CHAPTER 1

Numerical Descriptive Measures

This chapter introduces the areas of statistics, basic concepts and definitions, population and sample, types of variables, graphical representations of qualitative and quantitative data, numerical summary measures, measure of central tendency for ungrouped data (mean, median and mode), measures of dispersion for ungrouped data (range, variance and standard deviation), mean, variance and standard deviation for grouped data, measures of position including quartiles, percentiles, box-and-whisker plot, skewness and kurtosis.

1.1 INTRODUCTION

Statistics refers to the field or discipline of study. *Statistics* is a group of methods that are used to collect, analyse, present, and interpret data and to make decisions. Decisions made by using statistical methods are called *educated guesses. Statistics* has two areas, *theoretical* or *mathematical statistics* and *applied statistics* as shown in Fig.1.1. *Theoretical* or *mathematical statistics* deals with the development, derivation, and proof of statistical theorems, formula, rules, and laws. *Applied statistics* concerns the applications of those theorems, formulae, rules and laws to solve real-world problems. This chapter deals with *applied statistics.*

Applied statistics is divided into two branches, *Descriptive statistics* and *Inferential statistics*. *Descriptive statistics* involves methods for organising, displaying, and describing data by using tables, graphs and summary measures. The collection of all elements of interest is called a *population* in statistics. The selection of a few elements from this population is called a *sample*. Statistics that deals with making-decision, inferences, predictions, and forecasts about populations based on results obtained from samples. Statistics that deals with decision-making procedures is called *inferential statistics*. This branch of statistics is also known as *inductive reasoning* or *inductive statistics*. *Inferential statistics* consists of methods that employ sample results to help make decisions or predictions about a population.

Probability, which measures the likelihood that an event will occur, is an important part of statistics. It is the basis of inferential statistics, where decisions are made under conditions of uncertainty. Probability theory is used to evaluate the uncertainty involved in those decisions. Probability statements are about occurrence or non-occurrence of a certain event under certain conditions.

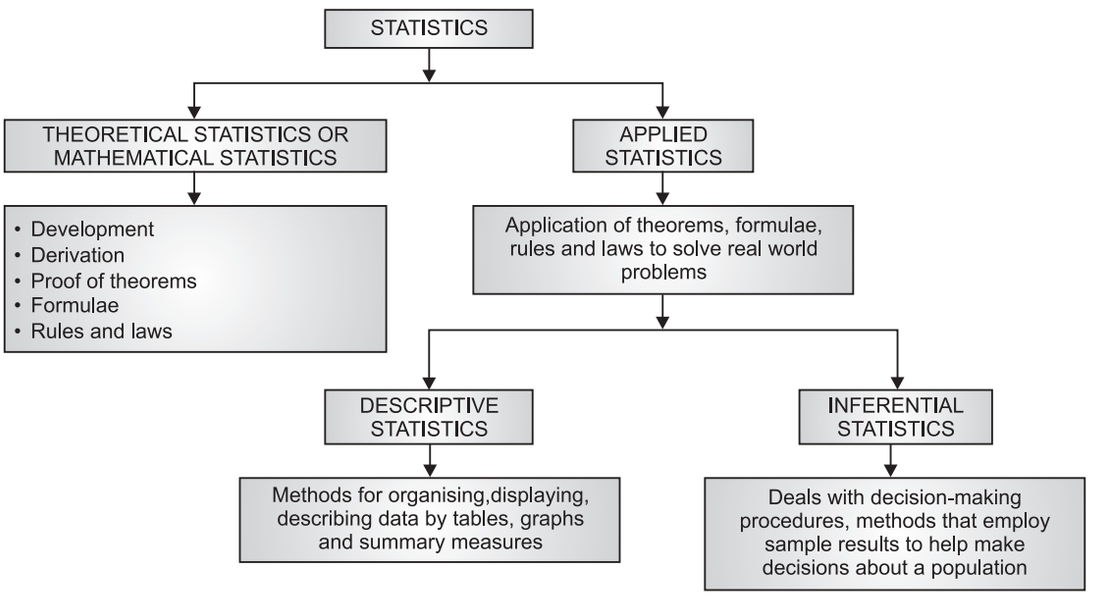


Fig. 1.1: Areas of statistics

1.1.1 Population and Sample

A *population* consists of all elements: individuals, items, or objects, whose characteristics are being studied. The population that is being studied is also called the *target population*. The collection of a few elements selected from a population is called a sample. Thus, a portion of the population selected for study is referred to as a *sample*. The collection of information from the elements of a population or a sample is called a survey. A *survey* that includes each and every element of the population is called a *census*. A survey conducted on a sample is called a *sample survey*. A sample that represents the characteristics of the population as closely as possible is called a *representative sample*.

A sample drawn in such a way that each element of the population has same chance of being selected is called a *random sample*. If the chance of being selected is the same for each element of the population, it is called a *simple random sample*. An *element* or *member* of a sample or population is a specific subject or object about which the information is collected. A *variable* is a characteristic under study that assumes different values for different elements. In relation to a variable, the value of a *constant* is fixed. The value of a variable for an element is called an *observation* or *measurement*. A *data set* is a collection of observations on one or more variables.

1.1.2 Types of Variables

A variable that can be measured numerically is called a *quantitative variable*. The data collected on a quantitative variable are known as *quantitative data*. Quantitative variable is classified as either *discrete variable* or *continuous variable*. A variable whose values are continuous is called a *continuous variable*. A *discrete variable* can assume only certain values with no intermediate values. On the other hand, a variable that can assume any numerical value over a certain interval or intervals is called a *continuous variable*. The time taken to complete a trip to a corner store is an example of a continuous random variable because it

can assume any value, say, between one and two hours. A variable that cannot assume a numerical value but can be classified into two or more non-numeric categories is called a *qualitative* or *categorical* variable. The data collected on such a variable are known as *qualitative data*. The types of variables and types of data are classified as shown in Figs.1.2 *(a)* and *(b)*.

Quantitative

Continuous: Weight, Height, Line width measurements

Discrete: Number of students, Number of accidents

Qualitative

Nominal: Black, Green, Yellow, …

Ordinal: High, Medium, Low

Ranked: 1 2 3 etc.

Data collected on different elements at the same point in time or for the same period of time are called *cross-section data.*

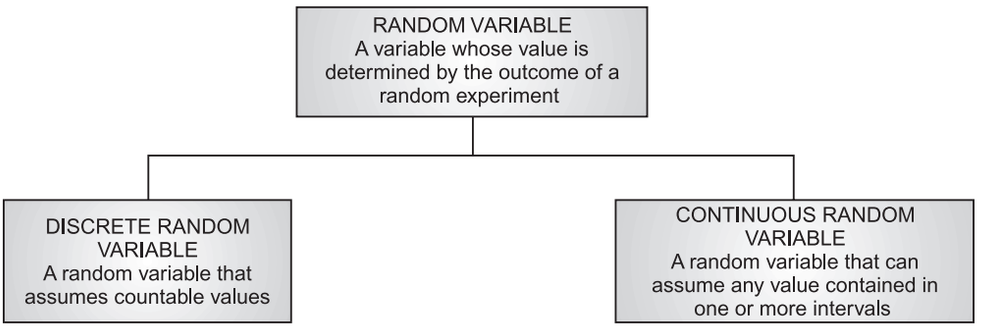


Fig. 1.2(a): Types of variables

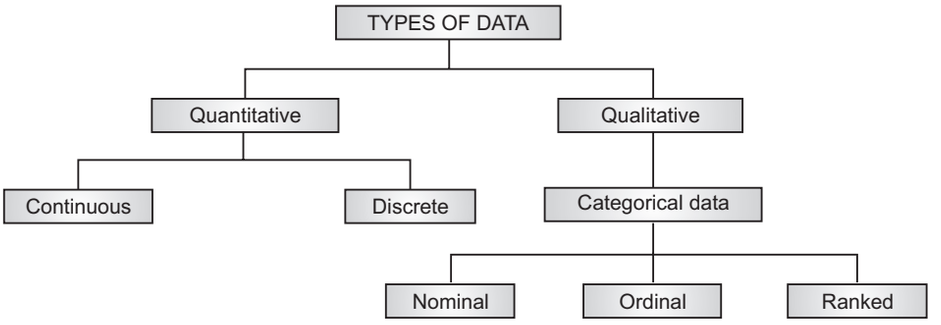


Fig. 1.2(b): Types of data

1.1.3 Organising Data

Data recorded in the sequence in which they are collected and before they are processed or ranked are called *raw data*.

1.1.3.1 Qualitative Data

A frequency *distribution* for qualitative data lists all categories and the number of elements that belong to each of the categories. A *relative frequency distribution* lists the relative frequencies for all categories.

$$\text{Relative frequency of a category} = \frac{\text{frequency of that category}}{\text{sum of all frequencies}} \tag{1.1}$$

A *percentage distribution* lists the percentages for all categories.

$$\text{Percentage} = \text{relative frequency} \times 100 \tag{1.2}$$

1.1.3.2 Graphical Representation of Qualitative Data

The *bar graph* and the *pie chart* are two types of graphs used to represent or display qualitative data. A graph made of bars whose heights represent the frequencies of representative categories is called a *bar graph*. A *pie chart* is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The pie chart represents the total sample or population. The pie chart is divided into different portions that represent the percentages of the population or sample belonging to different categories.

Example E 1.1

The following data give the results of a sample survey of three categories *A*, *B* and *C*.

A B C C A B A B C C A B C A C

A C A B C A B C C C C A C C C

(a) prepare a frequency distribution table

(b) compute the relative frequencies and percentages for all three categories *A*, *B* and *C*

(c) what percentage of the elements in this sample belongs to category *B*?

(d) what percentage of the elements in this sample belongs to categories *A* or *C*?

(e) draw a bar graph of the frequency distribution.

SOLUTION:

(a) and *(b)*

Category	Frequency	Relative frequency	Percentage
A	9	0.30	30
B	6	0.20	20
C	15	0.50	50
Totals	30	1.00	100

(c) 20% of the element in this sample belongs to category *B*.

(d) 30% + 50% = 80% of the elements in this sample belong to categories *A* or *C*.

(e) The bar graph of the frequency distribution is shown in Fig. E1.1.

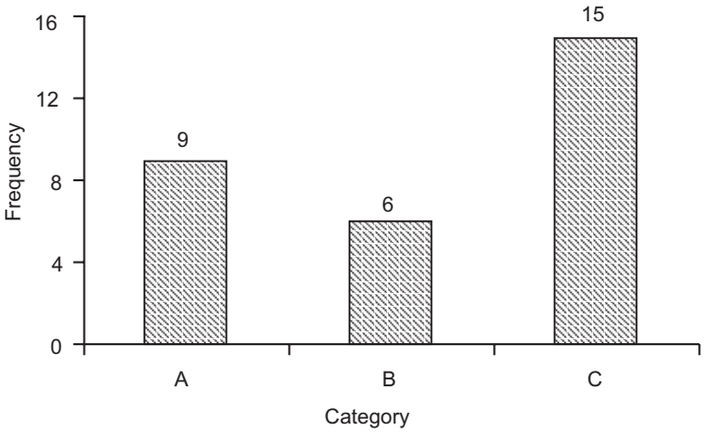


Fig. E1.1: Bar graph of the frequency distribution

Example E1.2

Draw the pie chart for the percentage distribution in Example E1.1.

SOLUTION:

The pie chart is shown in Fig. E1.2.

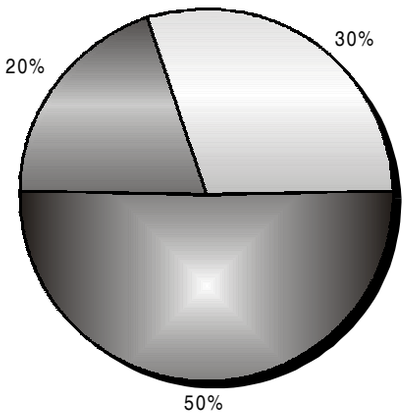


Fig. E1.2: Pie chart

1.1.3.3 Graphical Representation of Quantitative Data

A *frequency distribution* for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called *grouped data*. The *class boundary* is the midpoint of the upper limit of one class and the lower limit of the next class, while class width = upper boundary – lower boundary.

The *class width* is also known as the *class size*.

$$\text{The } \textit{class} \text{ midpoint or } \textit{mark} = \frac{\text{lower limit } + \text{ upper limit}}{2} \tag{1.3}$$

$$\text{Approximate class width} = \frac{\text{largest value} - \text{smallest value}}{\text{number of classes}} \tag{1.4}$$

The relative frequencies and percentages for a quantitative data set are calculated as follows:

$$\text{Relative frequency of a class} = \frac{\text{frequency of that class}}{\text{sum of all frequencies}} = \frac{f}{\sum f} \tag{1.5}$$

$$\text{Percentage} = \text{relative frequency} \times 100$$

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

The most common shapes of histograms are *symmetric, skewed,* and *uniform* or *rectangular*. A *symmetric histogram* is identical on both sides of its central point. A *skewed histogram* is nonsymmetric. For a skewed histogram, the tail on one side is longer than the tail on the other side. A *skewed-to-the-right* histogram has a longer tail on the right side while a skewed-to-the-left histogram has a longer tail on the left side. A *uniform* or *rectangular histogram* has the same frequency for each class.

A *polygon* is another method that can be used to represent quantitative data in graphical form. A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a *polygon*. A polygon with relative frequencies marked on the vertical axis is called a *relative frequency polygon*. In a similar manner, a polygon with percentages marked on the vertical axis is called a *percentage polygon*.

A *cumulative frequency distribution* gives the total number of values that fall below the upper boundary of each class.

$$\text{Cumulative relative frequency} = \frac{\text{cumulative frequency}}{\text{total observations in the data set}} \tag{1.6}$$

$$\text{Cumulative frequency} = \text{cumulative relative frequency} \times 100 \tag{1.7}$$

When plotted on a graph, the cumulative frequencies give a curve that is called an *ogive*. Thus, an *ogive* is a curve drawn for the cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at height equal to the cumulative frequencies of respective classes.

Another technique that is used to represent quantitative data in condensed form is the *stem-and-leaf-display*. In a *stem-and-leaf display* of quantitative data, each value is divided into two portions: a stem and a leaf. The leaves for each stem are shown separately in a display.

Example E1.3

Out of 27 randomly selected families in a small town, the following data give the number of children less than 18 years of age:

1	0	2	4	4	2	3	2	0
2	0	0	0	0	1	4	0	2
0	2	0	2	3	2	1	2	2

(a) construct a frequency distribution table using single-valued classes

(b) compute the relative frequencies and percentages for all classes

(c) find the number of families with one or two children less than 18 years of age

(d) find the number of families with two or three children less than 18 years of age

(e) draw a bar graph for the frequency distribution.

SOLUTION:

(a) and *(b)* The frequency distribution, relative frequencies and percentages are shown in Table E1.3.

Table E1.3

Number of children less than 18 years of age	Frequency	Relative frequency	Cumulative relative frequency	Percentage (probability)
0	9	0.334	0.334	33.3
1	3	0.111	0.445	11.1
2	10	0.370	0.815	37.0
3	2	0.074	0.889	7.4
4	3	0.111	1.0	11.1
Total	27			100.0

(c) Number of families with one or two children under 18 years of age = 3 + 10 = 13.

(d) Number of families with two or three children under 18 years of age = 10 + 2 = 12.

(e) The bar graph for the frequency distribution is shown in Fig. E1.3.

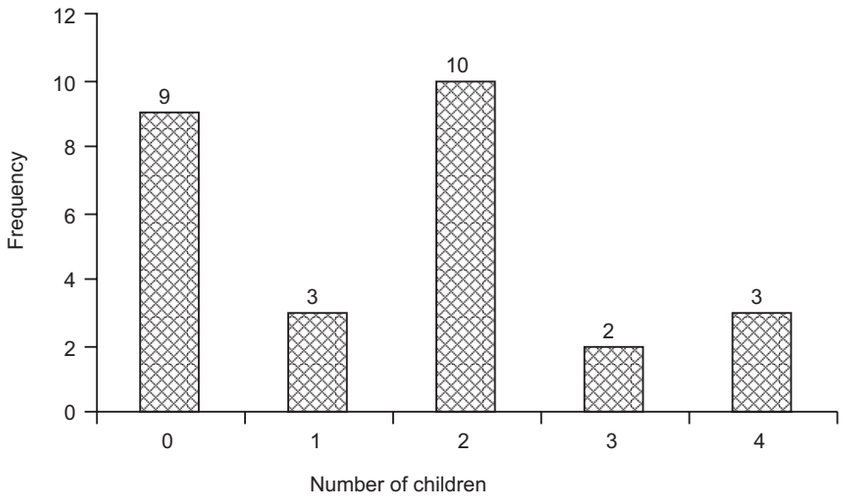


Fig. E1.3: Bar graph of frequency distribution

Example E1.4

The frequency distribution of ages of all 60 part-time graduate students in a Master of Science in Mechanical Engineering degree program of a university is given below:

Age (years)	Number of students
21-25	22
26-30	13
31-35	11
36-40	8
41-45	6

(a) prepare a cumulative frequency distribution table

(b) calculate the cumulative relative frequencies and cumulative percentages for all classes

(c) find the percentage of students who are 31 years of age or older

(d) draw a ogive for the cumulative percentage distribution

(e) using the ogive, determine the percentage of students who are 35 years or younger.

SOLUTION:

(a) and *(b)* The cumulative frequency, cumulative relative frequency and percentages are shown in Table E1.4.

