2. T.S. = 1; reject H_0 ; 3. T.S. = 6; reject H_0 ;
4. T.S. = 3; reject H_0 virtually at any levels; 5. T.S. = -2.04; p-value = 4.12%;
6. T.S. = -1.48; reject H_0 at levels higher than 6.94%; 7. T.S. = 2.97; accept H_0 at any levels; 8. T.S. = -0.71; reject H_0 .

9.4. The Mann-Whitney test

Suppose two independent random samples are to be used to compare two populations. We may be unwilling to make assumptions about the form of the underlying population probability distributions or we may be unable to obtain exact values of the sample measurements. If the data can be ranked in order of magnitude for either of these situations, the *Mann-Whitney* test (sometimes called *Mann-Whitney U test*) can be used to test the hypothesis that the probability distributions associated with the two populations are equivalent.

Assume that apart from any possible differences in central location, that the two population distributions are identical. Suppose that n_1 observations are available from the first population and n_2 observations from the second population. The two samples are pooled and the observations are ranked in ascending order, with ties assigned the average of the next available ranks.

Let R_1 denote the sum of the ranks from the first population. The Mann-Whitney statistic is

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

In testing the null hypothesis that the central locations of the two population distributions are the same, we assume that the two population distributions are identical. It can be shown that if the null hypothesis is true, the random variable U has mean

$$E(U) = \mu_U = \frac{n_1 \cdot n_2}{2}$$

and variance

$$Var(U) = \sigma_U^2 = \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}$$

Then for large sample sizes (both at least 10), the distribution of the random variable,

$$Z = \frac{U - \mu_U}{\sigma_U}$$

is well approximated by the standard normal distribution.

Decision rules for the Mann-Whitney test

Suppose that two population distributions are identical, apart from any possible differences in central location. In testing the null hypothesis the two population distributions have the same central location, the following test have significance level α :

H_0 : Two population distributions have the same central location

1. If the alternative hypothesis is one sided hypothesis that the location of population 1 is higher than the location of population 2, the decision rule is

$$\text{Reject } H_0 \text{ if } \frac{U - \mu_U}{\sigma_U} < -z_\alpha$$

2. If the alternative hypothesis is one sided hypothesis that the location of population 1 is lower than the location of population 2, the decision rule is

$$\text{Reject } H_0 \text{ if } \frac{U - \mu_U}{\sigma_U} > z_\alpha$$

3. If the alternative hypothesis is two sided hypothesis that the two population distributions differ, the decision rule is

$$\text{Reject } H_0 \text{ if } \frac{U - \mu_U}{\sigma_U} < -z_{\alpha/2} \text{ or } \frac{U - \mu_U}{\sigma_U} > z_{\alpha/2}$$

Example:

Let us demonstrate the methodology of the Mann-Whitney test by using it conduct a test on the population of account balances at two branches of some Bank. Data collected from two independent simple random samples, one from each branch, are shown in Table 2.2.

Table2.2

Branch 1		Branch 2	
Sampled <u>Account</u>	Account <u>balance</u>	Sampled <u>account</u>	Account <u>balance</u>
1	1 095	1	885
2	955	2	850

3	1 200	3	915
4	1 195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1 000
9	875	9	1 050
10	1 055	10	935
11	1 025		
12	975		

The first step in the Mann-Whitney test is to rank the *combined* (pooled) data from the two samples from low to high. Using the combined set of 22 observations shown in Table 2.2, the lowest value of \$750(item 6 of sample2) is ranked number 1. Continuing the ranking, we have

<u>Account balance</u>	<u>Item</u>	<u>Rank</u>
750	6 of sample 2	1
800	5 of sample 2	2
805	7 of sample 1	3
.....
1 195	4 of sample 1	21
1 200	3 of sample 1	22

Item 6 of sample 1 and item 4 of sample 2 both have the same account balance, \$950. We could give one of these items a rank 12 and the other a rank 13, but this could lead to an erroneous conclusion. In order to avoid this difficulty the

usual treatment for tied data values is to assign each value the rank equal to the average of the ranks associated with the tied items. Thus the tied observations of \$950 are both assigned ranks of 12.5. Table 2.3 shows the entire data set with the rank of each observation.

Table 2.3

Branch 1			Branch 2		
Sampled	Account		Sampled	Account	
<u>Account</u>	<u>balance</u>	<u>Rank</u>	<u>account</u>	<u>balance</u>	<u>Rank</u>
1	1 095	20	1	885	7
2	955	14	2	850	4
3	1 200	22	3	915	8
4	1 195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1 000	16
9	875	6	9	1 050	18
10	1 055	19	10	935	10
11	1 025	17			
12	975	15			
Sum of ranks		169.5			83.5

The next step in the Mann-Whitney test is to sum the ranks for each sample. These sums are shown in Table 2.3. The test procedure can be based upon the sum of the ranks for either sample. In the following discussion we use the sum of the ranks for the sample from branch 1. We will denote this sum by R_1 . Thus, in our example $R_1 = 169.5$.

The value observed for the Mann-Whitney test is

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1 = 12 \cdot 10 + \frac{12 \cdot 13}{2} - 169.5 = 28.5$$

Since two samples are selected from identical populations and n_1 and n_2 each is 10 or greater, the sampling distribution of U can be approximated by a normal distribution with mean

$$E(U) = \mu_U = \frac{n_1 \cdot n_2}{2} = \frac{12 \cdot 10}{2} = 60$$

and variance

$$Var(U) = \sigma_U^2 = \frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12} = \frac{12 \cdot 10 \cdot 23}{12} = 230$$

Suppose that we want to test the null hypothesis that the central locations of the distributions of account balance are identical against the two-sided alternative for $\alpha = 0.05$. The decision rule is to reject the null hypothesis if

$$\frac{U - \mu_U}{\sigma_U} < -z_{\alpha/2} \text{ or } \frac{U - \mu_U}{\sigma_U} > z_{\alpha/2}$$

Here

$$\frac{U - \mu_U}{\sigma_U} = \frac{28.5 - 60}{\sqrt{230}} = -2.08$$

$$z_{\alpha/2} = z_{0.025} = 1.96 \text{ and } -z_{0.025} = -1.96$$

Since -2.08 is less than -1.96, we reject the null hypothesis that two population account balances are identical. Thus we conclude that two populations are not identical. The probability distribution of account balances at branch 1 is not the same as that at branch 2.

Now, from Table1 of the Appendix, the value of $\alpha/2$ corresponding to a value (-2.08) is 0.0188, so the corresponding α is 0.0376

$$p\text{-value} = 2 \cdot (1 - F_z(\text{test statistics})) = 2(1 - 0.9812) = 0.0376$$

The null hypothesis will be rejected for any significance level higher than 3.76%. Thus, these data do not contain strong evidence against the hypothesis that the central locations of accounts at two branches are the same. There is very strong support that two branches account balances are not identical.

Exercises

1. Starting salaries were recorded for ten recent business administration graduates at each of two well-known universities. Use $\alpha = 0.1$ and test for the difference in the starting salaries from the two universities is zero against the alternative that starting salaries are higher for the university A.

<u>University A</u>		<u>University B</u>	
<u>Student</u>	<u>Monthly salary (\$)</u>	<u>Student</u>	<u>Monthly salary (\$)</u>
1	890	1	1 000
2	950	2	1 020
3	1 200	3	1 140
4	1 150	4	1 000
5	1 300	5	975
6	1 350	6	925
7	990	7	900
8	1 050	8	1 025
9	1 400	9	1 075
10	1 450	10	930

2. The following data show product weights for items produced on two production lines

Line 1: 13.6; 13.8; 14.0; 13.9; 13.4; 13.2; 13.3; 13.6; 12.9; 14.4

Line 2: 13.7; 14.1; 14.2; 14.0; 14.6; 13.5; 14.4; 14.8; 14.5; 14.3; 15.0; 14.9

Test that the difference between the product weights for the two lines is zero against the alternative that product weights of second line is higher.

Use $\alpha = 0.10$. Also find p -value.

3. A random sample of 14 male students and an independent random sample of 16 female students were asked to write essays at the conclusion of a writing course. Their grades were recorded below:

Male: 75; 80; 60; 80; 95; 100; 65; 70; 75; 60; 50; 55; 90; 95

Female: 85; 70; 90; 100; 95; 67; 50; 50; 67; 83; 78; 62; 43; 97; 89; 73

Test the 5% significance level null hypothesis that, in the aggregate the male and female students are equally ranked, against a two-sided alternative. Also find p -value.

4. For a random sample of 12 management department gradates and 14 economics department graduates were asked their starting salaries. Those salaries were then ranked from 1 to 26. The following rankings resulted

Management: 2; 6; 7; 1; 11; 20; 8; 14; 21; 12; 4; 26

Economics: 13; 3; 17; 25; 5; 9; 10; 24; 15; 23; 16; 22; 18; 19

Analyze the data using the Mann-Whitney test, and comment on the results.