Table E1.4

Age (years)	Frequency	Cumulative frequency	Cumulative relative frequency	Cumulative percentage
21–25	22	22	0.366	36.6
26–30	13	35	0.583	58.3
31–35	11	46	0.766	76.6
36–40	8	54	0.90	90.0
41–45	6	60	1.00	100.0

(c) Percentage of students who are 31 years of age or older = 100 – 58.3 = 41.7%.

(d) The frequency, cumulative relative frequency and the ogive diagram for the cumulative percentage distribution is shown in Fig. E1.4.

(e) The percentage of a student who is 35 years or younger is about 70%.

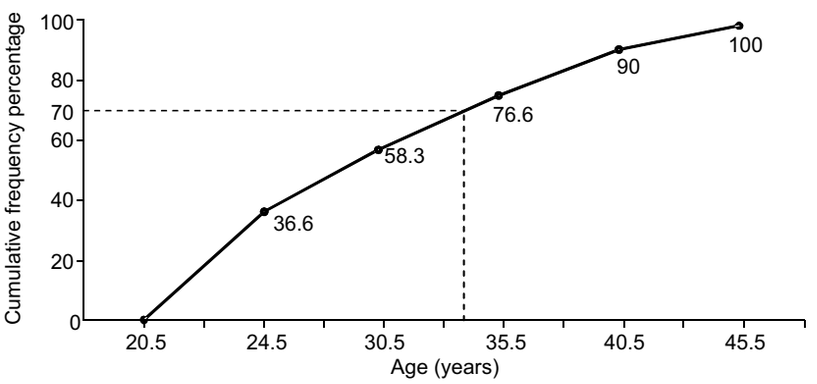


Fig. E1.4: Ogive diagram for the cumulative percentage distribution

Example E1.5

Following are the overtime monies earned by 15 employers of a company during the month of August 2008.

780, 892, 995, 1168, 883, 1041, 994, 1084, 776, 981, 1129, 1067, 932, 655, 642

Prepare a stem-and-leaf display by arranging the leaves for each stem in an increasing order.

SOLUTION:

The stem-and-leaf display and the stem-and-leaf display data are shown in Figs. E1.5(*a*) and E1.5(*b*) respectively.

6	55	42			
7	80	76			
8	92	83			
9	95	94	81	32	
10	41	84	67		
11	68	29			

Fig. E1.5(a): Stem-and-leaf display of data

6	42	55			
7	76	80			
8	83	92			
9	32	81	94	95	
10	41	67	84		
11	29	68			

Fig. E1.5(b): Stem-and-leaf display of data arranged in increasing order

1.2 NUMERICAL SUMMARY MEASURES

There are three basic numerical summary (descriptive) measures used to organize and display large data sets as shown in Fig.1.3. They are:

1. Measures of central tendency
2. Measures of dispersion or scatter and
3. Measures of position.

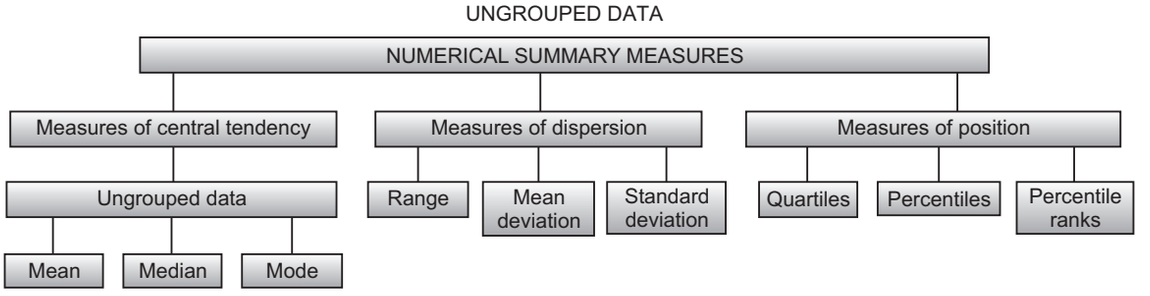


Fig. 1.3: Three basic numerical summary measures

1.2.1 Measures of Central Tendency for Ungrouped Data

A data set is generally represented by numerical summary measures called the *typical values*. A measure of central tendency gives the center of a histogram or a frequency distribution curve. There are three measures of central tendency: the mean, the median, and the mode.

1.2.1.1 Mean for Ungrouped Data

An *average* is a value which is typical or representative of a set of data. Since such values tend to lie centrally within a set of data arranged according to magnitude, averages are also called *measures of central tendency.*

The mean, also called the arithmetic mean or average, for ungrouped data is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data } \mu = \frac{\sum X}{N} \tag{1.8}$$

$$\text{Mean for sample data } \overline{X} = \frac{\sum X}{n} \tag{1.9}$$

where μ is the population mean, $\sum X$ the sum of all values, N the population size, $\overline{X}$ the sample mean, and n the sample size. Sometimes the data may contain a few very small or a few very large values. Such values are called *outliers* or *extreme values*.

1.2.1.2 Median

The *median* is the value of the middle term in a data set that has been ranked in either increasing or decreasing order. Median for ungrouped data = the value of the $\left(\frac{n+1}{2}\right)^{th}$ term in a ranked data set. If the given data set represents a population, replace n by N. If the number of observations in a data set is odd, then the median is given by the value of the middle term in the ranked data. Similarly, if the number of observations is even, then the median is the average of the values of the two middle terms.

1.2.1.3 Mode

The *mode* for an ungrouped data is the value that occurs with the highest frequency in a data set. A data set with only one value occurring with highest frequency has only one mode and such a data set is called *unimodal.* A data set with two values occurring with the same (highest) frequency has two modes. The distribution in this case is called *bimodal.* If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be *multi modal.* It should be noted here that a data set with each value occurring only once has no mode.

1.2.1.4 Empirical Relation among Mean, Median and Mode

For unimodal frequency curves which are moderately skewed (asymmetrical), the empirical relationship is given by

$$\text{mean} - \text{mode} = 3\,(\text{mean} - \text{median})$$

Figures 1.4*(a)* and *(b)* show the relative positions of the mean, median and mode for frequency curves which are skewed to the right and left respectively. It should be noted here that for symmetrical curves, the mean, median and mode and they all coincide.

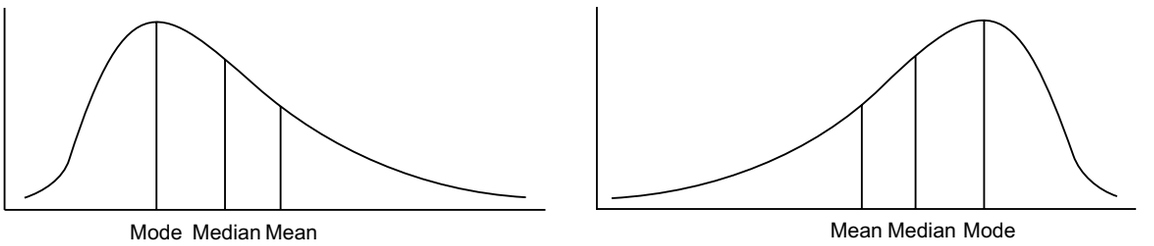


Fig. 1.4: Empirical relationship

Example E1.6

The following data gives the number of automobiles owned by each of the 30 families in a city.

2	3	1	1	1	2	4	3	1	1
3	2	2	1	2	3	2	3	1	3
4	1	3	4	1	1	1	3	2	2

(a) prepare a frequency distribution table for this data set using single-valued classes.

(b) calculate the relative frequency and percentage distributions.

(c) find the percentage of the families in this sample having two or three automobiles.

(d) draw a bar graph for the relative frequency distribution.

SOLUTION:

(a) and *(b)*

Number of automobiles in a family	Frequency	Relative frequency	Percentage
1	11	0.366	36.6
2	8	0.267	26.7
3	8	0.267	26.7
4	3	0.1	10

(c) Number of families in this sample who have two or three automobiles = 26.7 + 26.7 = 53.4%

(d) The bar graph for the relative frequency distribution is shown in Fig. E1.6.

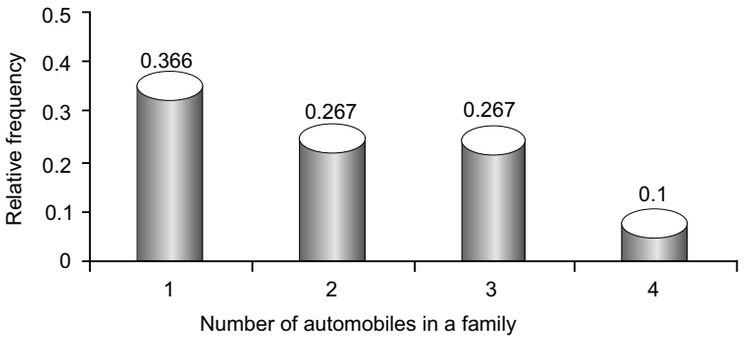


Fig. E1.6: Bar graph of relative frequency distribution

Example E1.7

The following are the ages of 15 employees of a manufacturing company.

65, 38, 52, 27, 24, 45, 49, 50, 26, 37, 44, 52, 41, 60, 39

Calculate the mean, median and mode.

SOLUTION:

$$\text{Mean} = \frac{\Sigma X}{n} = \frac{(65+38+52+27+24+45+49+50+26+37+44+52+41+60+39)}{15}$$

$$= \frac{649}{15} = 43.266 \text{ years.}$$

Rearrange the given data in an increasing order gives

24, 26, 27, 37, 38, 39, 41, 44, 45, 49, 50, 52, 52, 60, 65

$$\text{Position of the middle term} = \frac{n+1}{2} = \frac{15+1}{2} = 8$$

Hence, the median is the value of the middle term in the ranked data.

Median = 44 years

In the data set, 52 occur twice and each of the remaining values occurs only once.

Therefore, mode = 52.

1.2.2 Measures of Dispersion for Ungrouped Data

The mean, median and mode are not sufficient measures to reveal the shape of the distribution of a data set. The measures that show the spread of a data set are called the *measures of dispersion*. There are three measures of dispersion range, variance, and standard deviation as shown in Fig. 1.5.

1.2.2.1 Range

The range is the difference between the largest and smallest values in the data.

The range for ungrouped data = largest value – smallest value.

The range is not a very satisfactory measure of dispersion.

1.2.2.2 Variance and Standard Deviation

The standard deviation is the most used measure of dispersion. The formulae for calculating the variance and standard deviation are as follows:

Variance for ungrouped data

Population variance

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N} \tag{1.10}$$

Sample variance

$$s^2 = \frac{\sum\left(x - \overline{X}\right)^2}{n-1} = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} \tag{1.11}$$

where, σ^2 is the population variance, s^2 is the sample variance, $(x - \mu)$ or $(x - \bar{x})$ is the deviation of x value from the mean, μ is the population mean, $\bar{x}$ is the sample mean, N is the population size, and n is the sample size.

The population standard deviation $\sigma = \sqrt{\text{population variance}} = \sqrt{\sigma^2}$ (1.12)

The sample standard deviation $s = \sqrt{\text{sample variance}} = \sqrt{s^2}$ (1.13)

A numerical value such as the mean, median, mode, range, variance, or standard deviation computed for a population data set is called a *population parameter*, or simply a *parameter*. A summary measure calculated for a sample data set is called a *sample statistic* or simply a *statistic*.

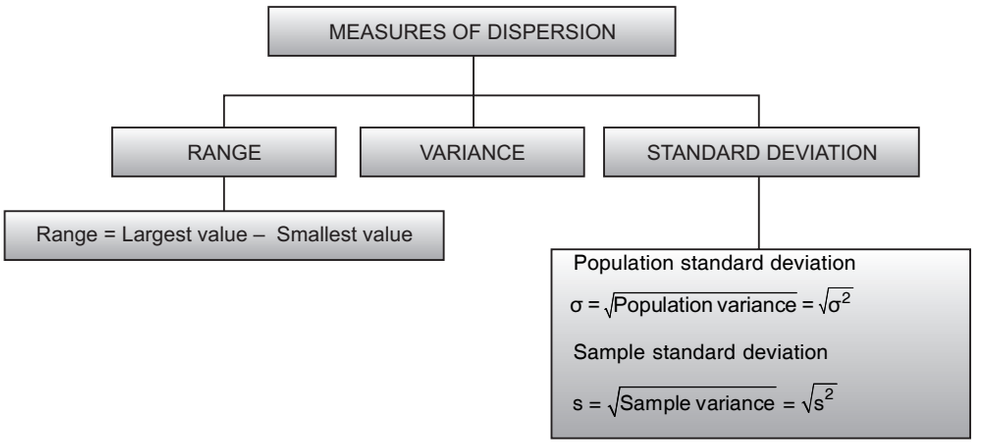


Fig. 1.5: Measures of dispersion

Example E1.8

The following data gives the number of years of service of 15 employees in a manufacturing company.

5, 9, 7, 6, 24, 11, 4, 13, 10, 9, 20, 8, 19, 17, 25

Calculate the range, variance and standard deviation.

SOLUTION:

Range = largest value – smallest value = 25 – 4 = 21 years

$$\text{Variance} = s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{2732 - \frac{178^2}{15}}{15-1} = \frac{2732 - 2112.266}{14} = 619.734$$

$$\text{Standard deviation} = s = \sqrt{s^2} = \sqrt{619.734} = 24.894$$

Example E1.9

The following data refers to the pay (in thousands of dollars) of 12 employees at a manufacturing company.

75 69 65 49 21 21 18 18 17 16 16 15

(a) calculate the mean, median and mode for this data

(b) find the range, variance and standard deviation.

SOLUTION:

(a) Mean $= \frac{\Sigma x}{n} = \frac{400}{12} = \$33{,}333$

Median is $\left[\frac{n+1}{2}\right]^{th}$ term in a ranked data set. Therefore,

$$\frac{n+1}{2} = \frac{12+1}{2} = 6.5$$

$$\text{Median} = \frac{21+18}{2} = \$19{,}500$$

Mode = \$21,000, \$18,000 and \$16,000

(b) Range = largest pay – smallest pay = \$75,000 – \$15,000 = \$60,000

$$\text{Variance} = s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{19{,}528 - \frac{400^2}{12}}{12-1} = 563.1515$$

Standard deviation $= s = \sqrt{s^2} = \sqrt{563.1515} = 23.73 = \$23{,}730$

1.2.3 Mean, Variance and Standard Deviation for Grouped Data

1.2.3.1 Mean for Grouped Data

$$\text{Mean for population data: } \mu = \frac{\sum mf}{N} \tag{1.14}$$

$$\text{Mean for sample data: } \overline{X} = \frac{\sum mf}{n} \tag{1.15}$$

where, m is the midpoint and f is the frequency of a class.

1.2.3.2 Variance and Standard Deviation for Grouped Data

Population variance

$$\sigma^2 = \frac{\sum f\ (m-\mu)^2}{N} = \frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N} \tag{1.16}$$

Sample variance

$$s^2 = \frac{\sum f\ (m - \overline{X})^2}{n-1} = \frac{\sum m^2 f - \frac{(\sum mf)^2}{n}}{n-1} \tag{1.17}$$

where σ^2 is the population variance, s^2 is the sample variance, m is the midpoint of a class, N is the population size, n is the sample size, μ is the population mean, $\overline{X}$ is the sample mean, and f the class frequency.

Population standard deviation $\sigma = \sqrt{\sigma^2}$ (1.18)

Sample standard deviation $s = \sqrt{s^2}$ (1.19)

Example E1.10

Calculate the mean, variance and standard derivation for the following population data.

x	0–3	4–7	8–11	12–15	16–19	20–23
f	7	4	18	12	7	6

SOLUTION:

Classes	f	m	mf	m^2 f
0-3	7	1.5	10.5	15.75
4-7	4	5.5	22	121
8-11	18	9.5	171	1624.5
12-15	12	13.5	162	2187
16-19	7	17.5	122.5	2143.75
20-23	6	21.5	129	2773.50
Total	54		617	8865.50

Population mean

$$\mu = \frac{\sum mf}{N} = \frac{617}{54} = 11.426$$

$$\text{Variance} = \sigma^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N} = \frac{8865.50 - \frac{617^2}{54}}{54} = \frac{8865.50 - 7049.796}{54} = 33.624$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{33.624} = 5.798$

1.2.4 Measures of Position

Here we examine several descriptive measures based on percentiles. Descriptive measures based on percentiles are not sensitive to the influence of a few extreme observations. Descriptive measures based on percentiles are generally preferred over those based on the mean and standard deviation.

A measure of position determines the position of a single value in relation to other values in a sample or a population data set. The measures of position considered here are the quartiles, percentiles, and percentile ranks as shown in Fig. 1.6.

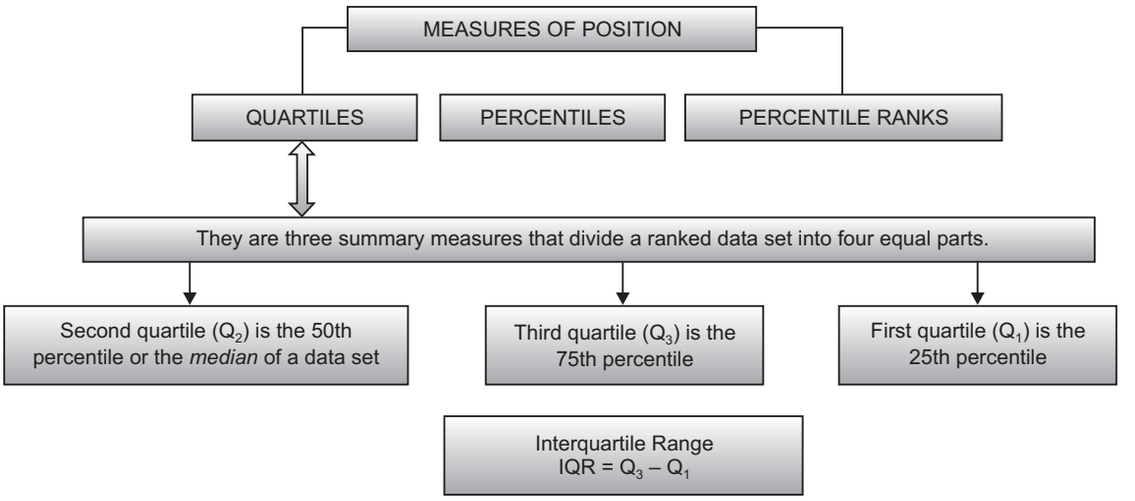


Fig. 1.6: Measures of position

1.2.4.1 Quartiles and Interquartile Range

The most commonly used percentiles are *quartiles*. Let n denotes the number of observations and if we arrange the data in an increasing order, then

- the first quartile is at position $\frac{n+1}{4}$
- the second quartile is the median, which is at position $\frac{n+1}{2}$
- the third quartile is at position $3\left(\frac{n+1}{4}\right)$

Thus, a data set has three quartiles, and denoted by Q_1, Q_2, and Q_3, Q_1, is the number that divides the bottom 25% of the data from the top 75%; the second quartile, Q_2 , the median, is the number that divides the bottom 50% of the data from the top 50%; the third quartile, Q_3, is the number that divides the bottom 75% of the data from the top 25%. Figure 1.7 shows the positions of the three-quartiles. Figure 1.8 shows the quartiles for uniform, bell-shaped, right-skewed, and left-skewed distributions. The difference between the third quartile and the first quartile of data set is called the *interquartile range* (IQR). Interquartile range is a preferred measure of variation when the median used as the measure of the center. The interquartile range is a resistant measure. IQR gives the range of the 50% of the observations.

Interquartile range IQR = $Q_3 - Q_1$

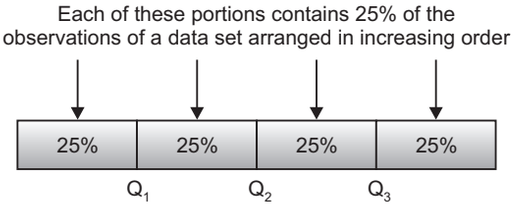


Fig. 1.7: Quartiles

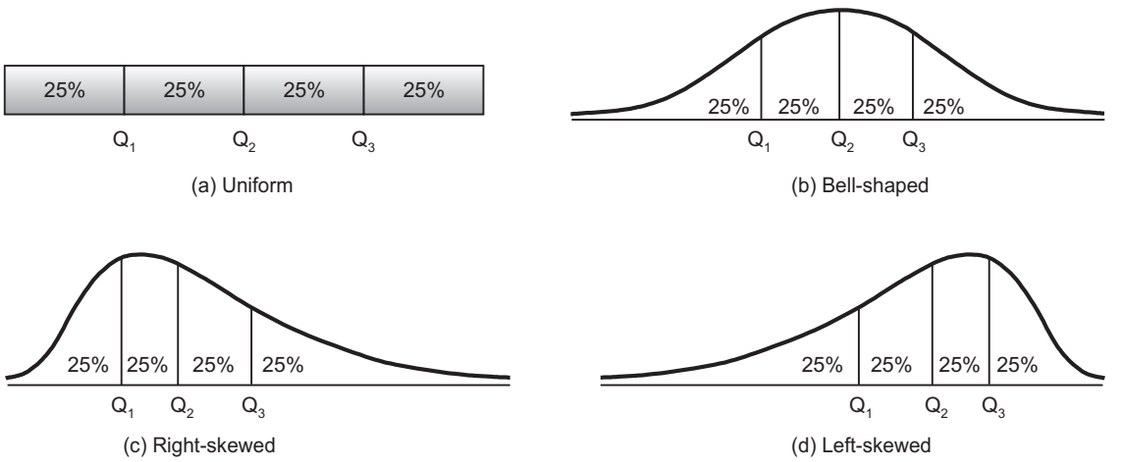


Fig. 1.8: Quartiles for (a) uniform (b) bell-shaped (c) right-skewed and (d) left-skewed distributions

1.2.4.2 Percentiles

Percentiles are the summary measures that divide ranked data set into 100 equal parts. Each ranked data set has 99 percentiles that divide it into 100 equal parts. The k^{th} percentile is denoted by P_k, where k is an integer in the range 1 to 99. Figure 1.9 shows the positions of the 99 percentiles.

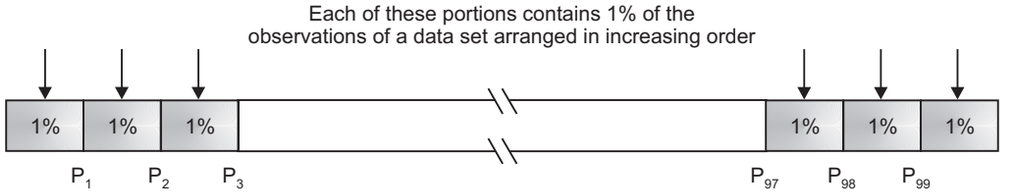


Fig. 1.9: Percentiles

Hence, the k^{th} percentile, P_k , can be described as a value in a data set such that about $k\%$ of the measurements are smaller than the value of P_k and about $(100 - k)\%$ of the measurements are greater than the value of P_k. The value of the k^{th} percentile is given by

$$P_k = \text{the value of the } \left(\frac{kn}{100}\right)^{th} \text{ term in a ranked data set}$$

where k denotes the number of the percentile and n represents the sample size.

The percentile rank for a particular value x_i of a data set is given by

$$\text{Percentile rank of } x_i = \frac{\text{number of values less than } x_i}{\text{total number of values in the data set}} \times 100$$

The percentile rank of x_i gives the percentage of values in the data set that are smaller than x_i.

Deciles: Deciles are also useful in statistics. The deciles of a data set divide into tenths, or 10 equal parts. A typical data set has nine deciles. If we denote them by D_1, D_2, … D_9, then the first decile, D_1 is the number that divide the bottom 10% of the data from the top 90%; the second decile, D_2, is the number that divide the bottom 20% of the data from the top 80%; and so on. Hence, the first decile is the 10th percentile, the second decile is the 20th percentile, etc.

1.2.4.3 Skewness and Kurtosis

Skewness is the degree of asymmetry, or departure from symmetry, of a distribution. For skewed distribution, the mean tends to lie on the same side of the mode as the longer tail as shown earlier in Figs. 1.4(*a*) and (*b*).

Skewness is defined as

$$\text{Skewness} = \frac{\text{mean–mode}}{\text{standard deviation}} = \frac{\bar{X} - \text{mode}}{s} \tag{1.20}$$

or alternatively,

$$\text{Skewness} = \frac{3(\text{mean–median})}{\text{standard deviation}} = \frac{3(\bar{X} - \text{median})}{s} \tag{1.21}$$

The above two measures are called, respectively, *Pearson's first and second coefficient of skewness.*

Other definitions for skewness in terms of quartiles and percentiles are as follows:

$$\text{Quartile coefficient of skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \tag{1.22}$$

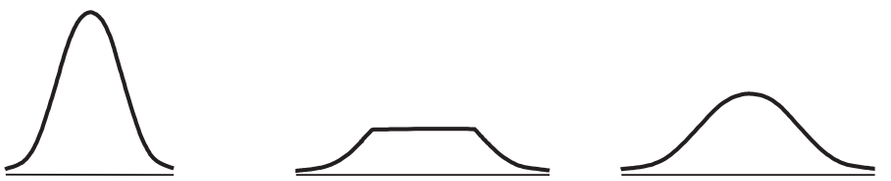
$$= \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{1.22a}$$

$$\text{10–90 percentile coefficient of skewness} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} \tag{1.23}$$

$$= \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \tag{1.23a}$$

If a frequency curve of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be *skewed to the right* or to have positive skewness. If the reverse is true, it is said to be *skewed to the left* or to have *negative skewness*.

Kurtosis is the degree of peakedness of a distribution taken relative to a normal distribution. A distribution having a relatively high peak such as the curve of Fig. 1.10(*a*) is called *leptokurtic*, while the curve of Fig. 1.10(*b*) which is flat-topped is called *platykurtic*. The normal distribution, Fig. 1.10(*c*), which is not very peaked or very flat-topped, is called *mesokurtic*.



(a) Leptokurtic (b) Platykurtic (c) Mesokurtic

Fig. 1.10: Kurtosis

Example E1.11

The number of hours worked by 24 employees of a company is given below:

40	43	40	39	36	44	40	39
39	52	27	50	41	47	40	48
38	36	25	41	35	36	16	40

(a) calculate the three quartiles and the Interquartile range

(b) find the approximate value of the 81st percentile

(c) calculate the percentile rank of 37.

SOLUTION:

(a) The ranked data is

16 25 27 35 36 36 36 38 39 39 39 40 40 40 40 40 41 41 43 44 47 48 50 52

$$Q_2 = \text{median} = 40$$

$$Q_1 = \frac{36+36}{2} = 36$$

$$Q_3 = \frac{41+43}{2} = 42$$

IQR = Interquartile range = $Q_3 - Q_1 = 42 - 36 = 6$

(b) $\frac{k\,n}{100} = \frac{81(24)}{100} = 19.44$

Thus, the 81st percentile can be approximated by the value of the 19th term in the ranked data, which is 43. Hence $P_{81} = 45$.

(c) Seven values in the given data are smaller than 37.

Therefore, the percentile rank of 37 = $\frac{7}{24} \times 100 = 29.17\%$.

1.2.4.4 Box-and-Whisker Plot

A *box-and-whisker plot* depicts a graphical representation of data using five measures: the median, the first quartile, the third quartile and the smallest and the largest values in the data set between the lower and the upper inner fences. A box-and-whisker plot shows the center, spread, and skewness of a data set. It is constructed simply by drawing a box and two whiskers that use the median, the first quartile, the third quartile and the smallest and the largest values in the data set between the lower and the upper inner fences.

The box plot utilizes three fundamental measures of dispersion in a graphic manner for a set of data. The simplest measure of dispersion is the range. The box plot as shown in Fig.1.11 also incorporates the interquartile range, $Q_3 - Q_1$, and the semi-interquartile range, $(Q_3 - Q_1)/2$. The semi-interquartile range also corresponds to the median of the data set. Outliers are observations that fall well outside the overall pattern of the data. An outlier may be the result of a measurement or recording error, an observation from a different

population, or an unusual extreme observations. We can utilize quartiles and the interquartile range (IQR) to identify potential outliers, that is, to spot observations that may be outliers. Referring to Fig.1.11 we define the lower and upper limits, the numbers that lie, respectively 1.5 IQRs below the first quartile and 1.5 IQRs above the third quartile.

Thus, Lower limit = $Q_1 - 1.5$ IQR

Upper limit = $Q_3 - 1.5$ IQR

Observations that lie outside the lower and upper limits are potential outliers. Typically outliers are observations that fall < ($Q_1 - 3$ IQR) and > ($Q_3 + 3$ IQR).

Box plots are useful for comparing two or more data sets and to identify the approximate shape of the distribution of a data set. Figure 1.12 shows the uniform, bell-shaped, right-skewed and left-skewed distributions and their corresponding box plots. The box width and whisker length relate to skewness and symmetry.

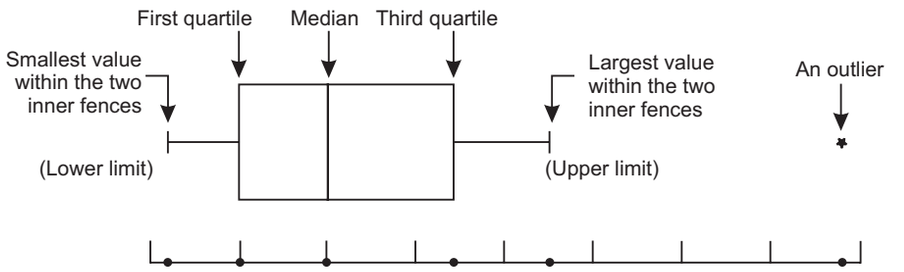


Fig. 1.11: Box-and-whisker plot

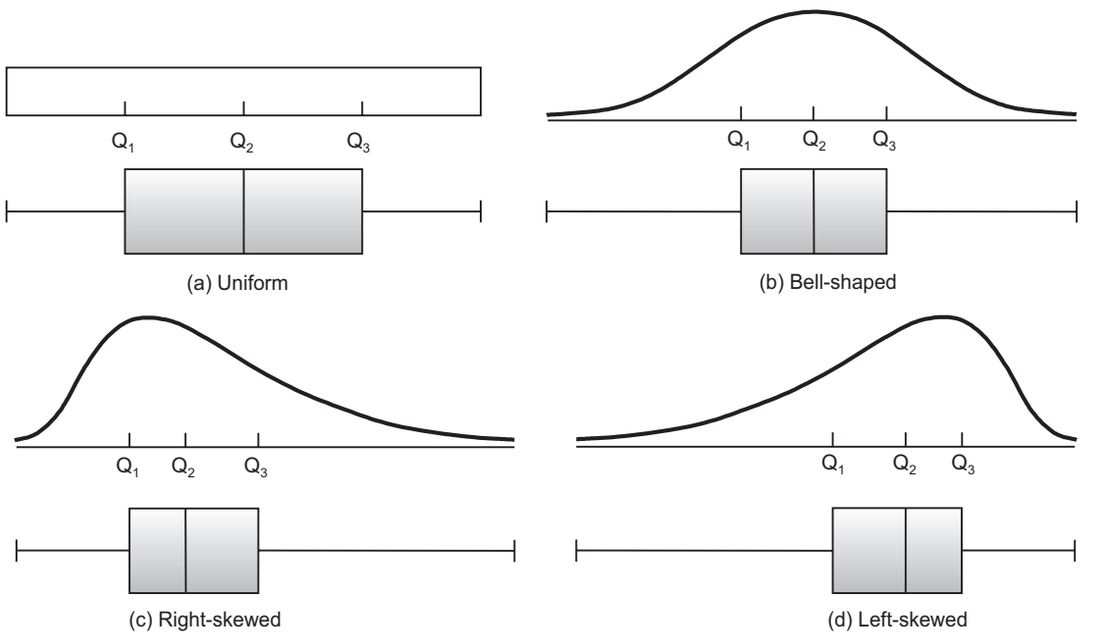


Fig. 1.12: Distribution shapes and box plots for (a) uniform, (b) bell-shaped, (c) right-skewed, (d) left-skewed distributions

Example E1.12

Refer to the data in the Example E1.11 and construct a box-and-whisker plot and comment on the skewness of the data.

SOLUTION:

Median = $Q_2 = 40$, $Q_1 = 36$, $Q_3 = 42$, IQR = $Q_3 - Q_1 = 42 - 36 = 6$

Lower inner fence = $Q_1 - 9 = 36 - 9 = 27$

Upper inner fence = $Q_3 + 9 = 42 + 9 = 51$

The box-and-whisker plot is shown in Fig. E1.12.

The longest and smallest values with the two inner fences are 47 and 35 respectively. There are two outliers 50 and 52 shown by an asterisk in Fig. E1.12. The data in this example is skewed to the right, because the lower 50% of the values are spread over a smaller range than the upper 50% of the value.

Descriptive Statistics

Variable	N	Mean	Median	Tr. Mean	St. Dev	SE Mean
C 1	24	38.83	40.00	39.27	7.81	1.59
Variable	Min.	Max.	Q_1	Q_3		
C 1	16.00	52.00	36.00	42		

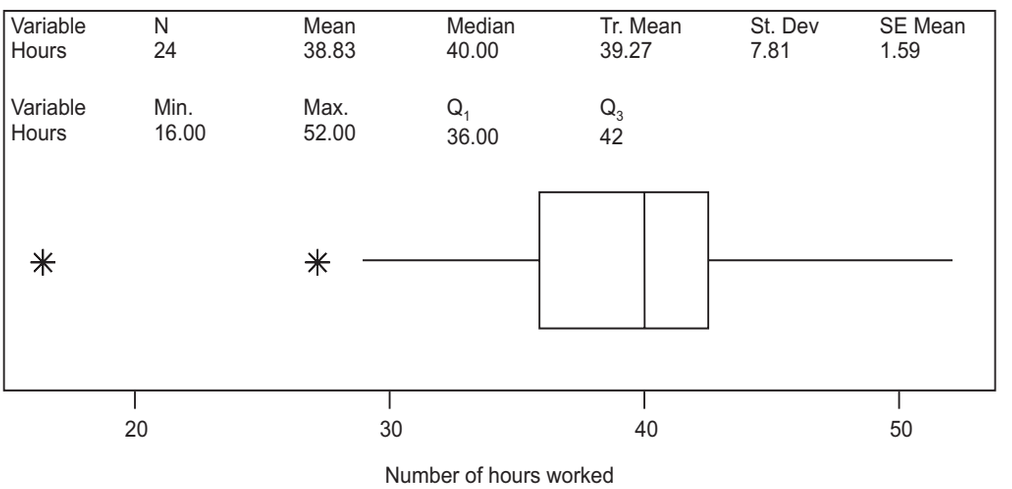


Fig. E1.12: Box plot

1.3 SUMMARY

In this chapter, we have introduced the types of statistics, namely, the descriptive statistics and the inferential statistics. Descriptive statistics consists of methods for organizing and summarizing information. Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

PROBLEMS

P1.1 The following data give the results of a sample survey. The letters *X*, *Y* and *Z* represent the three categories.

X Y Y X Z Y Z Z Z X Z Y Z Z Z

Z Y X X X X Y X X Y Z Y Y Z X

(a) prepare a frequency distribution table

(b) calculate the relative frequencies and percentages for all categories

(c) what percentage of the elements in the sample belongs to category *Y*?