5. Starting salaries of graduates from two leading universities were compared. Independent random samples of 40 from each university were taken, and the 80 starting salaries were pooled and ranked. The sum of the ranks for students from one of these universities was 1450. Test the null hypothesis that the central locations of the population distributions are identical against two sided alternative.

6. A stock market analyst produced at the beginning of the year a list of stocks to buy and another list of stocks to sell. For a random sample of ten stocks from the “buy list”, percentage returns over the year were as follows:

10.6; 5.2; 12.8; 16.2; 10.6; 4.3; 3.1; 11.7; 13.9;
11.3

For an independent random sample of ten stocks from the “sell list”, percentage returns over the year were as follows:

-2.6; 6.1; 9.9; 11.3; 2.3; 3.9; -2.3; 1.3; 7.9; 10.8

For $\alpha = 0.05$ use the Mann-Whitney test to interpret these data. Also find and interpret p -value.

Answers

1. T.S. = ; reject H_0 ; 2. T.S. = ; reject H_0 ; p -value = 0.3%;

3. T.S. = ; accept H_0 ; 4. T.S. = ; p - value = 12.36%; H_0 will be rejected at all levels higher than 12.36%; 5. T.S. = ; p -value = 0.101; H_0 will be rejected at any level higher than 10.1%; 6. T.S. = ; reject H_0 at 5%;

p - value = 2.58%.

Chapter 10

Simple linear regression

10.1. Introduction

In day-to-day decisions-making situations, businesspersons and economists frequently draw conclusions and make recommendations based on the relationship between two variables. For example, a marketing manager may project sales volume based upon observed relations between advertising expenditures and sales volume. Although in some instances the manager will rely on his or her intuition as to how the variables are related, the safest approach, by far, is to collect data on the two variables and then evaluate their relationship statistically. These relationships are expressed mathematically as

$$y = f(x)$$

where the function may follow linear and nonlinear forms.

10.2. The scatter diagram

As a first step in determining if a relationship exists between two variables we could plot or graph the available data for the two variables. Suppose that a sales manager has recorded containing data on annual sales and years of experience. The information is given in the following table:

Salesperson	1	2	3	4	5	6	7	8	9	10
Years of experience	1	3	4	4	6	8	10	10	11	13
Annual sales (\$1000's)	80	97	92	102	103	111	119	123	117	136

Let us plot these data on a graph with years of selling experience on the horizontal axis and annual sales on the vertical axis. We now have a **scatter diagram**. It is given this name because the plotted points are “scattered” over the graph or diagram. The scatter diagram for these data is shown in Figure 3.1.

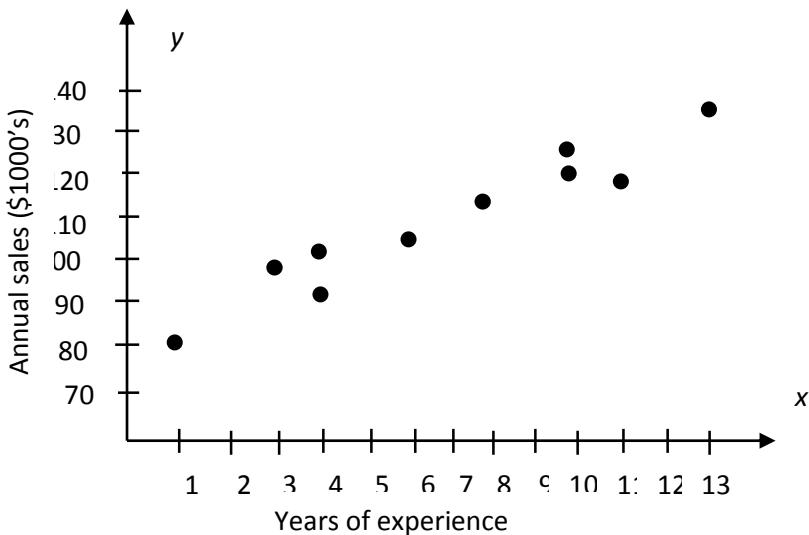


Fig.3.1. Scatter diagram of annual sales and years of

In

regression analysis statisticians commonly will classify a variable as an **independent** or a **dependent** variable. The classification is used to indicate which variable is doing the predicting or explaining (independent variable) and which variable is being predicted or explained (dependent variable). In our example, the years of selling experience is referred to as the independent variable. It is used to predict the sales volume, or dependent variable.

Does the scatter diagram in Fig. 3.1 allow us to draw conclusions?

It gives us an overview of the data. It indicates that in this case there is a good chance that the variables are related. In fact, it appears that the relationship between these two variables may be approximated by a straight line or linear function.

10.3. Correlation analysis

We will introduce some statistical measures that provide greater precision for describing relationships.

Let X and Y be a pair of random variables, with means μ_x and μ_y , and variances σ_x^2 and σ_y^2 . As a measure of the association between these variables, we introduced **the covariance**, defined as

$$Cov(x, y) = S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

where x_i and y_i are the observed values, \bar{X} and \bar{Y} are the sample means, and n is the sample size.

A positive value of the covariance indicates a direct or increasing linear relationships and a negative value of covariance indicates a decreasing linear relationship. Positive association indicates that the high values of X tend to be associated with high values of Y and low X with low Y . When there is a negative association, so that high values of X are associated with low values of Y and low X with high Y , the covariance is negative. If there is no linear association between X and Y , their covariance is 0.

Another measure of the relationship between two variables is the **correlation coefficient**. In this section we will consider the simple linear correlation, for short linear correlation, which measures the strength of the linear association between two variables.

Definition:

The simple linear correlation, denoted by, r_{xy} , measures the strength of the linear relationship between two variables for a sample and is calculated as

$$r_{xy} = \frac{Cov(x, y)}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.1)$$

An equivalent expression is

$$r_{xy} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2 \right) \cdot \left(\sum_{i=1}^n y_i^2 - n \cdot (\bar{y})^2 \right)}} \quad (3.2)$$

- 1.** The sample correlation coefficient ranges from -1 to $+1$ with,
 - a) $r_{xy} = +1$ indicates a perfect positive linear relationship;
 - b) $r_{xy} = 0$ indicates no relationships between X and Y
 - c) $r_{xy} = -1$ indicates a perfect decreasing linear relationship between X and Y .
- 2. Positive correlations** indicate positive or increasing linear relationship with values closer to $+1$, indicating data points closer to a straight line, and closer to 0 , indicating greater deviations from a straight line.
- 3. Negative correlations** indicate negative or decreasing linear relationship with values closer to -1 , indicating data points closer to a straight line, and closer to 0 , indicating greater deviations from a straight line.

If $r_{xy} = 1$, it is said to be a case of perfect positive linear correlation. In such cases, all points in the scatter diagram lie on a straight line that slopes upward from left to right, if $r_{xy} = -1$, the correlation is said to be a perfect negative linear correlation. In this case, all points in a scatter diagram fall on a straight line that slopes downward from left to right.

If the correlation between two variables is positive and close to 1 , we say that the variables have a *strong positive linear correlation*. If the correlation between two variables is positive but close to 0 , then the variables have a *weak positive linear correlation*. On the other hand, if the correlation between two variables is negative and close to -1 , then the variables are said to have a *strong negative linear correlation*. Also, if the correlation between two variables is negative and close to 0 , there exists a *weak negative linear correlation* between the variables.

Example:

An economist is interested in the relationship between food expenditure and income. Calculate the sample correlation coefficient for the data recorded on monthly incomes and food expenditure of seven households.

Household	1	2	3	4	5	6	7
Income (100's of \$)	35	49	21	39	15	28	25
Food expenditure (100's of \$)	9	15	7	11	5	8	9

Solution: The sample means are

$$\bar{x} = \frac{\sum x_i}{n} = \frac{212}{7} = 30.29; \quad \bar{y} = \frac{\sum y_i}{n} = \frac{64}{7} = 9.14$$

The sample correlation coefficient can be calculated either by (3.1) or (3.2)

It is more convenient to use (3.2) to calculate correlation coefficient:

Necessary calculations of the sample correlation for the data are set out in the following table 3.1

Table 3.1

Household	Income (x_i)	Food expenditure (y_i)	$x_i \cdot y_i$	x_i^2	y_i^2
1	35	9	315	1225	81
2	49	15	735	2401	225
3	21	7	147	441	49
4	39	11	429	1521	121
5	15	5	75	225	25
6	28	8	224	784	64
7	25	9	225	625	81
Sums	212	64	2150	7222	646

Hence, the sample correlation is:

$$r_{xy} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2 \right) \cdot \left(\sum_{i=1}^n y_i^2 - n \cdot (\bar{y})^2 \right)}} = \\ = \frac{2150 - 7 \cdot (30.29) \cdot (9.14)}{\sqrt{(7222 - 7 \cdot (30.29)^2) \cdot (646 - 7 \cdot (9.14)^2)}} = \frac{212.05}{221.25} = 0.96$$

The sample correlation, 0.96, indicates very strong positive relationships between monthly income and food expenditure. The high value of monthly income tends to be associated with the higher value of food expenditure.