(d) what percentage of the elements in the sample belongs to category *X* and *Z*?

(e) draw bar graph for the frequency distribution.

P1.2 Fifteen faculty members at a university were asked whether ethics course should be required of all students as a requirement for graduation. The responses of these faculty members are listed below. (F, A, and N indicate that a faculty member is in favour, against, or has no opinion, respectively).

N A A F N N F F F A A N F F A

(a) prepare a frequency distribution table

(b) find the relative frequencies and percentages for all categories

(c) what percentage of the faculty members in this sample are in favour of this issue?

(d) draw a pie chart for the percentage distribution.

P1.3 The following data give the weights (in kg) of a random sample of 30 senior students at a university. Construct a frequency distribution table.

81 77 79 75 80 83 82 83 85 77 80 83 74 72 86

73 78 73 76 82 84 79 80 79 84 76 78 81 79 78

P1.4 A sample of randomly selected households in a large city produced the following data on the number of vehicles owned.

5 1 1 1 1 1 1 2 2 1 1 1 3 2 2

2 2 4 1 1 0 3 2 1 1 1 0 1 1 4

1 2 2 4 1 2 4 2 3

Draw a bar graph for vehicles owned.

P1.5 The following table gives the frequency distribution of ages of all 50 employees of a company.

Age	Number of employees
18 to 30	12
31 to 43	18
44 to 56	14
57 to 69	6

(a) find the class boundaries and class width

(b) do all classes have the same width?

(c) prepare the relative frequency and percentage distributions

(d) find the percentage of the employees that are 43 years old or younger.

P1.6 The following data give the number of machines manufactured by a company for a sample of 30 days.

24	21	31	32	26	35	27	31	34	23	22	22	33	27	26
33	33	31	29	27	23	25	23	35	23	28	31	28	29	27

(a) construct a frequency distribution table using the classes 21–23, 24–26, 27–29, 30–32 and 33–35.

(b) calculate the relative frequencies and percentages for all classes

(c) construct a histogram and a polygon for the percentage distribution.

(d) for what percentage of the days is the number of machines produced in the interval 27 to 29?

P1.7 The academic ranks of fifteen faculty members in a department are listed below. F, A and AS indicate that a faculty member is a Full Professor, Associate Professor and Assistant Professor, respectively.

F AS A F A A F AS F A F AS F A F

(a) prepare a frequency distribution table

(b) compute the relative frequencies and percentages for all categories

(c) what percentages of the faculty in this sample is full professor rank?

(d) draw a pie chart for the percentage distribution.

P1.8 The results of a sample survey of three categories 1, 2 and 3 are given below:

1	1	3	2	3	1	2	1	2	2	3	1	3
2	1	2	1	1	3	3	3	2	3	3	2	2

(a) prepare a frequency distribution table

(b) calculate the relative frequencies and percentages for all three categories

(c) find the percentage of the elements in this sample belonging to category 2

(d) find the percentage of the elements in this sample belonging to category 1 or 3.

P1.9 The following data gives the number of car oil changes made at a service station for a sample of 30 days.

23	31	26	24	34	34	30	26	24	27
21	27	30	23	28	32	27	23	27	30
29	35	33	23	27	27	24	35	30	28

(a) construct a frequency distribution table using the classes

21–23, 24–26, 27–29, 30–32 and 33–35

(b) calculate the relative frequencies and percentages for all classes

(c) construct a histogram and a polygon for the percentage distribution

(d) for what percentage of the days is the number of oil changes made in the interval 30–32?

P1.10 The following data gives the 1999 autocollision claims paid by an insurance company to 24 claimants.

2301	1678	4634	4541	851	2306	1174	974
1132	1828	749	2585	971	2891	1275	725
2782	1161	2406	912	3510	820	1560	1208

(a) construct the cumulative frequency, cumulative relative frequency and cumulative percentage distribution table

(b) draw an ogive for the cumulative percentage distribution

(c) using the ogive, find the percentage of claimants in this sample who received $2500 or less.

P1.11 The following data gives the collision repair cost in dollars for 25 automobiles randomly selected form a list of 100 cars.

2410	739	2500	413	556
2470	1789	2109	899	987
2170	326	1343	1158	1234
6181	4983	5892	2312	3953
6721	4800	1325	2911	1926

(a) construct a frequency distribution table. Assume the width of each class as $1500

(b) calculate the relative frequencies and percentages for all classes

(c) draw a histogram and a polygon for the relative frequency distribution

(d) what are the class boundaries and the width of the fourth class?

P1.12 The following data give the number of unemployed adults in 2008 for twenty states.

95	328	132	49	48	37	167	231	90	67
32	23	270	24	287	347	35	20	16	107

(a) construct a frequency distribution table. Take the classes as 1–70, 71–140 and so on

(b) prepare the relative frequency and percentages.

P1.13 The following data give the 2008 total annual compensation (in thousands of dollars) for 29 chief executive officers of some large corporations.

2595	1678	4832	4433	952	2207	1075	4200
1135	2829	767	2555	977	2891	1276	2781
1160	2406	908	3501	811	1568	964	727
1208	1029	1385	1011	2815			

(a) construct a frequency distribution table. Take $1 thousand as the lower limit of the first class and $1000 thousand as the width of the class

(b) prepare the relative frequency and percentages.

P1.14 The following data give the number of children less than 18 years of age for 30 randomly selected families.

2	1	1	1	2	0	2	0	0	2	1	2	3	0	1
1	2	2	1	1	3	2	2	2	2	0	0	0	0	0

(a) prepare a frequency distribution table using single-valued classes
(b) calculate the relative frequencies and percentages for all classes
(c) how many families in this sample have 2 or 3 children less than 18 years of age?
(d) draw a bar graph for the frequency distribution.

P1.15 Refer to Problem P1.5.
(a) prepare a cumulative frequency distribution table
(b) calculate the cumulative relative frequencies and cumulative percentages for all classes
(c) what percentage of the employees are 44 years of age or older?
(d) draw an ogive for the cumulative percentage distribution
(e) using the ogive, find the percentage of employees who are 40 years old or younger.

P1.16 Refer to Problem P1.6. Prepare the cumulative frequency, cumulative relative frequency and cumulative percentage distribution.

P1.17 Refer to Problem P1.13. Prepare the cumulative frequency, cumulative relative frequency and cumulative percentage distribution by using the table in Problem P1.13.

P1.18 Following are the SAT scores (out of a maximum possible score of 1600) of 12 students who took this test.

786	892	997	1170	881	1042
995	1084	773	980	1128	1066

Prepare a stem-and-leaf display. Arrange the leaves for each stem in increasing order.

P1.19 The following data give the time (in minutes) taken to commute from residence to work for 20 workers of accompany.

10	50	64	34	47	5	12	23	37	27
26	32	17	7	13	19	29	43	21	22

Construct a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

P1.20 At the end of a particular training program, 400 students in a college are judged by a panel of examiners as *excellent*, *very good*, *good*, or *fair*. The following data give the distribution:

Excellent	80
Very good	140
Good	120
Fair	60

Represent the figures by a pie chart.

P1.21 The following are the test scores of 60 students on engineering ethics test.

43	46	48	50	51	51	52	52	53	53
55	55	56	56	57	58	60	66	67	67
68	68	70	71	72	73	73	74	75	75
75	75	77	77	78	78	76	79	79	80
80	80	80	81	82	82	82	83	83	85
84	84	85	86	86	86	88	91	92	93

Construct a ranked stem-and-leaf display of the data.

P1.22 Plot the stem-and-leaf diagram for the following cholesterol readings (mg/deciliter) of twenty patients.

184	231	195	186	241	190	239	254	225	237
211	224	214	197	204	233	198	215	216	205

P1.23 The following are the scores of 30 students on math test:

75	53	80	96	66	78	72	87	93	95
69	72	81	62	76	86	79	68	50	92
83	84	77	64	71	87	72	93	58	99

Construct a ranked stem-and-leaf display of the data.

P1.24 The following data give the monthly rent paid by a sample of 30 households selected from a city.

428	586	733	675	550	989	1021	620	750	660
540	578	956	1030	1070	930	871	765	880	975
651	1020	950	840	781	870	900	801	750	821

Construct a stem-and-leaf display for the above data.

P1.25 The following data gives the moneys donated by 32 university faculty members to a charity during the calendar year 2008.

473	417	173	109	178	287	117	175
281	474	589	379	89	235	414	357
412	608	454	317	355	611	469	601
453	287	489	484	616	354	189	388

(a) prepare a stem-and-leaf display for these data using the last two digits as leaves

(b) condense the stem-and-leaf display by grouping the stem as 0–1, 2–3 and 4–6.

P1.26 The following data gives the number of hours spent exercising in a gymnasium by 10 randomly selected members of a university during the past week.

8	13	6	0	1	6	11	4	0	7

Compute the mean, median and mode.

P1.27 The following data gives the number of vacuum cleaners produced by a manufacturing company for a sample of 12 days.

25	33	28	24	36	34	30	22	24	29	31	23

Calculate the mean, median and mode.

P1.28 The following data gives the number of heavy machinery sold by a large manufacturing company for the 12 months in the year 2008.

3	2	47	118	204	97
68	86	62	57	98	99

Find the mean, median and mode.

P1.29 The following data are the test scores (in points) of 9 students in a class out 6 points:

5.8 5.7 5.9 5.7 5.5 5.7 5.7 5.7 5.6

Find the mean, median and mode for this data.

P1.30 The following data gives the number of ball bearings manufactured by a company per day for a sample of 10 days.

24 32 27 23 35 33 29 21 23 28

Calculate the mean, median and mode for these data.

P1.31 The following data give the number of automobile accidents in a city during the past 12 days.

3 7 11 5 6 3 8 7 6 9 13 2

Calculate the range, variance and standard deviation.

P1.32 Compute the range, variance and standard deviation for the data given in Problem P1.26.

P1.33 The following data give the number of hours spent doing the homework and attending classes by 10 randomly selected college students during the past week.

18 16 27 34 39 35 42 29 31 24

Compute the range, variance and standard deviation.

P1.34 The following data give the number of new automobiles sold at a dealership during a 15 days period.

13 5 9 7 6 10 11 14 12 8 4 8 16 10 8

Find the range, mean, variance and standard deviation.

P1.35 Following are the 2008 earnings (in thousands of dollars) before taxes of all 10 employees of a small company:

30 35 22 28 19 47 52 43 39 32

Calculate the variance and standard deviation of these data.

P1.36 The following data give the number of automobiles that stopped at a service station during each of the 10 hours observed:

30 32 35 29 28 17 19 30 18 25

Find the range, variance and standard deviation.

P1.37 Consider the following two data sets:

Data set 1:	8	12	25	37	41
Data set 2:	15	19	32	44	48

Calculate the standard deviation of each data set. Comment on the results obtained.

P1.38 The following table gives the frequency distribution of the amounts of long distance telephone bills for the month of August 2008 for a sample of 60 families in a city.

Amount of long distance telephone bill (dollars)	Number of families
25 to less than 50	15
50 to less than 75	12
75 to less than 100	20
100 to less than 125	8
125 to less than 150	5

Calculate the mean, variance and standard deviation.

P1.39 Using the sample formulae, find the mean, variance and standard deviation for the grouped data given in the following table:

x	f
0 to less than 4	16
4 to less than 8	23
8 to less than 12	14
12 to less than 16	10
16 to less than 20	8
20 to less than 24	7

P1.40 The following table gives the frequency distribution of entertainment expenditure (in dollars) incurred by 50 families during the past week.

Entertainment expenditure (dollars)	Number of families
0 to less than 10	5
10 to less than 20	10
20 to less than 30	14
30 to less than 40	10
40 to less than 50	6
50 to less than 60	5

Find the mean, variance and standard deviation.

P1.41 The following table gives the grouped data on the weights of all 100 babies born at a hospital in 2005.

Weight (kg)	Number of babies
1 to less than 2	5
2 to less than 3	30
3 to less than 4	40
4 to less than 5	20
5 to less than 6	5

Find the mean, variance and standard deviation.

P1.42 The following table gives the frequency distribution of the number of personal computers sold during the past month at 40 computer stores in a big city. Calculate the mean, variance and standard deviation.

Computers sold	Number of stores
4 to 12	6
13 to 21	10
22 to 30	13
31 to 39	6
40 to 48	5

P1.43 The following table gives the number of television sets owned by 80 households.

Number of television sets owned	Number of households
0	4
1	32
2	27
3	11
4	6

Find the mean, variance, and standard deviation. Use the number of television sets owned as the values of m in the formulae.

P1.44 The following data gives the scores of 12 students in Design for Economy and Reliability class.

74 79 67 52 98 57 75 72 84 87 90 78

(a) calculate the values of the three quartiles and the interquartile range. Where does the value of 80 lies relative to these quartiles?

(b) calculate the value of the 93rd percentile. Comment on the calculated percentile.

(c) determine the percentile rank of 81. Comment on the percentile rank.

P1.45 The following data gives the number of hours worked in the last week by 30 employees of a manufacturing company.

38 40 42 45 35 47 28 40 43 39

53 40 51 23 42 49 40 51 36 40

49 33 21 31 40 33 39 16 36 41

(a) calculate the values of the three quartiles and the interquartile range

(b) calculate the approximate value of the 79th percentile

(c) calculate the percentile rank of 39.

P1.46 The following data gives the number of electronic components produced at ABC company for a sample of 30 days.

24 32 27 23 33 33 29 23 25 28

20 20 26 31 33 27 27 28 29 23

36 31 26 34 22 36 23 28 31 27

(a) calculate the values of the three quartiles and the interquartile range. Where does the value of 31 lie in relation to these quartiles?

(b) calculate the value of the 65th percentile

(c) calculate the percentage of days when the number of electronic components produced was 32 or higher.

P1.47 The following data gives the number of new computer systems sold during a 19 days period.

93 84 75 63 97 81 73 69 47 58

95 78 84 77 45 69 83 98 84

(a) calculate the values of the three quartiles and the interquartile range. Where does the value of 81 lies in relation to these quartiles?

(b) calculate the approximate value of the 93rd quartile

(c) find the percentile rank of 82.

P1.48 The following are the scores of 20 students in a Mathematics class:

48 82 82 83 70 73 78 74 46 81

78 74 64 80 79 80 80 79 80 80

(a) find the quartiles

(b) find the interquartile range

(c) find the five number summary.

P1.49 The random sample of 21 tourists yielded the following data on length of stay (in days) in a summer resort.

4 3 10 4 6 13 12 15 5 18 7

7 12 2 9 6 25 23 1 3 9

(a) find the quartiles

(b) find the interquartile range

(c) find the five number summary.

P1.50 The following are the scores of 12 students in a statistics class.

74 81 68 51 97 57 76 73 85 88 90 79

(a) find the values of the 3 quartiles

(b) find the interquartile range.

P1.51 The following are the ages of 9 employees of a small manufacturing company.

47 28 40 51 34 37 62 23 33

(a) find the values of the three quartiles

(b) find the interquartile range.

P1.52 Refer to Problem P1.50. Find the value of the 62nd percentile.

P1.53 Refer to Problem P1.50. Find the percentile rank for the score 85.

P1.54 A sample of 20 full-time students in a college yielded their weekly times spent on attending classes and doing their homework in hours. Determine the quartiles for these data.

43 6 16 21 32 14 38 20 27 31

32 30 25 64 34 30 41 35 26 38

P1.55 Refer to Problem P1.54.

(a) find the five-number summary

(b) find the interquartile range

(c) obtain the lower and upper limits

(d) identify potential outliers, if any.

P1.56 The following data gives the incomes in thousands of dollars for a sample of 12 employees in a company

22 16 31 59 21 51 28 37 41 91 26 45

Construct a box-and-whisker plot for this data set.

P1.57 The following table shows the number of used cars sold each week during 10 weeks period by a car dealership in a large city.

66 96 147 147 154 175 88 154 57 116

(a) find the 3 quartiles

(b) determine the interquartile range

(c) identify the potential outliers, if any

(d) construct a box plot.

P1.58 The following data give the time (in minutes) taken by 14 students to complete a Maths test.

73 91 82 71 74 98 95 89 88 79 93 96 87 77

Prepare a box-and-whisker plot. Comment on the skewness of these data.

P1.59 The following data give the income (in thousands of dollars) for a sample of 12 households.

17 23 32 22 52 60 29 38 92 42 27 46

Construct a box-and-whisker plot for these data.

P1.60 The following data give the hours worked by 30 employees of a company during one week.

43 39 27 40 42 40 35 45 38 47

40 51 36 40 40 23 42 53 51 48

36 41 39 16 48 21 31 34 40 34

Make a box-and-whisker plot.

P1.61 A corporation manufactures ball bearings. The following data give the number of bearings produced at the company for a sample of 30 days.

31 35 22 34 23 28 26 27 31 35

24 27 32 23 33 33 29 23 28 25

21 26 20 31 33 27 23 27 29 28

Prepare a box-and-whisker plot. Comment on the skewness of these data.

REVIEW QUESTIONS

1. Describe the need to group data in the form of a frequency table.
2. Describe how the relative frequencies and percentages of categories obtained from the frequencies of categories.
3. Describe the steps to be made to group a data set in the form of a frequency distribution table.
4. How are the relative frequencies and percentages of classes obtained from the frequencies of classes?