10.3.1. Hypothesis test for correlation

The sample correlation coefficient r_{xy} is useful as a descriptive measure of the strength of linear association in a sample. We can also use the correlation coefficient to test the null hypothesis that there is no linear association in the population between a pair of random variables; that is

$$H_0 : \rho = 0$$

We can show that when the null hypothesis is true and the random variable have a joint normal distribution then the random variable

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

follows a Student's t distribution with $(n-2)$ degrees of freedom. The following tests of the null hypothesis

$$H_0 : \rho = 0$$

have a significance level of α :

1. To test H_0 against the alternative

$$H_1 : \rho > 0$$

the decision rule is

reject H_0 if $T.S. > t_{n-2,\alpha}$

2. To test H_0 against the alternative

$$H_1 : \rho < 0$$

the decision rule is

reject H_0 if $T.S. < -t_{n-2,\alpha}$

3. To test H_0 against the two sided alternative

$$H_1 : \rho \neq 0$$

the decision rule is

reject H_0 if $T.S. > t_{n-2,\alpha/2}$ or $T.S. < -t_{n-2,\alpha/2}$

where $T.S. = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$, and $t_{n-2,\alpha}$ is the number for which

$$P(t_{n-2} > t_{n-2,\alpha}) = 2$$

where the random variable t_{n-2} follows a Student's t distribution with $(n-2)$ degrees of freedom.

Example:

A sample data set produced the following information

$$\begin{aligned} n &= 10; & \sum x_i &= 66; & \sum y_i &= 588; & \sum x_i y_i &= 2244; \\ \sum x_i^2 &= 396; & \text{and} & & \sum y_i^2 &= 58734 \end{aligned}$$

Find the sample correlation, and test against a two sided alternative the null hypothesis that the population correlation is 0. Take $\alpha = 0.05$.

Solution: Denoting by ρ the population correlation, we want to test

$$H_0 : \rho = 0$$

against the two sided alternative

$$H_1 : \rho \neq 0$$

the decision rule is

$$\text{reject } H_0 \text{ if } T.S. > t_{n-2,\alpha/2} \quad \text{or} \quad T.S. < -t_{n-2,\alpha/2}$$

Firstly, let us find the value of sample correlation coefficient

$$r_{xy} = \frac{2244 - 12 \cdot 5.5 \cdot 49}{\sqrt{(396 - 12 \cdot (5.5)^2) \cdot (58734 - 12 \cdot (49)^2)}} = -\frac{990}{993.69} = -0.996$$

The value of the test statistic is

$$T.S. = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{-0.996 \cdot \sqrt{10-2}}{\sqrt{1-(-0.996)^2}} = -\frac{2.817}{0.0894} = -31.5$$

$$t_{n-2,\alpha/2} = t_{8,0.025} = 2.306 \text{ and } -t_{8,0.025} = -2.306$$

Since $-31.5 < -2.306$ we reject H_0 . Virtually for any level of α we reject hypothesis that there is no association between x and y . These data contain very strong evidence of positive (linear) association between x and y .

Exercises

1. For the data set

x	0	1	6	3	5
y	4	3	0	2	1

- a) Construct a scatter diagram
 - b) Guess the sign and value of the correlation coefficient
 - c) Calculate the correlation coefficient.
2. A sample data set produced the following information.

$$n = 460; \quad \sum x_i = 9880; \quad \sum y_i = 1456;$$

$$\sum x_i y_i = 85080;$$

$$\sum x_i^2 = 485870; \quad \text{and} \quad \sum y_i^2 = 135675$$

Calculate the linear correlation coefficient r_{xy} .

3. Calculations from a data set of $n = 48$ pairs of (x, y) values have provided the following results

$$\sum (x_i - \bar{x})^2 = 260.2; \quad \sum (y_i - \bar{y})^2 = 403.7; \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 298.8$$

Calculate the linear correlation coefficient.

4. The following table gives the experience (in years) and monthly salaries (in hundred of dollars) of nine randomly selected secretaries

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	22	12	15	17	15	19	24	13	27

a) Do you expect the experience and monthly salaries to be positive or negatively related?

b) Compute the correlation coefficient.

c) Test at the 10% significance level, against a two sided alternative, the null hypothesis that the population correlation coefficient is zero.

5. The following data were collected regarding the starting monthly salary and the grade point average (GPA) for students who had obtained a degree in business administration and economics:

GPA	2.6	3.4	3.6	3.2	3.5	2.9
Monthly salary (\$)	900	1200	600	1100	1400	1000

- a) Develop a scatter diagram for the above data.
 b) Compute the sample correlation coefficient between grade point average and salary.
 c) Test at the 5% significance level the null hypothesis that the population correlation coefficient is zero against the alternative that it is positive.
- 6.** The management of a supermarket wanted to check the effect of the number of broadcast on TV on the gross sales at the store. The management experimented for eight weeks by broadcasting a different number of commercials each week on TV. The following table gives the number of commercials during each week and the gross sales (in 1000's of dollars)

Number of commercials	22	16	28	12	30	19	24	32
Gross sales per week	3.64	3.12	4.08	2.84	3.98	3.55	4.02	4.38

- a) Compute the sample correlation coefficient between number of broadcasts and gross sales.
 b) Test the null hypothesis that number of broadcasts and gross sales are uncorrelated in the population against the alternative that population correlation is positive.

Answers

- 1.** b) high negative correlation: c) -0.992; **2.** 8.99; **3.** 0.92; **4.** a) positively; b) 0.95; c) $T.S.=8.051$; fail to reject H_0 ; **5.** b) 0.114; c) $T.S.=0.23$; reject H_0 ; **6.** a) 0.96; b) $T.S.=8.40$; reject H_0 at virtually any level.

10.4. Spearman rank correlation

Suppose that a random sample $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ of n pairs of observations is taken. If x_i and y_i are each ranked in ascending order and the sample correlation of these ranks is calculated, the resulting coefficient is called the **Spearman rank correlation coefficient**. If there are no tied ranks, an equivalent formula for computing this coefficient is

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where the d_i are the differences of the ranked pairs:

$$d_i = \text{rank}(x_i) - \text{rank}(y_i)$$

Remark: Spearman rank correlation shares the properties of r_{xy} that $-1 \leq r_s \leq 1$ and that values near +1 indicate a tendency for the larger values of X to be paired with the larger values of Y .

The following test of the null hypothesis H_0 of no association in the population

$H_0 : x$ and y are independent

have significance level α :

1. To test against the alternative of positive association, the decision rule is

Reject H_0 if $r_s > r_{s,\alpha}$

2. To test against the alternative of negative association, the decision rule is

Reject H_0 if $r_s < -r_{s,\alpha}$

3. To test against the two sided alternative of some association, the decision rule is

Reject H_0 if $r_s > r_{s,\alpha/2}$ or $r_s < -r_{s,\alpha/2}$

The table of critical values of the Spearman rank correlation coefficient is given in the Table 5 of the Appendix.

Example:

Scores that 8 salesmen made on a test that measures their aggressiveness (x), and their sales in thousands of dollars for their second year with a certain company (y).

x (aggressiveness)	30	17	35	28	42	25	19	34
y (sales)	35	31	40	46	50	32	33	42

- a) Find and interpret Spearman rank correlation
- b) Test the null hypothesis that aggressiveness and sales are independent again the alternative that they are positively correlated. Take $\alpha = 0.05$.

Solution:

- a) First of all, let us rank **separately** x and y in ascending order. These two rank appear in third and fourth columns of the following table 3.2

Table 3.2

x	y	Rank x_i	Rank y_i	$d_i = x_i - y_i$	d_i^2
30	35	4	5	1	1
17	31	8	8	0	0
35	40	2	4	-2	4
28	46	5	2	3	9
42	50	1	1	0	0
25	32	6	7	-1	1
19	33	7	6	1	1
34	42	3	3	0	0
sum					16

The differences between ranks and squared differences between ranks are shown in the last two columns of the table. Substituting the values $n = 8$ and $\sum d_i^2 = 16$ into formula for Spearman rank correlation, we obtain

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 16}{8 \cdot 63} = 1 - 0.19 = 0.81$$

It means that there exists strong positive correlation between aggressiveness and sales volume.

- b) The null and alternative hypotheses are

$$H_0 : x \text{ and } y \text{ are independent}$$

$$H_1 : x \text{ and } y \text{ are positively correlated}$$

The decision rule is

Reject H_0 if $r_s > r_{s,\alpha}$

For a sample of size $n=8$, and $\alpha = 0.05$,

$$r_{s,\alpha} = r_{8,0.05} = 0.643$$

Since $0.81 > 0.643$ we reject H_0 , and accept the alternative hypothesis that x and y are positively correlated.