5. Describe the concept of cumulative frequency distribution. How are cumulative relative frequencies and cumulative percentages calculated?
6. Briefly explain how a stem-and-leaf display for a data set is prepared.
7. Describe very briefly the following:
   (a) bar graph *(b)* class
   (c) class boundary *(d)* class frequency
   (e) class width or size *(f)* grouped data
   (g) histogram *(h)* ogive
   (i) pie chart
8. Explain the difference between ungrouped and grouped data.
9. Briefly explain the meaning of the following terms:
   (a) median *(b)* outlier
   (c) mean *(d)* mode
10. Explain the relationship between mean, median and mode for symmetric and skewed histograms.
11. Explain the difference between a population parameter and a sample statistic.
12. Are the values of the mean and standard deviation that are calculated using grouped data exact or approximate values of the mean and standard deviation, respectively? Explain.
13. Describe how the three quartiles are calculated for a data set.
14. What is interquartile range?
15. Describe how the percentiles are calculated for a data set.
16. Describe the concept of the percentile rank for an observation of a data set.
17. Explain what summary measures are used to construct a box-and-whisker plot.
18. Define the following terms:
   (a) coefficient of variation *(b)* measures of dispersion
   (c) measures of position *(d)* parameter
   (e) range *(f)* first quartile
   (g) second quartile *(h)* statistic
   (i) third quartile *(j)* upper inner fence
   (k) upper outer fence *(l)* variance

STATE TRUE OR FALSE

1. Variable is a characteristic that varies from one person or thing to another. (True/False)
2. Qualitative variable is a numerically valued variable. (True/False)
3. Quantitative variable is a numerically valued variable. (True/False)
4. Discrete variable is a qualitative variable whose possible values can be listed. (True/False)
5. Continuous variable is a quantitative variable whose possible values form some interval of numbers. (True/False)
6. Data is values of a variable. (True/False)
7. Qualitative data is values of a qualitative variable. (True/False)

8. Continuous data is values of a continuous variable. (True/False)
9. Qualitative variable yields non-numerical data. (True/False)
10. Classes are categories for non-grouping data. (True/False)
11. Frequency is the number of observations that fall in a class. (True/False)
12. Frequency distribution is a listing of all classes and their relative frequencies. (True/False)
13. Relative-frequency distribution is a listing of all classes and their frequencies. (True/False)
14. Lower cut point is the smallest value that could go in a class. (True/False)
15. Upper cut point is the smallest value that could go in the next higher class (equivalent to the lower cut point of the next higher class). (True/False)
16. The middle of a class is found by summing its cut points. (True/False)
17. Width is the average between the cut points of a class. (True/False)
18. Grouping can help to make a large and complicated set of data more compact and easier to understand. (True/False)
19. The concepts of cut points and mid points make sense only for numerical data (for which doing arithmetic is meaningful). (True/False)
20. The frequency of a class is the number of observations in the class, whereas the relative frequency of a class is the ratio of the class frequency to the total number of observations. (True/False)
21. The percentages of a class equals 100 times the relative frequency of the class. (True/False)
22. The relative frequency of a class is the percentage of the class expressed as a decimal.(True/False)
23. Two data sets that have identical frequency distributions have identical relative-frequency distributions. (True/False)
24. Two data sets that have identical relative-frequency distributions have identical frequency distributions. (True/False)
25. The four elements of a grouped-data table are the classes, frequencies, relative frequencies, and mid points. (True/False)
26. The mid point of each class is not the same as the class. (True/False)
27. Relative frequencies always lie between 0 and 1 and hence provide a standard for comparison. (True/False)
28. A frequency histogram displays the class frequencies on the vertical axis, whereas a relative-frequency histogram displays the class relative frequencies on the vertical axis. (True/False)
29. Stem-and-leaf diagrams are generally useful with large data sets. (True/False)
30. The distribution of a data set is a table, graph, or formula that provides the values of the observations and how often they occur. (True/False)
31. Sample data are the values of a variable for a sample of the population. (True/False)
32. Population data are the values of a variable for an entire population. (True/False)
33. Census data is another name for sample data. (True/False)
34. A sample distribution is the distribution of population data. (True/False)

35. The population distribution is the distribution of population data. (True/False)
36. Distribution of a variable is another name for the population distribution. (True/False)
37. The measure of center is to indicate where the center or most typical value of data set lies. (True/False)
38. The three most important measures of center are the mean, the median and the mode. (True/False)
39. The mean, median and mode can all be used with quantitative data. (True/False)
40. If the number of observations is even, then the median is the observation exactly the middle of the ordered list. (True/False)
41. If the number of observations is odd, then the median is the mean of the two middle observations in the ordered list. (True/False)
42. The range of a data set is the difference between the maximum (largest) and minimum (smallest) observations. (True/False)
43. The more variation that there is in a data set, the larger is its standard deviation. (True/False)
44. Almost all the observations in any data set lie within three standard deviations to either side of the mean. (True/False)
45. The purpose of a measure of variation is to indicate the amount of variation in a data set. (True/False)
46. The first quartile is the median part of the entire data set that lies at or below the median of the entire data set. (True/False)
47. The second quartile is the median of the entire data set. (True/False)
48. The third quartile is the median part of the entire data set that lies at or below the median of the entire data set. (True/False)
49. The interquartile range is the difference between the first and third quartiles. (True/False)
50. The median and interquartile range are resistant measures, whereas the mean and standard deviations are not. (True/False)
51. Box plots are useful when the data set (number of observations) is large. (True/False)
52. Parameter is a descriptive measure for a population. (True/False)
53. Statistic is a descriptive measure for a sample. (True/False)
54. For an observed value of a variable x, the corresponding value of the standardised variable z is called the x-score of the observation. (True/False)
55. Numbers that are used to describe data sets are called descriptive measures. (True/False)
56. Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of dispersion. (True/False)
57. Descriptive measures that indicate the amount of variation or spread, in a data set are called measures of variation. (True/False)
58. Almost all the observations in any data set lie within 3 standard deviations to either side of the mean. (True/False)
59. An outlier is an observation that falls well outside the overall pattern of the data. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **False** 5. **True** 6. **True** 7. **True** 8. **True** 9. **True** 10. **False**
11. **True** 12. **False** 13. **False** 14. **True** 15. **True** 16. **False** 17. **False** 18. **True** 19. **True** 20. **True**
21. **True** 22. **True** 23. **True** 24. **True** 25. **True** 26. **False** 27. **True** 28. **True** 29. **False** 30. **True**
31. **True** 32. **True** 33. **False** 34. **False** 35. **True** 36. **True** 37. **True** 38. **True** 39. **True** 40. **False**
41. **False** 42. **True** 43. **True** 44. **True** 45. **True** 46. **True** 47. **True** 48. **False** 49. **True** 50. **True**
51. **True** 52. **True** 53. **True** 54. **False** 55. **True** 56. **False** 57. **True** 58. **True** 59. **True**

CHAPTER 2

Probability

In Chapter-1, we have discussed on the descriptive statistics, that is, methods for organising and summarising numerical data. Another important aspect is to present the fundamentals of inferential statistics, that is, methods of drawing conclusions about a population based on information from a sample of the population. Because inferential statistics involves utilising information from part of a population (a sample) to infer conclusions about the entire population, we can never be certain that our conclusions (inferences) are correct or true, that is, uncertainty is inherent in inferential statistics. Therefore, we need to be familiar with uncertainty before we can understand, develop, and apply the methods of inferential statistics.

The science of uncertainty is called *probability theory*. Probability theory enables us to evaluate and control the likelihood that a statistical inference is correct. More generally, probability theory provides the mathematical basis for inferential statistics.

This chapter introduces the basic concepts and definitions on probability, events (simple and compound), Venn diagram, tree diagram, approaches to probability (classical, relative frequency concept of probability, subjective probability, marginal probability and conditional probability. Special multiplication rule and Baye's formula are also briefly presented.

2.1 EXPERIMENT, OUTCOME AND SAMPLE SPACE

The tossing of a coin or the rolling of a die constitutes an *experiment*. In probability and statistics, the term *experiment* is used in a very wide sense and refers to any procedure that yields a collection of outcomes. The knowledge of all possible outcomes when a coin is tossed, or a die is rolled, is important. This is always the case for determining the probabilities. Measuring the length of a bolt, weighing the contents of a box of materials, and measuring the breaking strength of a metal component are all examples of experiments.

An *experiment* is a process that, when performed, results in one and only one of many observations. These observations are known as the *outcomes* of the experiment. The collection of all outcomes for an experiment is called a *sample space*. For tossing a coin, we can use the set {Heads, Tails} as the sample space. For rolling a six-sided die, we can use the set {1, 2, 3, 4, 5, 6}. These sample spaces are finite. Some experiments have sample spaces with an infinite number of outcomes. The elements of a sample space are called *sample points*. The *sample space* for an experiment can be described by drawing either a *Venn diagram*

or a *tree diagram*. A *Venn diagram* is a picture that depicts all the possible outcomes for an experiment. In a *tree diagram*, a branch of the tree represents each outcome.

2.2 SIMPLE AND COMPOSITE EVENTS

An *event* is a collection of one or more of the outcomes of an experiment. An event may be simple event or a compound event as shown in Fig. 2.1. A simple event is also called an *elementary event*, and a composite event is also called a *compound event*. An event that includes one and only one of the final outcomes for an experiment is called a *simple event* and is generally denoted by E_i. A compound event is a collection of more than one outcome of an experiment. The probability of a composite event is the sum of the probabilities of the simple events of which it is composed. When all the sample points or simple events of an experiment are equally likely and so have equal probabilities, then the probability of a composite event is easily found. For instance, if there are k sample points and a composite event A contains r of them, then $P(A) = \frac{r}{k}$.

The probability that a simple event E_1 will occur is denoted by $P(E_1)$, and the probability that a compound event A will occur is denoted by $P(A)$.

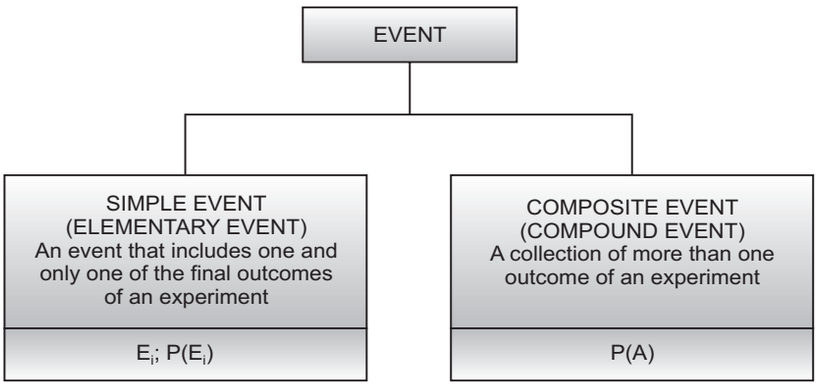


Fig. 2.1: Classification of an event

An *impossible event* is one that has no outcomes in it and consequently, cannot occur. On the other hand, a *sure event* is one that has all the outcomes of the sample space in it and will, therefore, definitely occur when the experiment is performed. Thus, the sample space constitutes a sure event.

Probability Definition: Let us assume that the sample space S has N outcomes e_1, e_2, e_N so that there are N simple events $\{e_1\}$, $\{e_2\}$, $\{e_N\}$.

The *probability of a simple event* $\{e\}$ is a number denoted by $P[\{e\}]$ and satisfies the following conditions:

1. $P[\{e\}]$ is always between 0 and 1 that is $0 \le P[\{e\}] \le 1$.
2. The sum of the probabilities of all the simple events is 1; that is, $P[\{e_1\}]+P[\{e_2\}] + \cdots + P[\{e_N\}] = 1$

The *probability of an event A*, denoted by $P(A)$, is defined as the sum of the probabilities assigned to the simple events that comprise the event A. The impossible event has probability 0 and the sure event has probability 1.

Example E2.1

Find the probability of getting *(a)* exactly two tails (event *A*), *(b)* at least two tails (event *B*) in tossing 3 balanced fair coins.

SOLUTION:

(a) The event *A* of exactly two tails comprises of the sample points (*TTH, THT, HTT*).

Hence, $$P(A) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

(b) The event *B* of at least two tails comprises of the sample points (*TTT*), (*TTH*), (*THT*) and (*HTT*).

Hence, $$P(B) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

Example E2.2

Find the probability that the sum of the numbers shown in the two faces, when two dice are thrown *(a)* is 8 and *(b)* 10.

SOLUTION:

(a) The event *A* that the sum of the numbers shown on the two faces is eight consists of the sample points:
(1, 7), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) and (7, 1)

There are 7 sample points. Therefore, the required probability is $\frac{7}{36}$.

(b) The event that the sum is 10 consists of the following sample points:
(4, 6), (5, 5) and (6, 4)

Hence, the required probability is $\frac{3}{36} = \frac{1}{12}$.

2.3 AXIOMS OF PROBABILITY

The subject of probability is based on three rules, known as axioms. They are:

1. Let *S* be a sample space. Then $P(S) = 1$.
2. For any event *A*, $0 \le P(A) \le 1$.
3. If *A* and *B* are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$
   More generally, if $A_1, A_2, A_3, \ldots$ are mutually exclusive events then $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + \cdots$

2.4 FINITE PROBABILITY SPACES

Consider a sample space *S* and the class *C* of all events. We assume here that if *S* is finite, then all subsets of *S* are events. *S* then become a probability space by assigning probabilities to the events in *C* so that they satisfy the probability axioms.

If S is a finite sample space with n elements and suppose the physical characteristics of the experiment suggest that the various outcomes of the experiment be assigned equal probabilities. Then S becomes a probability space, called a *finite equiprobable space*, if each point in S is assigned the probability $1/n$ and if each event A containing r points is assigned the probability r/n.

$$\text{Hence,} \quad P(A) = \frac{\text{number of elemens in } A}{\text{number of elements in } S} = \frac{n(A)}{n(S)} \tag{2.1}$$

$$\text{or} \qquad P(A) = \frac{\text{number of ways that the event } A \text{ can occur}}{\text{number of ways that the sample sapce } S \text{ can occur}} \tag{2.2}$$

Let $S = \{a_1, a_2, \ldots, a_n\}$ be a finite sample space. Then, a *finite probability space* or *finite probability model* is obtained by assigning each point a_i in S a real number p_i, called the probability of a_i, satisfying the following properties:

1. Each p_i is nonnegative, $p_i \geq 0$.
2. The sum of the $p_i = 1$, that is, $p_1 + p_2 + \cdots + p_n = 1$.

The *probability* $P(A)$ of an event A is defined to be the sum of the probabilities of the points in A.

2.5 INFINITE PROBABILITY SPACES

There are two cases: S is countably infinite and S is uncountable. A finite or countably infinite probability space S is said to be *discrete*. An uncountable space S which consists of a continuum of points is said to be *continuous*.

Case 1: Countably Infinite Sample Spaces

Suppose S is a countably infinite sample space: $S = \{a_1, a_2, \ldots\}$. Then, a probability space is obtained by assigning $a_i \in S$ a real number p_i, called its *probability*, such that

1. Each p_i is nonnegative, or $p_i \geq 0$.
2. The sum of the p_i is equal to 1

$$\text{or} \; p_1 + p_2 + \cdots = \sum_{i=1}^{\infty} p_i = 1.$$

The probability $P(A)$ of an event A is then equals the sum of the probabilities of its points.

Case 2: Uncountable Spaces

The probability of an event A is given by the ratio of $m(A)$ to $m(S)$:

$$P(A) = \frac{\text{length of } A}{\text{length of } S}$$

$$\text{or} \qquad P(A) = \frac{\text{area of } A}{\text{area of } S} \tag{2.3}$$

$$\text{or} \qquad P(A) = \frac{\text{volume of } A}{\text{volume of } S}$$

where the uncountable sample spaces S are those with some finite geometrical measurement such as length, area, or volume and in which a point is selected at random. Such a probability space is said to be *uniform.*

2.6 PROPERTIES OF PROBABILITY

The basic properties of probabilities are:

1. The probability of an event always lies in the range of 0 and 1. That is

   $0 \le P(E_1) \le 1$ Simple event

   $0 \le P(A) \le 1$ Compound event

2. The sum of the probabilities of all simple events or final outcomes for an experiment, denoted by $\Sigma P(E_i)$, is always equal to 1.

   Hence, $$\Sigma P(E_i) = P(E_1) + P(E_2) + \cdots = 1 \tag{2.4}$$

3. The probability of an event that cannot occur is 0. (An event that cannot occur is called an *impossible event*).
4. The probability of an event that must occur is 1. (An event that must occur is called a *certain event*).

2.7 VENN DIAGRAM

Graphical display of events are helpful for explaining and understanding probability. Venn diagrams, named after English logician John Venn (1834–1923) are considered one of the excellent ways to visually display events and relationships among events. The sample space is shown as a rectangle and the various events are drawn as circles (or other geometric shapes) inside the rectangle. Venn diagram for one event is shown in Fig. 2.2.

Each event A has a corresponding event defined by the condition that "A does not occur". That event is called the *complement* of A and is denoted (not A) or (A') as shown in Fig. 2.2(b). Event (not A) consists of all outcomes not in A, as shown in the Venn diagram in Fig. 2.2(a). With any two events, say, A and B, we can associate two new events. One new event is defined by the condition that "both event A and event B occur" and is denoted (A & B) or (A and B). Event (A and B) consists of all outcomes common to both event A and event B, as shown in Fig. 2.2(d). The other new event associated with A and B is defined by the condition "either event A or event B or both occur" or equivalently, that, "at least one of events A and B occur". That event is denoted (A or B) and consists of all outcomes in either event A or event B or both, as shown in Fig. 2.2(c).