Exercises

1. Specify the rejection region for Spearman's nonparametric test for rank correlation in each of the following cases

- a) $H_0 : \rho = 0$; $H_1 : \rho \neq 0$; $n = 10$; $\alpha = 0.05$
- b) $H_0 : \rho = 0$; $H_1 : \rho > 0$; $n = 20$; $\alpha = 0.025$
- c) $H_0 : \rho = 0$; $H_1 : \rho < 0$; $n = 30$; $\alpha = 0.01$

2. Compute Spearman's rank correlation coefficient for each of the following pairs of sample observations

- a)
- b)

x	33	61	20	19	40
y	26	36	65	25	35

x	5	20	15	10	3
y	80	83	91	82	87

3. A random sample of nine pairs of observations are recorded on two variables, x and y .

x	19	27	15	35	13	29	16	22	17
y	12	19	7	25	11	10	16	13	18

Do the data provide sufficient evidence to indicate that ρ , the rank correlation between x and y , differs from zero? Test using $\alpha = 0.05$.

4. Two expert wine testers were asked to rank six brands of wine. Their rankings are shown in the table.

Brand	Expert 1	Expert 2
A	6	5
B	5	6
C	1	2

D	3	1
E	2	4
F	4	3

Do the data present sufficient evidence to indicate a positive correlation in the rankings of the two experts?

5. Refer to the data of Exercise 6 of the previous section. Find Spearman's rank correlation coefficient, and use it to test, against two sided alternative, the null hypothesis of no association in the population between this pair of random variables.

6. Refer to the data of Exercise 5 of the previous section. Find Spearman's rank correlation coefficient, and use it to test the null hypothesis that these quantities are uncorrelated in the population against the alternative that population correlation is negative.

Answers

1. a) $r_s > 0.648$ or $r_s < -0.648$; b) $r_s > 0.450$; c) $r_s < -0.432$; **2.** a) $r_s = 0.4$; b) $r_s = 0.2$; **3.** $r_s = 0.48$; accept H_0 ; **4.** $r_s = 0.66$; reject H_0 virtually at any level; **5.** $r_s = 0.93$; reject H_0 virtually at any level; **6.** $r_s = 0.143$; can not reject H_0 at 5% level.

10.5. The linear regression model

Let us return to the example of an economist investigating the relationship between food expenditure and income. What factors or variables does a household consider when deciding how much money should be spent on food every week or every month? Certainly, income of household is one factor. Many other factors, say, the size of household, the preferences and tests of household members, are some of the variables that will influence a household's decision about food expenditure. These variables are called **independent variables** because they are all vary independently and they explain the variation in food expenditure among different households. In other words, these variables explain why different households spend different amounts of money on food. Food expenditure is called the **dependent variable** because it depends on the independent variables. Studying the effect

of two or more independent variables on a dependent variable using regression analysis is called **multiple regression**. If we choose only one (usually the most important) independent variable and study the effect of that single variable on a dependent variable, it is called a **simple regression**. Thus, simple regression includes only two variables: one independent and one dependent.

Definition: A regression model is a mathematical equation that describes relationship between two or more variables. A simple regression model includes only two variables: one independent and one dependent. The dependent variable is the one being explained and the independent variable is the one used to explain the variation in the dependent variable.

The relationship between two variables in a regression analysis is expressed by a mathematical equation called a **regression equation or model**.

A regression equation that gives a straight line relationship between two variables is called a **linear regression model**; otherwise, it is called a **nonlinear regression model**. In this chapter we will consider only linear regression model.

In a regression model, the independent variable is usually denoted by x and the dependent variable is usually denoted by y . Simple linear regression model is written as

$$y = \alpha + \beta x \quad (1)$$

In model (1), α gives the value of y for $x=0$, and β gives the change in y due to a change of one unit in x . This model simply states that y is determined exactly by x and for a given value of x there is one and only one value of y . For example, if y is food expenditure and x is income, then model (1) would state that food expenditure is determined by income only and that all households with the same income will spend the same amount on food. But as mentioned above, food expenditure is determined by many variables, only one of which is included in model (1). In reality, different households with the same income spend different amounts of money on food because of the differences in size of the household, their preferences and tastes. Hence, to take these variables into consideration and make model complete, we add another term to the right side of model (1). This term is called the **error term**. It is denoted by ε (Greek letter *epsilon*). The complete regression model is written as

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad (2)$$

Equation (2) is called the **population (or true) regression line** of y on x . In equation (2) α and β are the population model coefficients and ε is a random error term .

Population data are difficult to obtain. As a result, we almost always use sample data to estimate model (2). The estimated regression model is given by the equation

$$y_i = a + b \cdot x_i + e_i$$

where a and b are estimated values of the coefficients and e is the difference between the predicted value of y on the regression line, defined as

$$\hat{y}_i = a + b \cdot x_i$$

and the observed value y_i . The difference between y_i and \hat{y}_i for each value of x is defined as the residual

$$e_i = y_i - \hat{y}_i = y_i - (a + b \cdot x_i)$$

Thus for each observed value of x there is a predicted value of y from the estimated model and an observed value. The difference between the observed and predicted values of y is defined as the residual. The residual, e_i , is not the model error, ε , but is the combined measure of the model error and errors in estimating, a and b , and in turn the errors in estimating the predicted value.

10.5. 1. Least squares coefficient estimators

The population regression line is useful theoretical construct, but for applications we need to determine an estimate of the model using available data. Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \dots$

$\dots (x_n, y_n)$. We would like to find the straight line that best fits these points. To do this we need estimators of unknown coefficients α and β of the population regression line.

We obtain the coefficient estimators, a and b using equations derived by using the least squares procedure. As shown in Figure 3.2 there is a deviation, e_i

between the observed, y_i and the predicted value, \hat{y}_i , on the estimated regression equation for each value of x , where $e_i = \hat{y}_i - y_i$.

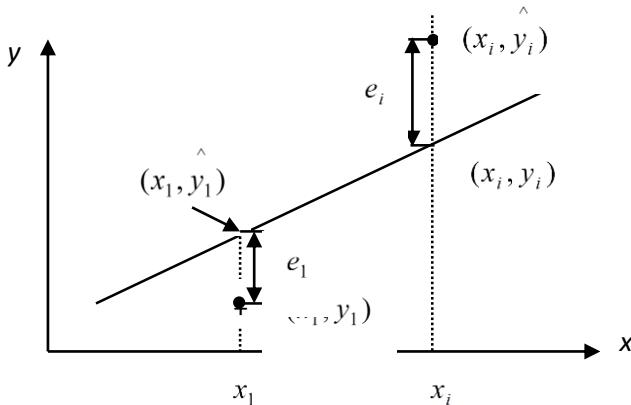


Fig. 3.2

Some of the e_i will be positive and some negative. We then compute a mathematical function that represents the effect of squaring all of the residuals and computing the sum of the squared residuals. This function- whose left side is labeled SSE –includes the coefficients, a and b . The quantity SSE is defined as the “Error Sum of Squares”. The coefficient estimators a and b are selected as the estimators that minimize the Error Sum of Squares.

105.2. Least square procedure

The least square procedure obtains estimates of the linear equation coefficients, a and b , in the model

$$\hat{y}_i = a + b \cdot x_i$$

by minimizing the sum of the squared residuals e_i

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

The coefficients a and b are chosen so that the quantity

$$SSE = \sum e_i^2 = \sum (y_i - (a + bx_i))^2$$

is minimized. It can be shown that the resulting estimates are

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2}$$

and

$$a = \bar{y} - b \cdot \bar{x}$$

where \bar{x} and \bar{y} are the respective sample means.

The line

$$\hat{y} = a + b \cdot x$$

is called the **sample regression line** or **the least squares regression line** of y on x .

Example:

Find the least squares regression line for the data on incomes (in hundreds of dollars) and food expenditures of seven households given in the table below.

Household	1	2	3	4	5	6	7
Income x	35	49	21	39	15	28	25
Food expenditure y	9	15	7	11	5	8	9

Use income as an independent variable and food expenditure as a dependent variable.

Solution:

We are to find the values of a and b for the regression model $\hat{y}_i = a + b \cdot x_i$. The following table shows the calculations required for the computations of a and b .

Using data from the table 3.3 we find

$$\bar{x} = \frac{212}{7} = 30.2857;$$

$$\bar{y} = \frac{64}{7} = 9.1429$$

Table3.3

Household	Income (x_i)	Food expenditure (y_i)	$x_i \cdot y_i$	x_i^2
1	35	9	315	1225
2	49	15	735	2401
3	21	7	147	441
4	39	11	429	1521
5	15	5	75	225
6	28	8	224	784
7	25	9	225	625
Sums	212	64	2150	7222

$$b = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2} = \frac{2150 - 7 \cdot (30.2857) \cdot (9.1429)}{7222 - 7 \cdot (30.2857)^2} = 0.2642$$

$$a = \bar{y} - b \cdot \bar{x} = 9.1429 - (0.2642) \cdot (30.2857) = 1.1414$$

Thus, our estimated regression model $\hat{y} = a + b \cdot x$ is

$$\hat{y} = 1.1414 + 0.2642 \cdot x$$

This regression line is called the least squares regression line. It gives the *regression of food expenditure on income*.

Using this estimated model, we can find the predicted value of y for a specific value of x . For example, suppose that we randomly select a household whose monthly income is \$3500 so that $x = 35$ (x denotes income in hundred of dollar in our example). The predicted value of food expenditure for this household is

$$\hat{y} = 1.1414 + 0.2642 \cdot 35 = \$10.3884 \text{ hundred}$$

In other words, based on our regression line, we predict that a household with a monthly income of \$3500 is expected to spend \$1038.84 per month on food.

10.5.3. Interpretation of a and b

a) Interpretation of a

Consider a household with zero income. Using the estimated regression line obtained above, the predicted value of y for $x=0$ is

$$\hat{y} = 1.1414 + 0.2642 \cdot 0 = \$1.1414\text{hundred}$$

Thus, we can state that a household with no income is expected to spend \$114.4 per month on food. We should be very careful while making this interpretation of a . In example of seven households, the incomes vary from a minimum of \$1500 to a maximum of \$4900. Hence, our regression line is valid only for the values of x between 15 and 49. If we predict y for a value of x outside this range, the prediction usually will not hold true. Thus, since $x=0$ is outside the range of household incomes that we have in the sample data, the prediction that a household with zero income spends \$114.14 per month on food does not carry much credibility.

b) Interpretation of b

The value of b in a regression model gives the change in y (dependent variable) due to a change of one unit in x (independent variable).

For example, by using the regression line $\hat{y} = 1.1414 + 0.2642 \cdot x$

$$\text{when } x = 30; \quad \hat{y} = 1.1414 + 0.2642 \cdot 30 = 9.0674$$

$$\text{when } x = 31; \quad \hat{y} = 1.1414 + 0.2642 \cdot 31 = 9.3316$$

Hence, when x increased by one unit, from 30 to 31, \hat{y} increased by $9.3316 - 9.0674 = 0.2642$, which is the value of b . Because of unit of measurement in hundred of dollars, we can state that, on average, a \$100 increase in income will cause a \$26.42 increase in food expenditure. We can also state that, on average, a \$1 increase in income of household will increase the food expenditure by \$0.2642.

Note that when b is positive, an increase in x will lead to an increase in y and decrease in x will lead to a decrease in y . Such a relationship between x and y is called a positive linear relationship. On the other hand, if the value of b is

negative, an increase in x will cause a decrease in y and a decrease in x will cause an increase in y . Such a relationship between x and y is called a negative linear relationship.

The values of y - intercept and slope calculated from sample data on x and y are called estimated values of α and β and denoted by a and b . Using a and b we can write estimated model as

$$\hat{y} = a + b \cdot x$$

where \hat{y} (read as y hat) is the **estimated** or **predicted** value of y for a given value of x .

105.4. Assumptions of the regression model

Like any other theory, the linear regression analysis is also based on certain assumptions. Consider the population regression model

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

There are four assumptions made about this model.

Assumption1: The random error term ε has a mean equal to zero for each x . In other words, among all households with the same income, some spend more than predicted food expenditure; others spend less than predicted food expenditure. Some of positive errors equal to the sum of negative errors so that the mean of errors for all households with the same income is zero.

Assumption 2: The errors associated with different observations are independent. According to this assumption, the errors for any two households are independent. All households decide independently how much spend on food.

Assumption 3: For any given x , the distribution of errors is normal. In other words, food expenditure for all households with the same income are normally distributed.

Assumption 4: The distribution of population errors for each x has the same (constant) standard deviation, which is denoted by σ_ε . This assumption indicates that the spread of points around the regression line is similar for all x values.

Exercises

1. Plot the following straight lines. Give the values of the y -intercept and slope for each of these lines and interpret them. Indicate whether each of the lines gives a positive or negative relationships between x and y .

$$\text{a) } y = 53 + 7x; \quad \text{b) } y = 75 - 6x$$

2. The following information is obtained from a sample data

$$n=10; \quad \sum_{i=1}^{10} x_i = 100; \quad \sum_{i=1}^{10} y_i = 220; \quad \sum_{i=1}^{10} x_i y_i = 3680; \quad \sum_{i=1}^{10} x_i^2 = 1140$$

Find the estimated regression line.

3. Computing from a data set of (x, y) values we obtained the following summary statistics

$$n=14; \quad \bar{x} = 3.5; \quad \bar{y} = 5.1; \quad \sum_{i=1}^{14} (x_i - \bar{x})^2 = 10.82;$$

$$\sum_{i=1}^{14} (x_i - \bar{x})(y_i - \bar{y}) = 2.677; \quad \sum_{i=1}^{14} (y_i - \bar{y})^2 = 2.01$$

Obtain the equation of the estimated regression line.

4. Given the five pairs of (x, y) values,

x	0	1	6	3	5
y	4	3	0	2	1

a) Construct a scatter diagram

b) Calculate the least squares estimates a and b .

c) Determine the fitted line and draw the line on the scatter diagram.

5. A researcher took a sample of 36 electronic companies and found the following relationship between x and y where x is the amount of money (in thousands of dollars) spent on advertising by a company during a year and y represents the total gross sales (in thousands of dollars) of that company in that year.

$$\hat{y} = 5.6 + 22.5 \cdot x$$

a) An electronic company spent 2000\$ on advertising during a year. What are its expected gross sales for that year?

b) Suppose five electronic companies spent 2000\$ each on advertising during that year. Do you expect these five companies to have the same actual gross sales for that year? Explain.

6. An economist wanted to determine whether or not the amount of phone bills and income of households are related. The following table gives information on the monthly incomes (in hundreds of dollars) and monthly telephone bills (in dollars) for a random sample of 10 households

Income	16	45	36	32	30	13	41	15	36	40
Phone bill	35	78	102	56	75	26	130	42	59	85

a) Find the regression line with income as an independent variable and the amount of the phone bill as a dependent variable.

b) Give an interpretation of the values of a and b calculated in part a.

c) Estimate the amount of the monthly phone bill for a household with a monthly income of \$2500.

7. An auto manufacturing company wanted to investigate how the price of one of its car models depreciates with age. The research department at the company took a sample of 9 cars of this model and collected the following information on the ages (in years) and prices (in hundreds of dollars) of these cars.

Age	8	3	7	10	3	5	6	9
Price	16	74	38	21	98	56	49	30

a) Construct a scatter diagram for these data. Interpret your results.

b) Find the regression line with price as a dependent variable and age as an independent variable.

c) Give a brief interpretation of the values of a and b calculated in part b.

d) Predict the price of a 4 year-old car of this model.

e) Estimate the price of a 19-year-old car of this model. Comment on this finding.

8. Construct a scatter diagram for the data in the following table

x	0.5	1	1.5
y	2	1	3

a) Plot the following two lines on your scatter diagram

- 1) $y = 3 - x$ and 2) $y = 1 + x$
- b) Which of these lines would you choose to characterize the relationship between x and y ? Explain
- c) Show that the sum of errors for both of these lines equals 0.
- d) Which of these lines has smaller SSE?
- e) Find the least squares regression line for the data and compare it to two lines described in part a.

Answers

2. $\hat{y} = -83.714 + 10.571 \cdot x$; 3. $\hat{y} = 4.225 + 0.247 \cdot x$; 4.c) $\hat{y} = 3.845 - 0.615 \cdot x$;
;
5. a)\$50.6 thousand; b)different amounts; 6.a) $\hat{y} = 2.3173 + 2.1869 \cdot x$;
c) \$56.99; 7. $\hat{y} = 111 - 9.84 \cdot x$; 8. b) The second line; d) The second line;
e) $\hat{y} = 1 + x$.