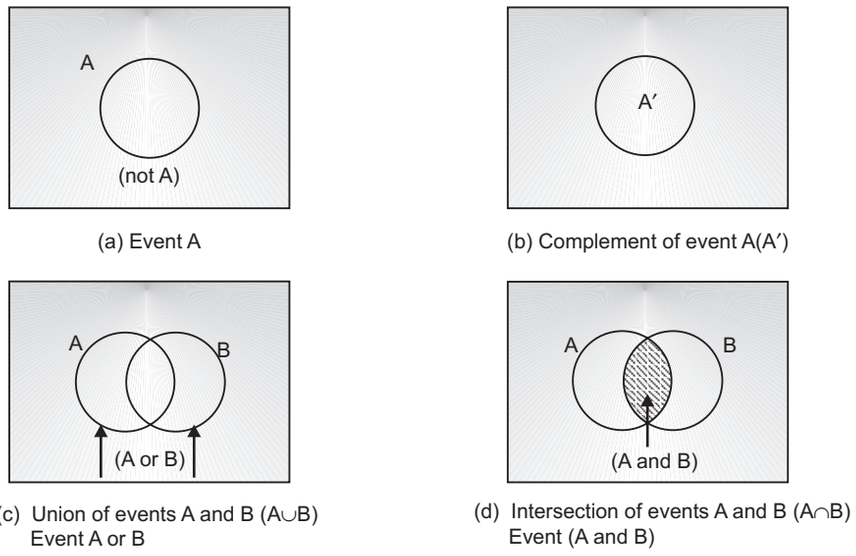
Relationship among Events

(not A): The event "A does not occur".

(A and B) or (A & B): The event "both A and B occur".

(A or B): The event "either A or B or both occur".

Note here that event "both A and B occurs" is the same as the event "both B and A occur", event (A and B) or (A & B) is the same as event (B and A) or (B & A). Similarly, event (A or B) is the same as event (B or A).



(a) Event A

(b) Complement of event A(A′)

(c) Union of events A and B (A∪B) Event A or B

(d) Intersection of events A and B (A∩B) Event (A and B)

Fig. 2.2: Venn diagram

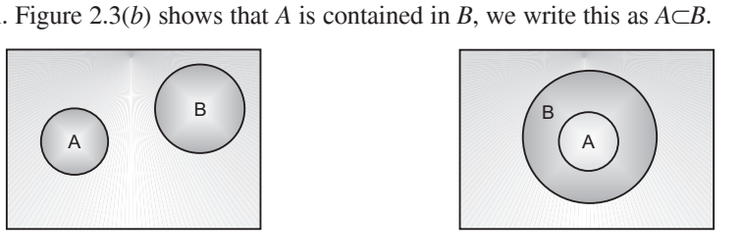
2.8 PROBABILITY TREE OR TREE DIAGRAM

In a *(finite) stochastic process*, in which a finite sequence of experiments where each experiment has a finite number of outcomes with given probabilities, the most convenient way to describe such a process is by means of a *labeled tree diagram.*

Various events are drawn as circles (or other geometric shape such as a rectangle, a square, or a circle) that depicts all the possible outcomes for an experiment. For instance, the shaded regions of the four Venn diagrams of Fig. 2.2 represents respectively, event A, the complement of event A, the union of events A and B, and the intersection of events A and B.

If A and B are any two subsets of a sample space S, their union $A \cup B$ is the subset of S that contains all the elements that are either in A, in B, or in both; their intersection $A \cap B$ is the subset. S that contains all the elements that are in both A and B; on the complement A' of A is the subset of S that contains all the elements of S that are not in A.

Figure 2.3(a) indicates that events A and B are *mutually exclusive*, that is, the two sets have no elements in common (or the two events cannot both occur). This is written $A \cap B = \phi$ denotes the *empty set*, which has no elements at all. Figure 2.3(b) shows that A is contained in B, we write this as $A \subset B$.



(a) Events A and B mutually exclusive

(b) Event A contains in Event B

Fig. 2.3: Venn diagram sharing special relationships among events A and B

Example E2.3

A bin contains a certain number of manufactured mechanical components, a few of which are defective. Two components are selected at random from this bin and inspected to determine, if they are non-defective or defective. How many total outcomes are possible? Draw a tree diagram for this experiment. Show all the outcomes in a Venn diagram.

SOLUTION:

Let G = the selected component is good

D = the selected component is defective

The four outcomes for this experiment are: GG, GD, DG and DD.

The Venn and tree diagrams are shown in Figs. E2.3(a) and (b).

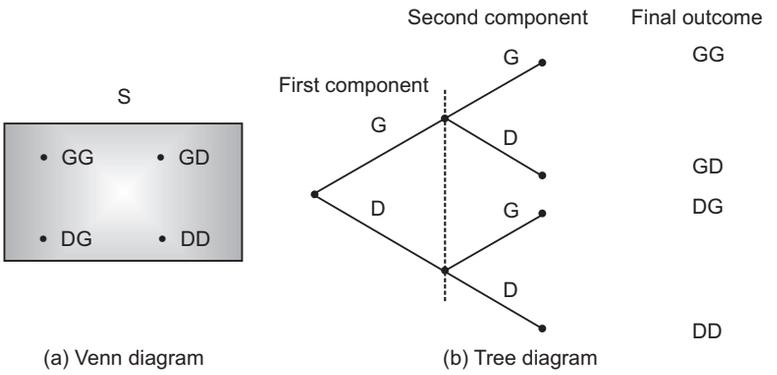


Fig. E2.3: Venn and tree diagrams

Example E2.4

In Fig. E2.4, H is the event that an employee has health insurance and D is the event that the employee has disability insurance.

(a) express in words what events are represented by regions 1, 2, 3 and 4

(b) what events are represented by regions 1 and 2 together

(c) regions 2 and 4 together

(d) regions 1, 2 and 3 together

(e) regions 2, 3 and 4 together.

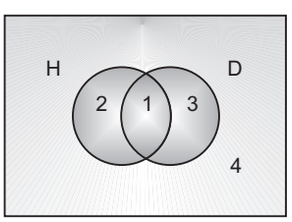


Fig. E2.4: Venn diagram

SOLUTION:

(a) 1 : The employee has health insurance and disability insurance
2 : The employee has health insurance but no disability insurance
3 : The employee has disability insurance but no health insurance
4 : The employee has neither health insurance nor disability insurance

(b) The employee has health insurance.

(c) The employee does not have health insurance.

(d) The employee has either health or disability insurance, but not both.

(e) The employee does not have both kinds of insurance.

Example E2.5

A small box contains a few red and a few blue balls. If two balls are randomly drawn and the colours of these balls are observed, how many total outcomes are possible? Draw a tree diagram for this experiment. Show all the outcomes in a Venn diagram.

SOLUTION:

Let R = a red ball is selected

B = a blue ball is selected

The experiment has 4 outcomes: *RR*, *RB*, *BR* and *BB*.

The Venn diagram and tree diagram are shown in Figs. E2.5(*a*) and (*b*).

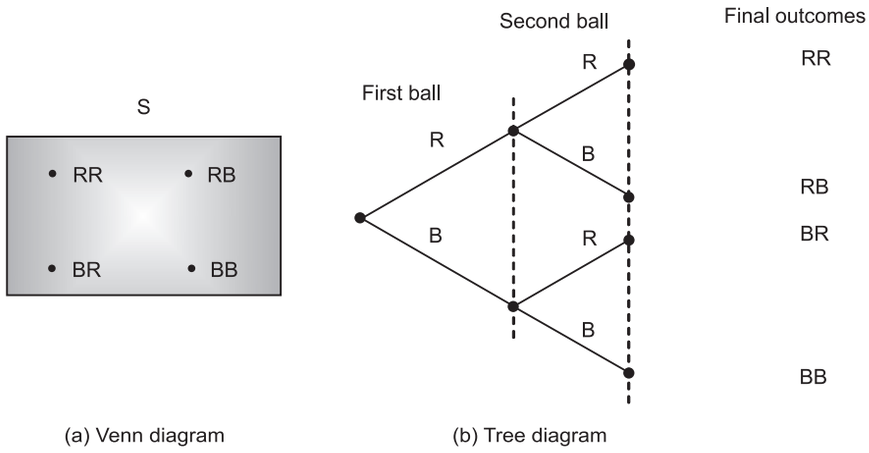


Fig. E2.5

Example E2.6

Draw a tree diagram for 3 tosses of a balanced (fair) coin. List all outcomes for this experiment in a sample space *S*.

SOLUTION:

Let H = a toss results in head

T = a toss results in tail.

Hence, the sample space is written as $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

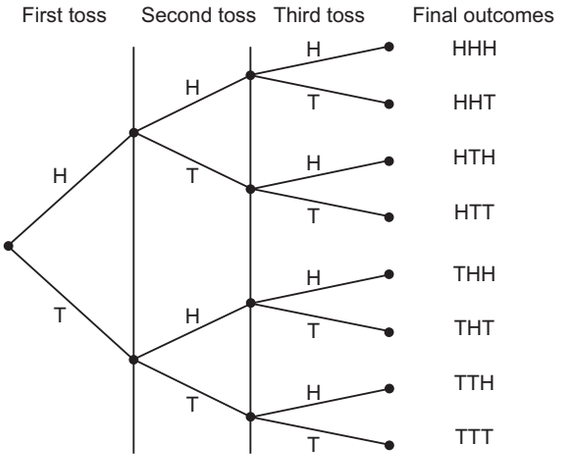


Fig. E2.6: Tree diagram

Example E2.7

Refer to Example E2.5. List all the outcomes included in each of the following events. Indicate which are simple and which are compound events.

(a) both balls are different colours

(b) at least one ball is red

(c) not more than one ball is blue

(d) the first ball is blue and the second is red.

SOLUTION:

(a) $\{RB, BR\}$; a compound event

(b) $\{RR, RB, BR\}$; a compound event

(c) $\{RR, RB, BR\}$; a compound event

(d) $\{BR\}$; a simple event.

2.9 APPROACHES TO PROBABILITY

There are three conceptual approaches to probability. They are

1. classical probability,
2. the relative frequency concept of probability and
3. the subjective probability concept.

2.9.1 Classical Probability

Outcomes that have the same probability of occurrence are called *equally likely outcomes*.

Two or more outcomes or events that have the same probability of occurrence are said to be *equally likely outcomes* or *events*. If E is an event, then $P(E)$ stands for the probability that event E occurs.

$$P(E_i) = \frac{1}{\text{total number of outcomes for the experiment}} \tag{2.5}$$

$$P(A) = \frac{\text{number of outcomes favourable to } A}{\text{total number of outcomes for the experiment}} \tag{2.6}$$

2.9.2 Relative Frequency Concept of Probability

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

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

The relative frequencies are not probabilities but approximate probabilities.

The *law of large numbers* states that if an experiment is repeated again and again, the probability of an event obtained from the relative frequency approaches the actual or theoretical probability.

2.9.3 Subjective Probability

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

2.9.4 Marginal Probability

Marginal probability is the probability of a single event without consideration of any other event. Marginal probability is also known as *simple probability*.

2.9.5 Conditional Probability

It 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 and B is written as $P(A|B)$ and read as "the probability of A given that B has already occurred".

The conditional probability of an event is the probability that the event occurs under the assumption that another event has occurred. The probability that event B occurs given that event A has occur is called a *conditional probability*. It is denoted by the symbol $P(B/A)$, which is read as "the probability of B given A". A is called the given event.

Joint and Marginal Probabilities

Data obtained by observing values of one variable of a population are known as *univariate data*. Data obtained by observing two variables of a population are called *bivariate data*, and a frequency distribution for bivariate data is called a *contingency table* or *two-way table*.

Example E2.8

A small bin contains 50 rubber balls. Of them, 26 are red and 24 are green. If one ball is randomly selected out of this bin, what is the probability that this ball is

(a) red

(b) green

SOLUTION:

(a) $P(\text{ball selected is red}) = \frac{26}{50} = 0.52$

(b) $P(\text{ball selected is green}) = \frac{24}{50} = 0.48$

Example E2.9

A multiple-choice question in a test contains five answers. If a student chooses one answer based on "pure guess", what is the probability that this student answer is

(a) correct

(b) wrong

Does these probabilities add up to 1? If yes, why?

SOLUTION:

(a) $P(\text{student's answer is correct}) = \frac{1}{5} = 0.2$

(b) $P(\text{student's answer is wrong}) = \frac{4}{5} = 0.8$

Yes, these probabilities add up to 1 because this experiment has two and only two outcomes, and according to the second property of probability, the sum of their probabilities must be equal to 1.

Example E2.10

A sample of 1000 families showed that 35 of them own no automobiles, 207 own one automobile each, 377 own two automobiles each, 264 own 3 automobiles each and 117 own 4 or more automobiles each. Write a frequency distribution for this problem. If one family is chosen randomly from these 1000 families, find the probability that this family owns

(a) two automobiles

(b) four or more automobiles.

SOLUTION:

Refer to Table E2.10.

(a) $P(\text{family selected owns 2 automobiles}) = 0.377$

(b) $P(\text{family selected owns 4 or more automobiles}) = 0.117$

Table E2.10

Automobiles owned	Frequency	Relative frequency
0	35	0.035
1	207	0.207
2	377	0.377
3	264	0.264
4 or more	117	0.117
Total	1000	1.000

2.10 MUTUALLY EXCLUSIVE EVENTS

Events that cannot occur together are known as *mutually exclusive events*. Such events do not have any common outcomes. If two or more events are mutually exclusive, then at most one of them will occur every time we repeat the experiment. Hence, the occurrence of one event excludes the occurrence of the other event or events. For example, consider tossing a balanced coin twice. This experiment has 4 outcomes: *HH*, *HT*, *TH* and *TT*. These outcomes are mutually exclusive since one and only one of them will occur when we toss this coin twice.

Two or more events are *mutually exclusive events* if no two of them have outcomes in common. The Venn diagram shown in Fig. 2.4 show the difference between two events that are mutually exclusive and two events that are not mutually exclusive. Similarly Fig. 2.5 shows three mutually exclusive events and two cases of three events that are not mutually exclusive. Two events are said to be *independent*, if the occurrence of one does not affect the probability of the occurrence of the other. Thus, if A and B are *independent events*, then either $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$. 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.

Hence, two events will be dependent if either $P(A \mid B) \neq P(A)$ or $P(B \mid A) \neq P(B)$. Two events are either *mutually exclusive* or *independent*. That is, mutually exclusive, events are always dependent, and dependent and independent events are never exclusive. Similarly, dependent events may or may not be mutually exclusive.

Two mutually exclusive events that are taken together include all the outcomes for an experiment are called *complementary event*. Thus, the complement of event A, denoted by $\bar{A}$ is the event that includes all the outcomes for an experiment that are not in A. It is clear that

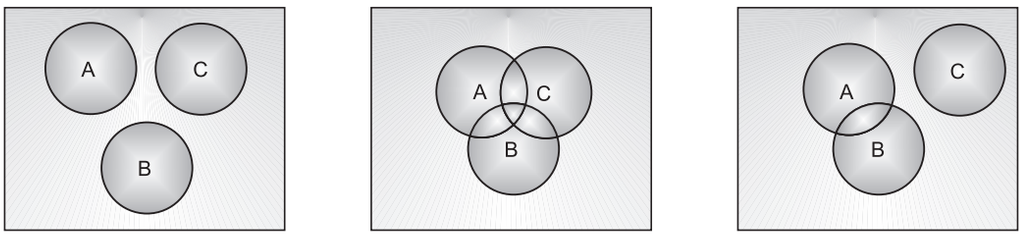
$$P(A) + P(\bar{A}) = 1 \tag{2.7}$$

Also $$P(A) = 1 - P(\bar{A}) \text{ and } P(\bar{A}) = 1 - P(A) \tag{2.8}$$

The *intersection* of two events is given by the outcomes that are common to both events. If A and B are two events defined in a sample space, then, the intersection of A and B represent the collection of all outcomes that are common to both A and B and is denoted by A and B or AB or BA.

(a) (b)

Fig. 2.4



(a) (b) (c)

Fig. 2.5

2.11 INDEPENDENT AND DEPENDENT EVENTS

Two events are said to be *independent* if the occurrence of one does not affect the probability of the occurrence of the other. In other words, A and B are *independent events* if either

$$P(A \mid B) = P(A) \text{ or } P(B \mid A) = P(B) \tag{2.9}$$

It can be shown that if one of these two conditions is true, then the second will also be true, and if one is not true then the second will also not be true.

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 the probability notation, the two events will be *dependent* if either

$$P(A \mid B) \neq P(A) \tag{2.10}$$

or $$P(B \mid A) \neq P(B) \tag{2.11}$$

2.12 COMPLEMENTARY EVENTS

The complement 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.

Events A and $\bar{A}$ are complements of each other. The Venn diagram in Fig. 2.6 shows the complementary events A and $\bar{A}$.

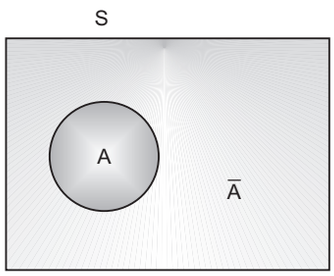


Fig. 2.6: Venn diagram of two complementary events

Since two complementary events are taken together, they include all the outcomes for an experiment and because the sum of the probabilities of all outcomes is 1, it is clear that

$$P(A) + P(\bar{A}) = 1 \tag{2.12}$$

Also $$P(A) = 1 - P(\bar{A}) \tag{2.13}$$

and $$P(\bar{A}) = 1 - P(A) \tag{2.14}$$

Example E2.11

A group of 2000 randomly selected adults were asked if they are in favour or against building a nuclear power plant in the city. The following table gives the results of this survey.

	In favour	Against	Total
Male	500	400	900
Female	650	550	1100
Total	1150	950	2000

(a) If one person is selected at random from these 2000 adults, find the probability that this person is
- *(i)* in favour of the plant
- *(ii)* against the plant
- *(iii)* in favour of the plant given the person is a female
- *(iv)* a male given the person is against the plant

(b) Are the events "male" and "in favour" mutually exclusive? What about the events "in favour" and "against"? Why or why not?