3.6. The explanatory power of a linear regression equation

In Figure 3.3 it is shown that the deviation of an individual y value from its mean can be

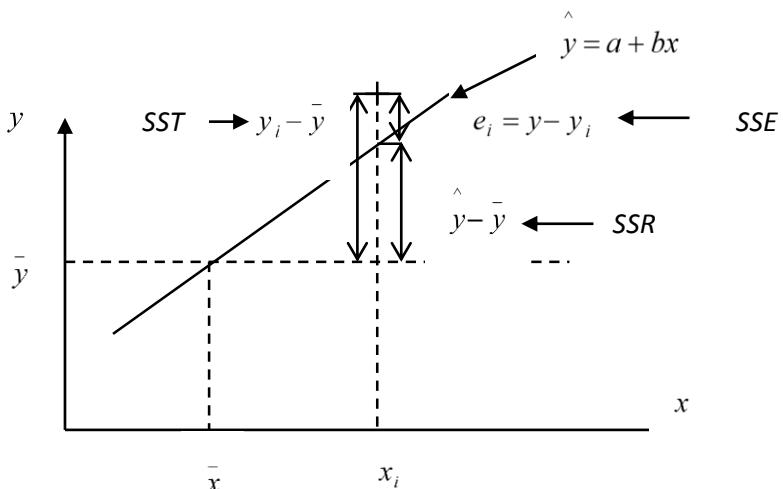


Fig.3.3

partitioned into deviation of the predicted value from the mean and the deviation of the observed value from the predicted value

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y} - \bar{y})$$

We square each side of the equation-because the sum of deviations about the mean is equal to zero-and sum the results over all n points

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y})^2 + \sum_{i=1}^n (\hat{y} - \bar{y})^2$$

Some of you may note the squaring of the right-hand side should include the cross product of the two terms in addition to their squared quantities. It can be shown that the cross predicted term goes to zero. This equation is expressed as

$$SST = SSR + SSE$$

We see that the total variability-SST- consists of two components-SSR-the amount of variability explained by the regression equation- named

“Regression Sum of Squares” and $-SSE$ -random or unexplained deviation of points from the regression line-named “Error Sum of Squares”. Thus

Total sum of squares: $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

Regression Sum of Squares: $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2$

Error	Sum	of
	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2 = \sum_{i=1}^n (e_i)^2$	

For a given set of observed values of the dependent variables, y , the SST is fixed as the total variability of all observations from the mean. We see that in the partitioning larger values of SSR and hence smaller value of SSE indicate a regression equation that “fits” or comes closer to the observed data. This partitioning is shown graphically in Figure 3.3.

Example:

Let us find SST , SSR and SSE for the data on incomes and food expenditure. Using calculation given in the table 3.3 we find the value of total sum of squares as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^7 y_i^2 - \frac{\left(\sum_{i=1}^7 y_i \right)^2}{n} = 646 - \frac{64^2}{7} = 60.8571$$

Table 3.4

x	y	\hat{y}	y^2	e_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	e_i^2
35	9	10.3884	81	-1.3884	4.7143	22.2246	1.9277
49	15	14.0872	225	0.9128	18.7143	350.225	0.8332
21	7	6.6896	49	0.3104	-9.2857	86.2242	0.0963
39	11	11.4452	121	-0.4452	8.7143	75.9390	0.1982
15	5	5.1044	25	-0.1044	-15.286	233.653	0.0109

28	8	8.5390	64	-0.5390	-2.2857	5.2244	0.2905
25	9	7.7464	81	1.2536	-5.2857	27.9386	1.5715
		646				801.429	4.9283

The error sum of squares SSE is given in the sum of the eight column in Table 3.4. Thus,

$$SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (e_i)^2 = 4.9283$$

The regression sum of squares can be found from $SST = SSR + SSE$.

Thus

$$SSR = SST - SSE = 60.8571 - 4.9283 = 55.9288.$$

The value of SSR can also be computed by using the formula.(Check!!)

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2 .$$

The total sum of squares SST is a measure of the total variation in food expenditures, SSR is the portion of total variation explained by the regression model (or by income), and the error sum of squares SSE is the portion of total variation not explained by the regression model.

3.6.1. Coefficient of determination R^2

If we divide both side of the equation

$$SST = SSR + SSE$$

by SST , we obtain

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

We have seen that the fit of the regression equation to the data is improved as

SSR increases and SSE decreases. The ratio $\frac{SSR}{SST}$ provides a descriptive measure of the proportion or percent of the total variability that is explained by the regression model. This measure is called the *coefficient of determination*-or more generally R^2 .

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The coefficient of determination is often interpreted as the percent of variability in y that is explained by the regression equation. We see that R^2 increases directly with the spread of the independent variable.

R^2 can vary from 0 to 1 since SST is fixed and $0 < SSE < SST$. A larger R^2 implies a better regression, everything else being equal.

Interpretation of R^2 : About $(100 \cdot R^2)\%$ of the sample variation in y (measured by the total sum of squares of deviations of the sample y values about their mean \bar{y}) can be explained by using x to predict y in the straight line model.

Example:

Calculate the coefficient of determination for the data on monthly incomes and food expenditures of seven households.

Solution:

From earlier calculations

$$SSR = 55.9288 \text{ and } SST = 60.8571$$

Hence,

$$R^2 = \frac{SSR}{SST} = \frac{55.9288}{60.8571} = 0.92$$

We can state that 92% of the variability in y is explained by linear regression, and the linear model seems very satisfactory in this respect. In other words, we can state that 92% of the total variation in food expenditures of households occurs because of the variation in their incomes, and the remaining 8% is due to other variables, like differences in size of the household, preferences and tastes and so on.

3.6.2. Estimation of model error variance

When we consider income and food expenditures, all households with the same income are expected to spend different amounts on food. Consequently, the random error ε_i will have different values for these households. The

variance σ_i^2 measures the spread of these errors around the population

regression line. Note that σ_i denotes the variance of errors for the population.

However, usually σ_i is unknown. In such cases, it is estimated by s_e^2 , which is the standard deviation of errors for the sample data.

An estimator for the variance of the population model error is

$$\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SSE}{n-2}$$

Division by $(n-2)$ instead of $(n-1)$ results because the simple regression model uses two estimated parameters, a and b , instead of one.

The formula for SSE is

$$SSE = \sum_{i=1}^n (e_i)^2 = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

If we introduce the following notations

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

$$SS_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

$$SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}$$

where SS stands for “sums of squares”, then

$$SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}.$$

Exercises

- 1.** The following information is obtained from a sample data set

$$n=12; \quad \sum_{i=1}^{12} x_i = 66; \quad \sum_{i=1}^{12} y_i = 588; \quad \sum_{i=1}^{12} x_i y_i = 2244;$$

$$\sum_{i=1}^{12} x_i^2 = 396; \quad \text{and} \quad \sum_{i=1}^{12} y_i^2 = 58734$$

Find the values of s_e^2 and R^2 .

2. A sample data set produced the following information

$$n=460; \quad \sum_{i=1}^{460} x_i = 3920; \quad \sum_{i=1}^{460} y_i = 2650; \quad \sum_{i=1}^{460} x_i y_i = 26570;$$

$$\sum_{i=1}^{460} x_i^2 = 48530; \quad \text{and} \quad \sum_{i=1}^{460} y_i^2 = 39347$$

Find the values of s_e^2 and R^2 .

3. Computing from a data set of (x, y) values produced the following summary statistics

$$n=14; \quad \bar{x}=3.5; \quad \bar{y}=2.32;$$

$$SS_{xx}=10.82; \quad SS_{xy}=2.677; \quad SS_{yy}=1.035$$

a) Obtain equation of the best fitting straight line.

b) Estimate σ_e^2 .

4. Computing from a data set of (x, y) values produced the following summary statistics

$$n=14; \quad \bar{x}=1.2; \quad \bar{y}=5.1;$$

$$SS_{xx}=14.10; \quad SS_{xy}=2.31; \quad SS_{yy}=2.01$$

Determine the proportion of variation in y that is explained by linear regression.

5. A calculation shows that $SS_{xx}=10.1$, $SS_{yy}=16.5$, and $SS_{xy}=9.3$, determine the proportion of variation in y that is explained by linear regression.

6. The following table lists the sizes of offices (in hundreds of square meters) and the rents (in dollars) paid for those offices.

Size of offices	22	17	19	28	35	24
Monthly rent	710	590	730	880	1080	820

- a) Find the regression line $y = a + bx$ with the size of an office as an independent variable and monthly rent as a dependent variable.
- b) Give a brief interpretation of the values of a and b .
- c) Predict the monthly rent for the office with 2400 square meters.
- d) One of the offices is 2600 square meters and its rent is \$850. What is the predicted rent for this office? Find the error for this office.
- e) Compute the standard deviation of errors.
- f) Calculate the coefficient of determination. What percentage of the variation in monthly rents explained by the sizes of the offices? What percentage of this variation is not explained?

7. Refer to exercise 7 of previous chapter. The following table which gives the ages (in years) and prices (in hundred of dollars) of eight cars of specific model, is reproduced from that exercise.

Age	8	3	7	10	3	5	6	9
Price	16	74	38	21	98	56	49	30

- a) Calculate the standard deviation of errors.
- b) Compute the coefficient of determination and give a brief interpretation of it.

Answers

- 1.** 22.2; 0.99; **2.** 50.06; 0.04; **3.** a) $\hat{y} = 1.454 + 0.247 \cdot x$; b) $s_e^2 = 0.031$;
4. 0.188; **5.** 0.5190; **6.** a) $\hat{y} = 194 + 25.1 \cdot x$ e) 40.2; f) 0.953; ;
7. a) 11.39; b) 0.856

3.7. Statistical inference: Hypothesis tests and confidence intervals

One of the main purposes for determining a regression line is to find the true value of the slope β of the population regression line. However, in almost all cases, the regression line is estimated using sample data. Then based on the sample regression line, inferences are made about the population regression

line. The slope b of a sample regression line is a point estimator of the slope β of the population regression line. The different sample regression lines estimated for different samples taken from the same population will give different values of b . If only one sample is selected, then the value of b will depend on which elements are included in the sample. Thus, b is a random variable and it possesses a probability distribution called a sampling distribution.

Assume that assumptions 3.5.4 are hold. Then b is an unbiased estimator of β and has a population variance

$$\sigma_b^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2}$$

and unbiased estimator of σ_b^2 is provided by

$$s_b^2 = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_e^2}{\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2} = \frac{s_e^2}{SS_{xx}}$$

In applied regression analysis we first would like to know if there is a relationship. We see that if β is zero then there is no relationship- y would not continuously increase or decrease with increase in x .

3.7.1. Hypothesis testing about β

Let β be a population regression slope and b its least square estimate based on n pairs of sample observations. Assume that assumptions 3.5.4 hold and also assume that the errors ε_i are normally distributed. Then the random variable

$$t = \frac{b - \beta}{s_b}$$

is distributed as Student's t distribution with $(n - 2)$ degree of freedom.

If we use notation

$$T.S. = t = \frac{b - \beta}{s_b}$$

for the test statistic then the following tests have a significance level α

1. To test either null hypothesis

$$H_0 : \beta = \beta_0 \text{ or } H_0 : \beta \leq \beta_0$$

against the alternative

$$H_1 : \beta > \beta_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-2,\alpha}$

2. To test either null hypothesis

$$H_0 : \beta = \beta_0 \text{ or } H_0 : \beta \geq \beta_0$$

against the alternative

$$H_1 : \beta < \beta_0$$

the decision rule is

Reject H_0 if $T.S. < -t_{n-2,\alpha}$

3. To test null hypothesis

$$H_0 : \beta = \beta_0$$

against the two sided alternative

$$H_1 : \beta \neq \beta_0$$

the decision rule is

Reject H_0 if $T.S. > t_{n-2,\alpha/2}$ or $T.S. < -t_{n-2,\alpha/2}$

Remark1: To test the hypothesis that x does not determine y linearly and there is no linear relationship, we will test the null hypothesis that the slope of the regression line is zero, that is $H_0 : \beta = \beta_0 = 0$; the alternative hypothesis that $H_1 : \beta \neq \beta_0 \neq 0$ means x determines y linearly; $H_1 : \beta > \beta_0 = 0$ means x determines y positively; $H_1 : \beta < \beta_0 = 0$ means x determines y negatively.

Remark2: The null hypothesis does not always have to be $\beta = 0$. We may test the null hypothesis that β is equal to a value different from zero.

Example:

Test at the 5% significance level if the slope of the population regression line for the example on incomes and food expenditure of seven households is positive.

Solution:

From earlier calculations we have

$$n = 7; \quad b = 0.2642 \quad \text{and} \quad s_e = 0.9922$$

$$s_b^2 = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{0.9856}{801.429} = 0.001229; \text{ and } s_b = 0.0350.$$

We are to test whether or not slope β of the population regression line is positive. The two hypotheses are

$$H_0 : \beta = \beta_0 \text{ (Slope is zero)}$$

$$H_1 : \beta > \beta_0 \text{ (Slope is positive)}$$

The decision rule is

reject H_0 if $T.S. > t_{n-2,\alpha}$.

The value of the test statistic is

$$T.S. = t = \frac{b - \beta}{s_b} = \frac{0.2642 - 0}{0.0350} = 7.549$$

The significance level is 0.05. Therefore,

$$t_{n-2,\alpha} = t_{5,0.05} = 2.015$$

The value of the test statistic $T.S. = 7.549$ is greater than the critical value of $t = 2.015$ and it falls in the rejection region. Hence, we reject the null hypothesis and conclude that x (income) determines y (food expenditure)

positively. That is, food expenditure increases with an increase in income and it decreases with a decrease in income.

3.7.2. Confidence intervals for the population regression slope β

We can derive confidence intervals for the slope β of the population regression line by using coefficient b and variance estimators we have developed.

If the assumptions 3.5.4 hold, and if the regression errors, ε_i , are normally distributed, then $100(1-\alpha)\%$ confidence interval for the population regression slope β is given by

$$b - t_{n-2,\alpha/2} \cdot s_b < \beta < b + t_{n-2,\alpha/2} \cdot s_b$$

where $t_{n-2,\alpha/2}$ is the number for which $P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$ and the random variable t_{n-2} follows Student's t distribution with $(n - 2)$ degrees of freedom.

Example:

Construct a 95% confidence interval for β for the data on incomes and food expenditures of seven households.

Solution:

From earlier calculations we have

$$n = 7; \quad b = 0.2642; \quad s_e = 0.9922 \text{ and } s_b = 0.0350$$

The confidence level is 95%. So

$$100(1-\alpha)\% = 95\%$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

$$t_{n-2,\alpha/2} = t_{5,0.025} = 2.571$$

The 95% confidence interval for β is

$$b - t_{n-2,\alpha/2} \cdot s_b < \beta < b + t_{n-2,\alpha/2} \cdot s_b$$

$$0.2642 - 2.571 \cdot 0.0350 < \beta < 0.2642 + 2.571 \cdot 0.0350$$

$$0.17 < \beta < 0.35$$

Thus, we are 95% confident that slope β for the population regression line is

between 0.17 and 0.35.

Exercises

1. The following information is obtained for a sample of 16 observations taken from a population

$$SS_{xx} = 340.700; \quad s_e = 1.951; \quad \text{and } \hat{y} = 12.45 + 6.32 \cdot x$$

- Make a 99% confidence interval for β .
- Using a significance level of 0.025, test the null hypothesis that β is zero against the alternative that β is positive.
- Using a significance level of 0.01, can you conclude that β is zero against the alternative that it is different from zero?
- Using a significance level of 0.02, test whether β is different from 4.50.

2. The following information is obtained for a sample of 100 observations taken from a population. (Note that because $n > 30$, we can use the normal distribution to make a confidence interval and test a hypothesis about β)

$$SS_{xx} = 524.884; \quad s_e = 1.464; \quad \text{and } \hat{y} = 5.48 + 2.50 \cdot x$$

- Make a 98% confidence interval for β
- Test at the 2% significance level whether β is zero against the alternative that it is positive.
- Can you conclude that β is zero? Use $\alpha = 0.01$.
- Using a significance level of 0.01, test whether β is 1.75 against the alternative that it is greater than 1.75.

3. Refer to exercise 7 of previous chapter. The following table which gives the ages (in years) and prices (in hundred of dollars) of eight cars of specific model, is reproduced from that exercise.

Age	8	3	7	10	3	5	6	9
Price	16	74	38	21	98	56	49	30

- Construct a 95% confidence interval for β .
- Test at the 5% significance level if β is zero against the alternative that it is negative.

4. The following table gives the experience (in years) and monthly salaries (in thousands of tenge) of nine randomly selected secretaries

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	22	12	15	17	15	19	24	13	27

- a) Find the least squares regression line with experience as an independent and monthly salary as dependent variables.
- b) Construct a 95% confidence interval for β .
- c) Test at the 2.5% significance level if β is zero against the alternative that it is positive.

5. The data on the size of six offices (in hundreds of square meters) and the monthly rents (in dollars) paid by firms for those offices are reproduced below from exercise 6 of the previous section.

Size of offices	22	17	19	28	35	24
Monthly rent	710	590	730	880	1080	820

- a) Construct a 99% confidence interval for β . You can use the calculations made in exercise 6 of previous section here.
- b) Test at the 5% significance level the null hypothesis that β is zero against the alternative that it is different from zero.

6. The following data give information on the ages (in years) and the number of breakdowns during the past year for a sample of six machines at a large company.

Age	9	14	18	15	10	11
Number of breakdowns	34	46	52	64	42	44

- a) Find the least squares regression line $\hat{y} = a + b \cdot x$
- b) Give a brief interpretation of the values a and b .
- c) Compute and interpret R^2 .
- d) Compute the standard deviation of errors.
- e) Construct a 98% confidence interval for β .

f) Test at the 2.5% significance level the null hypothesis that β is zero against the alternative that it is positive.

7. The following table gives information on the temperature in a city and volume of the ice cream (in thousands) sold at the supermarket for a random sample of eight days during the summer.