(c) Are the events "female" and "male" independent? Why or why not?

SOLUTION:

(a) *(i)* $P(\text{in favour}) = \dfrac{1150}{2000} = 0.575$

(ii) $P(\text{against}) = \dfrac{950}{2000} = 0.475$

(iii) $P(\text{in favour}|\text{female}) = \dfrac{650}{1100} = 0.59091$

(iv) $P(\text{male}|\text{against}) = \dfrac{400}{950} = 0.42105$

(b) The events "male" and "in favour" are not mutually exclusive because they can occur together. The events "in favour" and "against" are mutually exclusive because they cannot occur together.

(c) $P(\text{female}) = \dfrac{1100}{2000} = 0.55$

$P(\text{female}|\text{in favour}) = \dfrac{650}{1150} = 0.56522$

Since these two probabilities are not equal, the events "female" and "in favour" are not independent.

Example E2.12

There are a total of 170 practicing physicians in a big city. Of them, 60 are female and 30 are pediatricians. Of the 60 females, 10 are pediatricians. Are the events "female" and "pediatrician" independent? Are they mutually exclusive? Explain why or why not.

SOLUTION:

$$P(\text{pediatrician}) = \frac{30}{170} = 0.17647$$

$$P(\text{pediatrician}|\text{female}) = \frac{10}{60} = 0.16667$$

Since these two probabilities are not equal, the events "female" and "pediatrician" are not independent. The events are not mutually exclusive because they can occur together.

Example E2.13

A company hired 50 new graduates last month. Of these 25 are male and 15 are business majors. Of the 25 males, 15 are business majors. Are the events "male" and "business major" independent? Are they mutually exclusive? Explain why or why not.

SOLUTION:

$$P(\text{business major}) = \frac{15}{50} = 0.30$$

$$P(\text{business major}|\text{male}) = \frac{15}{25} = 0.60$$

Since these two probabilities are not equal, the events "male" and "business major" are not independent. The events are not mutually exclusive because they can occur together.

Example E2.14

Let A be the event that a number less than 3 is obtained if a die is rolled once. What is the probability of A? What is the complementary event of A, and what is its probability?

SOLUTION:

Event A will occur if either 1-spot or a 2-spot is obtained on the die. Hence,

$$P(A) = \frac{2}{6} = 0.3334$$

The complementary event of A is that either a 3-spot or 4-spot or a 5-spot, or a 6-spot is obtained on the die. Hence,

$$P(\bar{A}) = 1 - 0.3334 = 0.6666$$

2.13 INTERSECTION OF EVENTS AND MULTIPLICATION RULE

In this section, we introduce the intersection of two events and the application of multiplication rule to find the probability of the intersection of events.

2.13.1 Intersection of Events

Let A and B be two events defined in a sample space. The *intersection* of A and B represents the collection of all outcomes that are common to both A and B and is denoted by A and B. The intersection of events A and B is also denoted by either $A \cap B$ or AB. Figure 2.7 shows the intersection of events A and B. The shaded area in Fig. 2.7 gives the intersection of events A and B.

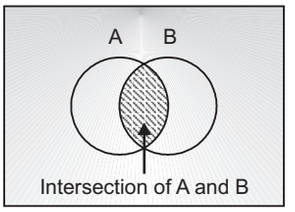


Fig. 2.7: Intersection of events A and B

2.13.2 Multiplication Rule

The probability that events A and B happen together is called the *joint probability* of A and B and is written as $P(A \text{ and } B)$. The probability of the intersection of two events is obtained by multiplying the marginal probability of one event by the conditional probability of second event. The rule is called the *multiplication rule.* Hence,

$$P(A \text{ and } B) = P(A)\, P(B|A) \tag{2.15}$$

The joint probability of events A and B can also be denoted by $P(A \cap B)$ or $P(AB)$.

Hence, for any two events, their joint probability is equal to the probability that one of the events occurs times the conditional probability of the other given event.

Conditional Probability

If A and B are two events, then,

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \tag{2.16}$$

and

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2.17}$$

given that $P(A) \neq 0$ and $P(B) \neq 0$.

Hence, for any two events A and B, the conditional probability that one event occurs given that the other event has occurred equals the joint probability of the two events divided by the probability of the given event.

Multiplication Rule for Independent Events

The probability of the intersection of two independent events A and B is

$$P(A \text{ and } B) = P(A)\, P(B) \tag{2.18}$$

The Special Multiplication Rule

The general multiplication rule for any two events A and B is given by Eq. (2.16)

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

If A and B are independent events, then

$$P(B \mid A) = P(B) \tag{2.19}$$

Hence, for the special case of independent events, we replace the term $P(B \mid A)$ in the general multiplication rule by the term $P(B)$.

If A and B are independent events, then

$$P(A \text{ and } B) = P(A) \cdot P(B) \tag{2.20}$$

and conversely, if

$$P(A \text{ and } B) = P(A)\, P(B)$$

then A and B are independent events. Hence, two events are independent if and only if their joint probability equals the product of their marginal probabilities.

Similarly, if events A, B, C,… are independent, then

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots) = P(A) \cdot P(B) \cdot P(C) \ldots \tag{2.21}$$

Mutually exclusive events are those that cannot occur simultaneously. *Independent events* are those for which the occurrence of some does not affect the probabilities of the others occurring. Two or more (non-impossible) events cannot be both mutually exclusive and independent.

Total Probability Rule

Events $A_1, A_2, A_3, \ldots, A_k$ are said to be *exhaustive*, if one or more of them must occur. Suppose events $A_1, A_2, A_3, \ldots, A_k$ are mutually exclusive and exhaustive; that is exactly one of the events must occur. Then for any event B

$$P(B) = \sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j) \tag{2.22}$$

Joint Probability of Mutually Exclusive Events

The joint probability of two mutually exclusive events is always 0. If A and B are two mutually exclusive events, then

$$P(A \text{ and } B) = 0 \tag{2.23}$$

Example E2.15

In a group of 10 adults, 4 have type A personality and 6 have a type B personality. If two persons are selected at random from this group of 10, what is the probability that the first of them has a type A personality and the second has a type B personality? Draw a tree diagram for this experiment.

SOLUTION:

Let C = first person selected has a type A personality

D = first person selected has a type B personality

E = second person selected has a type A personality

F = second person selected has a type B personality

Figure E2.15 shows the tree diagram for the experiment.

The probability that the first person has a type A personality and the second has a type B personality is given by

$$P(C \text{ and } F) = P(C)\, P(F|C) = \left(\frac{4}{10}\right)\left(\frac{6}{9}\right) = 0.2667.$$

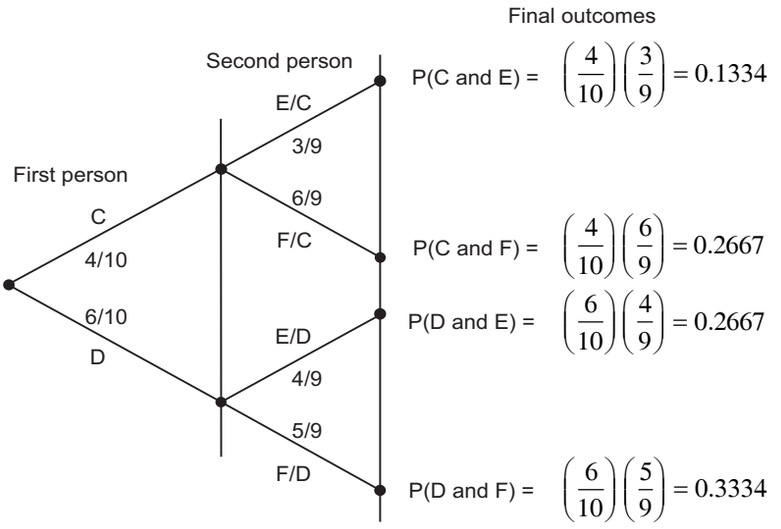


Fig. E2.15: Tree diagram

Example E2.16

Ten per cent of all items sold by a mail order company are returned by customers for a refund. Find the probability that of two items sold during a given hour by this company.

(a) both will be returned for a refund

(b) neither will be returned for a refund

Draw a tree diagram for this experiment.

SOLUTION:

Let A = first item is returned

B = first item is not returned

C = second item is returned

D = second item is not returned

(a) $P(A \text{ and } C) = P(A)\, P(C) = (0.10)(0.10) = 0.01$

(b) $P(B \text{ and } D) = P(B)\, P(D) = (0.90)(0.90) = 0.81$

The tree diagram for this equipment is shown in Fig. E2.16.

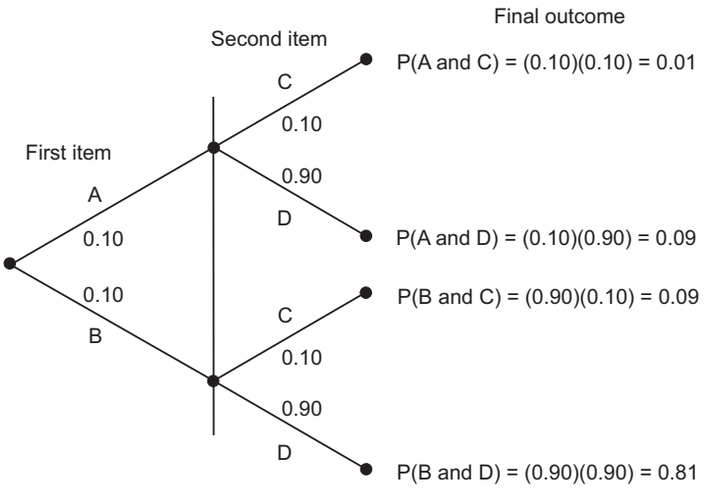


Fig. E2.16: Tree diagram

Example E2.17

The probability that a farmer is in debt is 0.80. What is the probability that three randomly selected farmers are all in debt? Assume independence of events.

SOLUTION:

Let D_1 = the first farmer is in debt

D_2 = the second farmer is in debt

D_3 = the third farmer is in debt

Hence, $P(D_1 \text{ and } D_2 \text{ and } D_3) = P(D_1)P(D_2)P(D_3) = (0.80)(0.80)(0.80) = 0.512$

Example E2.18

According to some private estimates, the probability that a randomly selected student from the population of all students enrolled in all institutions of higher education in a particular state is a female is 0.60, and the probability that this student is a female and a part-time student is 0.25. What is the probability that this student is part-time given that the student is a female?

SOLUTION:

Let F = student selected is a female

PT = student selected is a part-time

It is given that

$P(F) = 0.60$ and $P(PT \text{ and } F) = 0.25$

Hence, $$P(PT \mid F) = \frac{P(PT \text{ and } F)}{P(F)} = \frac{0.25}{0.60} = 0.41667$$

2.14 UNION OF EVENTS AND THE ADDITION RULE

In this section, we will discuss the union of events and the addition rule to compute the probability of the union of events.

2.14.1 Union of Events

The *union of two events A* and *B* includes all outcomes that are either in *A* or in *B* or in both *A* and *B*. Let *A* and *B* be two events defined in a sample space. The *union of events A* and *B* is the collection of all outcomes that belong either to *A* or to *B* or to both *A* and *B* and is denoted by *A* or *B*.

The union of events *A* and *B* is also denoted by "$A \cup B$".

2.14.2 Addition Rule

The method used to calculate the probability of the union of events is called the addition rule.

The probability of the union of two events *A* and *B* is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where $P(A)$ and $P(B)$ are the marginal probabilities of *A* and *B* respectively and $P(A \text{ and } B)$ is the joint probability of *A* and *B*.

Addition Rule for Mutually Exclusive Events

The probability of the union of two mutually exclusive events *A* and *B* is

$$P(A \text{ or } B) = P(A) + P(B) \tag{2.24}$$

Special Addition Rule

If event *A* and event *B* are mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B) \tag{2.25}$$

Generalizing, if events *A*, *B*, *C*, … are mutually exclusive, then

$$P(A \text{ or } B \text{ or } C \text{ or } \ldots) = P(A) + P(B) + P(C) + \cdots \tag{2.26}$$

Therefore, for mutually exclusive events, the probability that one or another of the events occurs equals to the sum of the individual probabilities.

The General Addition Rule

For events that are not mutually exclusive, the general addition rule is applied. Referring to Fig. 2.8, we see that

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{2.27}$$

For any two events, the probability that one or the other occurs equals the sum of the individual probabilities less the probability that both occur.

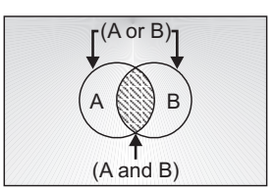


Fig. 2.8: Non-mutually exclusive events

Example E2.19

All 500 employees of a company were asked whether they are smokers or non-smokers of cigarettes and whether or not they are college graduates. Based on this information, the following two-way classification table was prepared:

	College graduate	Not a college graduate	Total
Smoker	55	100	155
Non-smoker	150	195	345
Total	205	295	500

Suppose one employee is selected at random from this company. Find the following probabilities:

(a) P(college graduate or smoker)

(b) P(smoker or not a college graduate)

(c) P(smoker or non-smoker)

SOLUTION:

Let S = smoker
N = non-smoker
C = college graduate
D = not a college graduate

(a) $P(C \text{ or } S) = P(C) + P(S) - P(C \text{ and } S) = \frac{205}{500} + \frac{155}{500} - \frac{55}{500} = 0.60$

(b) $P(S \text{ or } D) = P(S) + P(D) - P(S \text{ and } D) = \frac{155}{500} + \frac{295}{500} - \frac{100}{500} = \frac{350}{500} = 0.7$

(c) Since S and N are mutually exclusive events

$$P(S \text{ or } N) = P(S) + P(N) = \frac{155}{500} + \frac{345}{500} = 1.0$$

Example E2.20

The probability that a family owns a desktop computer is 0.78 that it owns a DVD player is 0.70, and that it owns both the computer and a DVD is 0.60. Find the probability that a randomly selected family owns a computer or a DVD player.

SOLUTION:

Let C = family selected owns a computer
D = family selected owns a DVD player

Then, $P(C \text{ or } D) = P(C) + P(D) - P(C \text{ and } D) = 0.78 + 0.70 - 0.60 = 0.88$

Example E2.21

The probability that an open-heart surgery is successful is 0.85. What is the probability that in two randomly selected open-heart surgeries at least one will be successful? Draw a tree diagram for this experiment.

SOLUTION:

Let A = first open-heart surgery is successful

B = first open-heart surgery is not successful

C = second open-heart surgery is successful

D = second open-heart surgery is not successful

P(at least one open-heart surgery is successful) = $P(A \text{ and } C) + P(A \text{ and } D) + P(B \text{ and } C)$

$= (0.85)(0.85) + (0.85)(0.15) + (0.15)(0.85) = 0.7225 + 0.1275 + 0.1275 = 0.9775$

The tree diagram is shown in Fig. E2.21.

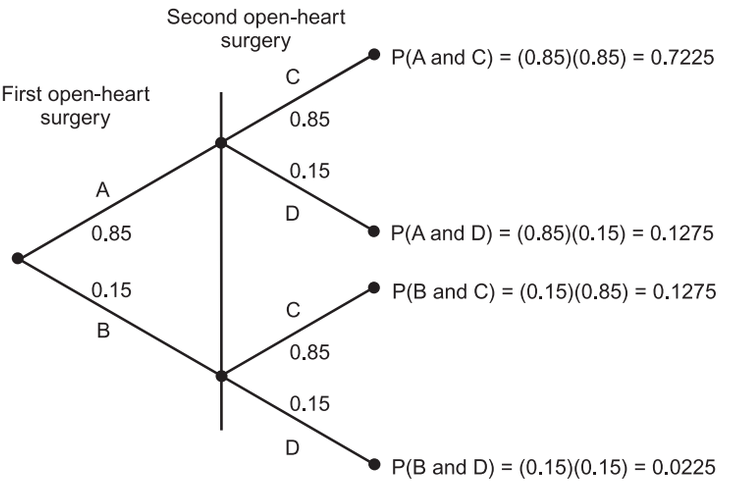


Fig. E2.21

Example E2.22

(a) In a class of 35 students, 11 are seniors, 8 are juniors, 9 are sophomores, and 7 are freshmen. If one student is selected at random from this class, what is the probability that this student is

(i) a junior

(ii) a freshman

(b) If two students are selected at random from this class of 35 students, find the probability that the first student selected is a junior and the second is a sophomore.

SOLUTION:

(a) *(i)* P(student selected is a junior) $= \dfrac{8}{35} = 0.22857$

(ii) P(student selected is a freshman) $= \dfrac{7}{35} = 0.2$

(b) Let J_1 = first student selected is a junior

S_2 = second student selected is a sophomore

$$P(J_1 \text{ and } S_2) = P(J_1)\, P(S_2|J_1) = \left(\frac{8}{35}\right)\left(\frac{9}{35}\right) = 0.05878$$

Example E2.23

A manufacturing company makes automobile steering wheels. The manufacturing system involves two independent processing machines so that each steering wheel passes through these two processes. The probability that the first processing machine is not working properly at any time is 0.09, and the probability that the second machine is not working properly at any time is 0.065. Find the probability that both machines will not be working properly at any given time.