Temperature	22	16	28	12	30	19	24	32
Ice cream sold	3.64	3.12	4.08	2.84	3.98	3.55	4.02	4.38

- a) Find the least squares regression line $\hat{y} = a + b \cdot x$. Take temperature as an independent variable and volume of ice cream sold as a dependent variable.
- b) Give a brief interpretation of the values a and b .
- c) Compute and interpret R^2 .
- d) Compute the standard deviation of errors.
- e) Construct a 95% confidence interval for β .
- f) Test at the 1% significance level the null hypothesis that β is zero against the alternative that it is positive.

Answers

- 1.** a) 6.01 to 6.63; b) $T.S. = t = 59.792$; reject H_0 ; c) $T.S. = t = 59.792$; reject H_0
 d) $T.S. = t = 17.219$; reject H_0 ; **2.** a) 2.35 to 2.65; b) $T.S. = z = 39.12$; reject H_0 ;
 c) $T.S. = z = 39.12$; reject H_0 ; d) $T.S. = z = 11.74$; reject H_0 ; **3.** a) -15.53 to -7.17; b) $T.S. = t = -6.645$; reject H_0 ; **4.** a) $\hat{y} = 10.4986 + 0.8689 \cdot x$; b) 0.5559 to 1.1819; c) $T.S. = t = 8.323$; reject H_0 ; **5.** a) 12.34 to 37.92; b) $T.S. = t = 4.604$; reject H_0 **6.** a) $\hat{y} = -1.4337 + 0.8916 \cdot x$; c) $R^2 = 0.94$; d) $s_e = 0.9285$; e) 0.5708 to 1.2124; f) $T.S. = t = 9.356$; reject H_0 ;
7. a) $\hat{y} = 2.0680 + 0.0714 \cdot x$; c) $R^2 = 0.92$; d) $s_e = 0.1537$; e) 0.0511 to 0.0917; f) $T.S. = t = 8.602$; reject H_0 .

3.8. Using the regression model for prediction a particular value of y

The second major use of a regression model is to predict a particular value of y for a given value of x , say x_0 . For example, we may want to predict the food expenditure of a randomly selected household with a monthly income of \$3000. In this case, we are not interested in the mean food expenditure of all households with a monthly income of \$3000 but in the food expenditure of one particular household with a monthly income of \$3000. This predicted value of y is denoted by \hat{y}_p . To predict a single value of y for $x = x_0$ from estimated sample regression line, we use the value of \hat{y} as *a point estimate of y_p* . Using the estimated regression line, we find \hat{y} for $x = 30$ as

$$\hat{y} = 1.1414 + 0.2642 \cdot (30) = 9.0674$$

Thus, based on our regression line, the point estimate for the food expenditure of a given household with a monthly income of \$3000 is \$906.74 per month. Different regression lines estimated by using different samples of seven households each taken from the same population will give different values of the point estimator for the predicted value of y for $x = 30$. Hence, a confidence interval constructed for \hat{y}_p based on one sample will give a more reliable estimate of \hat{y}_p than will a point estimate. The confidence interval constructed for \hat{y}_p is more commonly called a prediction interval.

Suppose that the population regression model is

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad (i = 1, 2, \dots, n+1)$$

and that the standard regression assumptions hold, and that the ε_i are normally distributed. Let a and b be the least squares estimates of α and β .

It can be shown that the following are $100(1 - \alpha)\%$ intervals:

1. For the forecast of the single value resulting for y_{n+1} at a given x_{n+1} , the prediction interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \cdot s_e^2}$$

2. For the forecast of the conditional expectation, $E(y_{n+1} | x_{n+1})$, the confidence interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \cdot s_e^2}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n+1} x_i}{n} \quad \text{and} \quad \hat{y}_{n+1} = a + b \cdot x_{n+1}.$$

Example:

- For the data on incomes and food expenditures of seven households, find
- 99% prediction interval for the predicted food expenditure for a single household with a monthly income of \$3500;
 - Obtain a 99% confidence interval for the expected food expenditure for all households with a monthly income of \$3000.

Solution:

- a) The point estimate of the predicted food expenditure for $x = 35$ is given by

$$\hat{y} = 1.1414 + 0.2642 \cdot (35) = 10.3884$$

$$100(1 - \alpha)\% = 99\%$$

$$\alpha = 0.01$$

$$\alpha/2 = 0.005$$

$$t_{n-2,\alpha/2} = t_{5,0.005} = 4.032$$

Using data from the previous chapters

$$s_e = 0.9922; \quad \bar{x} = 30.2857; \quad \text{and} \quad SS_{xx} = 801.4286$$

Hence, the 99% prediction interval for y_p for $x=35$ is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \cdot s_e}$$

$$10.3884 \pm 4.032 \cdot \sqrt{\left[1 + \frac{1}{7} + \frac{(35 - 30.2857)^2}{801.4286} \right]} \cdot 0.9922 =$$

$$= 10.3884 \pm 4.3284 = 6.0600 \text{ to } 14.7168$$

Thus, with 99% confidence we can state that the predicted food expenditure of a household with a monthly income of \$3500 is between \$606.00 and \$1471.68.

b) Once again, the point estimate of the expected food expenditure for $x=35$ is

$$\hat{y} = 1.1414 + 0.2642 \cdot (35) = 10.3884$$

Hence, the 99% confidence interval for $E(y_{n+1}/35)$ is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \cdot \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \cdot s_e}$$

$$10.3884 \pm 4.032 \cdot \sqrt{\left[\frac{1}{7} + \frac{(35 - 30.2857)^2}{801.4286} \right]} \cdot 0.9922 =$$

$$= 10.3884 \pm 1.6523 = 8.7361 \text{ to } 12.0407$$

Thus, with 99% confidence we can state that the mean food expenditure for all households with monthly income of \$3500 is between \$873.61 and \$1204.07.

As we can observe, the interval in part a) [606.00 to 1471.68] is much wider than the one for the mean value of y for $x=35$ calculated in part b) [873.61 to 1204.04]. This is always true. The prediction interval for predicting a single value of y is always larger than the confidence interval for estimating the mean value of y for a certain value of x .

Exercises

1. Construct a 99% confidence interval for the mean value of y and a 99% prediction interval for the predicted value of y for the following

$$a) \hat{y} = 3.25 + 80 \cdot x \text{ for } x = 15 \text{ given } s_e = 0.954;$$

$$\bar{x} = 18.52; SS_{xx} = 144.65; \text{ and } n = 10$$

$$b) \hat{y} = -27 + 7.67 \cdot x \text{ for } x = 12 \text{ given } s_e = 2.46; \bar{x} = 13.43;$$

$$SS_{xx} = 369.77; \text{ and } n = 10$$

2. Refer to Exercise 4 of the previous section. Construct a 90% confidence interval for the mean monthly salary of secretaries with 10 years of experience. Construct a 90% prediction interval for the monthly salary of a randomly selected secretary with 10 years of experience.

3. Refer to Exercise 6 of the previous section. Construct a 95% confidence interval for the mean number of breakdowns for all cars which are 16 years old. Determine a 95% prediction interval for y_p for $x = 16$.

4. The following data give information on the lowest cost price (in dollars) and the average attendance (thousand) for the past year for eight football teams

Ticket price	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
Attendance	24	21	22	22	18	13	9	6

- a) Taking ticket price as an independent variable and attendance as a dependent variable, estimate the regression of attendance on the ticket price.
- b) Interpret the slope of the estimated regression line.
- c) Find and interpret the coefficient of determination.
- d) Find and interpret a 90% confidence interval for the slope of the population regression line.

e) Find a 90% confidence interval for expected number of attendance for which the price of ticket is 20.

5. A sample of 25 employees at a production plant was taken. Each employee was asked to assess his or her own job satisfaction (x), on scale from 1 to 10. In addition, the number of days absent (y) from work during the last year were found for these employees. The sample regression line

$$\hat{y} = 13.6 - 1.2 \cdot x$$

was estimated by least squares for these data. Also found that

$$\bar{x} = 6.0; \quad \sum_{i=1}^{25} (x_i - \bar{x})^2 = 130; \quad SSE = 80.6$$

- Test at the 1% significance level against the appropriate one sided alternative the null hypothesis that job satisfaction has no linear effect on absence.
- A particular employee has job satisfaction level 4. Find a 90% confidence interval for the number of days this employee would be absent from work in a year.

Answers

- 1.** a) 13.871 to 16.629; b) 11.765 to 18.735; **2.** 18.108 to 20.267; 15.838 to 22.537; **3.** **4.** a) $\hat{y} = 2.029 + 0.0464 \cdot x$; c) $R^2 = 0.4194$; d) $0.003 < \beta < 0.0928$; e) 2.6525 to 3.2615; **5.** a) $T.S. = t = -7.303$; reject H_0 ; b) 5.4798 to 12.1202.