SOLUTION:

P(both machines not working properly)
= P(first machine is not working properly) P(second machine not working properly)
= (0.09)(0.065) = 0.00585

2.15 BAYE'S FORMULA

Baye's formula is quite useful in modifying a probability estimate as additional information becomes available. We have seen that a conditional probability is one in which the probability of the event depends upon whether the other event has occurred.

$$P(A/B) = \frac{P(AB)}{P(B)} \tag{2.28}$$

or

$$P(B/A) = \frac{P(AB)}{P(A)} \tag{2.29}$$

From Eqs. (2.28) and (2.29), we have

$$P(AB) = P(A/B)\, P(B) = P(B/A)\, P(A) \tag{2.30}$$

since $AB = BA$

and

$$P(A/B) = \frac{P(A/B)\, P(A)}{P(B)} \tag{2.31}$$

Since $P(A) + P(\bar{A}) = 1$, it follows that event B occurs jointly with either A or $\bar{A}$.

$$P(B) = P(AB) + P(\bar{A}\, B)$$

and from Eq. (2.30)

$$P(B) = P(A)\, P(B/A) + P(\bar{A})\, P(B/\bar{A}) \tag{2.32}$$

Substituting Eq. (2.32) into Eq. (2.31), we get

$$P(A/B) = \frac{P(A)\, P(B/A)}{P(A)\, P(B/A) \;+\; P(\bar{A})\, P(B/\bar{A})} \tag{2.33}$$

If event A has more than two available alternatives, we can write Eq. (2.33) as

$$P(A/B) = \frac{P(A_i)\, P(B/A_i)}{\sum_i P(A_i)\, P(B/A_i)} \tag{2.34}$$

Example E2.24

Find the joint probability of A and B for the following:

(a) $P(B) = 0.60$ and $P(A/B) = 0.70$

(b) $P(A) = 0.2$ and $P(B/A) = 0.15$

SOLUTION:

(a) $P(A \text{ and } B) = P(B \text{ and } A) = P(B)\, P(A/B) = (0.60)(0.70) = 0.42$

(b) $P(A \text{ and } B) = P(A)\, P(B/A) = (0.2)(0.15) = 0.03$

Example E2.25

A certain manufacturing company has 456 employees. Based on the information obtained by asking all the employees whether they are smokers or non-smokers and whether or not they are married, the following two-way classification table of data was obtained.

Smoker/Non-smoker	Married	Not married	Total
Smoker	37	72	109
Non-smoker	121	226	347
Total	158	298	456

(a) If one employee is selected at random from this company, find the probabilities of the following:

(i) P(Married and non-smoker)

(ii) P(Smoker and not married)

(b) Draw a tree diagram and compute all the joint probabilities for the given data.

SOLUTION:

(a) Let M = Married

NM = Not married

S = Smoker

NS = Non-smoker

(i) $P(M \text{ and } NS) = P(M)\, P(NS/M) = \left(\frac{158}{456}\right)\left(\frac{121}{158}\right) = 0.2653$

(ii) $P(S \text{ and } NM) = P(S)\, P(NM/S) = \left(\frac{109}{456}\right)\left(\frac{72}{109}\right) = 0.1579$

(b) The tree diagram is shown in Fig. E2.25.

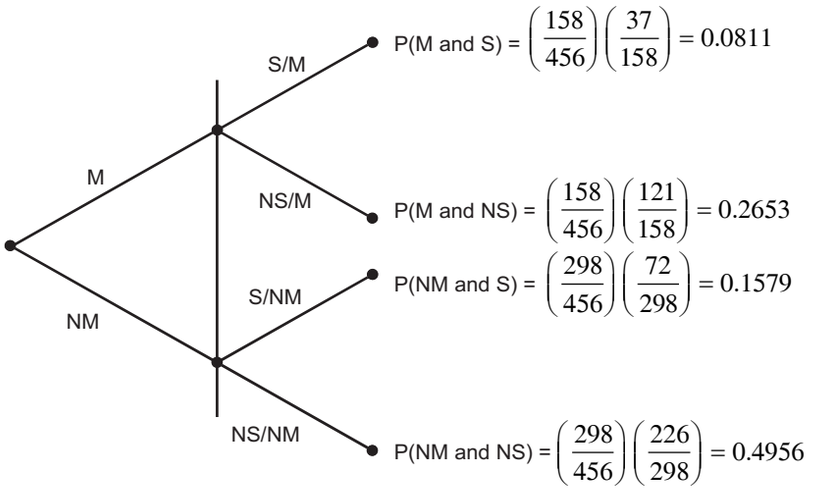


Fig. E2.25: Tree diagram

Example E2.26

Refer to the data of Example E2.25, if one employee is selected at random from the company, calculate the following probabilities.

(a) P(married or smoker)

(b) P(smoker or not married)

(c) P(smoker or non-smoker)

SOLUTION:

Referring to the solution of Example E2.25, we have

(a) $P(M \text{ or } S) = P(M) + P(S) - P(M \text{ and } S) = \left(\frac{158}{456}\right) + \left(\frac{109}{456}\right) - \left(\frac{37}{456}\right) = 0.5044$

(b) $P(S \text{ or } NM) = P(S) + P(NM) - P(S \text{ and } NM) = \left(\frac{109}{456}\right) + \left(\frac{298}{456}\right) - \left(\frac{72}{456}\right) = 0.7346$

(c) Since S and NS are mutually exclusive events,

$$P(S \text{ or } NS) = P(S) + P(NS) = \left(\frac{109}{456}\right) + \left(\frac{347}{456}\right) = 1.0$$

Example E2.27

The probability that a manufacturer is in a debt is 0.7. What is the probability that three randomly selected manufacturers are all in debt? Assume independence of events.

SOLUTION:

Let M_1 = first manufacturer in debt

M_2 = second manufacturer in debt

M_3 = third manufacturer in debt

Then, $P(M_1 \text{ and } M_2 \text{ and } M_3) = P(M_1)\,P(M_2)\,P(M_3) = (0.7)(0.7)(0.7) = 0.343$

Example E2.28

Find $P(A \text{ or } B)$ for the following:

(a) $P(A) = 0.17$, $P(B) = 0.5$ and $P(A \text{ and } B) = 0.12$

(b) $P(A) = 0.8$, $P(B) = 0.7$ and $P(A \text{ and } B) = 0.6$.

SOLUTION:

(a) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.17 + 0.5 - 0.12 = 0.55$

(b) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.8 + 0.7 - 0.6 = 0.9$

Example E2.29

The average acceptability of three automobile transmissions from three suppliers is given below:

Supplier	Relative probability of supply	Average acceptability
1	0.56	0.97
2	0.32	0.89
3	0.12	0.77

(a) Find the probability that any transmission received will perform acceptably

(b) Find the probability that a particular transmission delivered by supplier 2 when it is known to have performed acceptably.

SOLUTION:

(a) $P(A) = (0.56)(0.97) + (0.32)(0.89) + (0.12)(0.77) = 0.5432 + 0.2848 + 0.0924 = 0.9204$

(b) $$P(T_2 / A) = \frac{P(T_2)P(A/T_2)}{\Sigma P(T_1)P(A/T_1)} = \frac{(0.32)(0.89)}{(0.56)(0.97)+(0.32)(0.89)+(0.12)(0.77)} = \frac{0.2848}{0.9204} = 0.3094$$

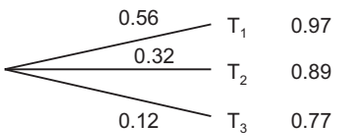


Fig. E2.29

2.16 ADDITIONAL EXAMPLES AND SOLUTIONS

Example E2.30

Following is a percentage distribution for the number of years of school completed by adult employees, 25 years old or over in a manufacturing company.

Years completed	Percentage	Event
0–4	2	A
5–7	4	B
8	5	C
9–11	11	D
12	40	E
13–15	20	F
16 or more	18	G

Determine the probability that an employee randomly selected

(a) has completed 8 years of school or less

(b) has completed 12 years of school or less

(c) has completed 13 years of school or more

(d) interpret each of your answers in *(a)*–*(c)* in terms of percentages.

SOLUTION:

(a) P(the employee has completed 8 years of school or less) $= P(A) + P(B) + P(C)$
$= 0.02 + 0.04 + 0.05 = 0.11$

(b) P(the employee has completed 12 years of school or less) $= P(A) + P(B) + P(C) + P(D) + P(E)$
$= 0.02 + 0.04 + 0.05 + 0.11 + 0.40 = 0.62$

(c) P(the employee has completed 13 years of school or more) $= P(F) + P(G)$
$= 0.20 + 0.18 = 0.38$

(d) The interpretation of each of the results above in terms of percentage is as follows:

(i) 11% of the adults aged 25 years old and over have completed 8 years of school or less

(ii) 62% of the adults aged 25 years old and over have completed 12 years of school or less

(iii) 38% of the adults aged 25 years old and over have completed 13 years of school or more.

Example E2.31

In a particular corporation, 56% of the employees are white, 95% are male, and 53% are white males. For a randomly selected employee, let

W = Event the employee selected is white

M = Event the employee selected male

(a) Find $P(W)$, $P(M)$ and $P(W \text{ and } M)$

(b) Determine $P(W \text{ or } M)$ and express the answer in percentages

(c) Find the probability that a randomly selected employee is female.

SOLUTION:

(a) $P(W) = 0.56$
$P(M) = 0.95$
$P(W \text{ or } M) = 0.53$

(b) $P(W \text{ or } M) = 0.56 + 0.95 - 0.53 = 0.98$
98% of the employees are either white or male.

(c) $P(F) = 1 - 0.95 = 0.05$

Example E2.32

The number of two door cars and four door vehicles in use by the employees of a company by age are as shown in the following Table E2.32.

Table E2.32

Age (years)		Type		Total
		2 Door T	4 Door F	
Under 6	A_1	25	45	70
6–8	A_2	15	5	20
9–11	A_3	20	10	30
12 and over	A_4	13	7	20
Total		73	67	140

For a randomly selected vehicle (2 Door or 4 Door), determine the probability that the vehicle selected

(a) is under 6 years old

(b) is under 6 years old, given that it is a 2 Door

(c) is a 2 Door

(d) is a 2 Door; given that it is under 6 years old

(e) interpret your answers in *(a)* to *(c)* in terms of percentages.

SOLUTION:

The probability that the vehicle selected is

(a) Under 6 years old is 70/140 = 0.5

(b) Under 6 years old, given that it is a 2 Door car is 25/73 = 0.343

(c) A 2 door car is 73/140 = 0.521

(d) A 2 door car; given that it is under 6 years old is 25/70 = 0.357

(e) *(i)* 50% of all vehicles are under 6 years old

(ii) 34.3% of all 2 door cars are under 6 years old

(iii) 52.1% of all vehicles are 2 door

(iv) 35.7% of all vehicles under 6 years old are 2 door cars.

Example E2.33

Table E2.33 presents a joint probability distribution for engineers and scientists by highest degree obtained in a particular *R&D* corporation.

Table E2.33

Highest degree	Type		
	Engineer, T_1	Scientist, T_2	$P(D_1)$
Bachelors, D_1	0.35	0.29	0.64
Masters, D_2	0.10	0.2	0.30
Doctorate, D_3	0.02	0.025	0.045
Other, D_4	0.01	0.005	0.015
$P(T_1)$	0.48	0.52	1.0

Determine the probability that the person selected

(a) is an engineer
(b) has a doctorate
(c) is an engineer with a doctorate
(d) is an engineer given the person has a doctorate
(e) has a doctorate, given the person is an engineer
(f) interpret your answers in *(a)–(e)* in terms of percentages.

SOLUTION:

The probability that the person selected

(a) is an engineer is 0.48.
(b) has a doctorate is 0.045.
(c) is an engineer with a doctorate is 0.02.
(d) is an engineer given the person has a doctorate is 0.02/0.045 = 0.444.
(e) has a doctorate, given that the person is an engineer is 0.02/0.48 = 0.0417.
(f) *(i)* 48% of all engineers and scientists are engineers.
(ii) 4.5% of all engineers and scientists have doctorates.
(iii) 2% of all engineers and scientists are engineers with a doctorate.
(iv) 44.4% of all engineers and scientists with doctorates are engineers.
(v) 4.17% of all engineers have doctorates.

Example E2.34

A frequency distribution for the class level of students in "Introduction to JAVA programming" course is as given in Table E2.34.

Table E2.34

Class	Frequency
Freshman	5
Sophomore	14
Junior	11
Senior	6
Total	36

Two students are randomly selected without replacement. Find the probability that

(a) the first student selected is a junior and the second a senior
(b) both students selected are sophomores
(c) draw a tree diagram for the solution of all the probabilities
(d) what is the probability that one of the students selected is a freshman and the other a sophomore?

SOLUTION:

Let F_i = Freshman on selection i, where $i = 1, 2$
S_i = Sophomore on selection i, where $i = 1, 2$

J_i = Junior on selection i, where $i = 1, 2$

H_i = Senior on selection i, where $i = 1, 2$

(a) $P(J_1 \text{ and } H_2) = (11/36)/(6/35) = 0.0524$

(b) $P(S_1 \text{ and } S_2) = (14/36)/(13/35) = 0.1444$

(c) The tree diagram is shown in Fig. E2.34.

(d) The probability that one of the students selected is a freshman and the other student selected is a sophomore is:

$$P(F_1 \text{ and } S_2) + P(S_1 \text{ and } F_2) = 0.0555 + 0.0555 = 0.111$$

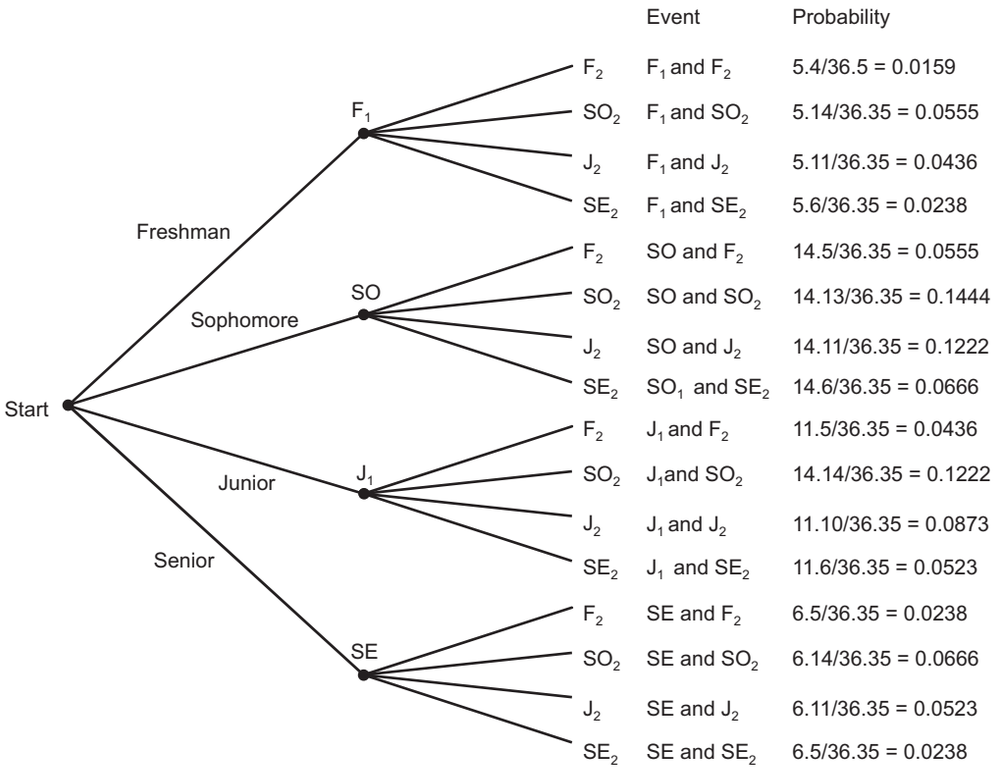


Fig. E2.34: Tree diagram

Example E2.35

Refer to Example E2.33

(a) determine $P(T_2)$, $P(D_3)$ and $P(T_2 \text{ and } D_3)$

(b) are T_2 and D_3 independent events? Why?

SOLUTION:

(a) $P(T_2) = 0.52$

$P(D_3) = 0.045$

$P(T_2 \text{ and } D_3) = 0.025$

(b) The special multiplication rule states that two events T_2 and D_3 are independent if,

$P(T_2 \text{ and } D_3) = P(T_2) \cdot P(D_3)$

Using the data in *(a)*

$P(T_2) \cdot P(D_3) = (0.52)(0.045) = 0.0234$

Since this product does not equal $P(T_2 \text{ and } D_3) = 0.025$, the events T_2 and D_3 are not independent.

Example E2.36

The two-way classification of data of 2000 randomly selected employees of a manufacturing company from a city based on gender and commuting time from residence to work place is given in Table E2.36.

Table E2.36

Gender	Commuting time from residence to work place			Total
	Less than 30 min	30 min to 1 hour	More than 1 hour	
Men	525	450	225	1200
Women	410	250	140	800
Total	935	700	365	2000

(a) If one employee is selected at random from these 2000 employees, find the probability that this employee

(i) commutes from more than 1 hour

(ii) commute for less than 30 minutes

(iii) is a man commuting for 30 min to 1 hour

(iv) is a woman commuting for more than 1 hour.

(b) Determine if the following events are mutually exclusive

(i) 'men' and 'commuting for more than 1 hour'

(ii) 'less than 30 min' 'more than 1 hour'

(iii) 'woman' and 'commutes for 30 minutes to 1 hour'. Are these events independent?

SOLUTION:

(a) *(i)* $P(\text{commutes for more than 1 hour}) = \dfrac{365}{2000} = 0.1825$

(ii) $P(\text{commutes for less than 30 min.}) = \dfrac{935}{2000} = 0.4675$

(iii) $P(\text{commutes for 30 min to 1 hour}) = \dfrac{450}{700} = 0.6428$

(iv) $P(\text{commutes for more than 1 hour/woman}) = \dfrac{140}{800} = 0.175$

(b) *(i)* The term 'man' and 'commutes for more than 1 hour' are not mutually exclusive. The two events can occur together.

(ii) The events 'less than 30 minutes' and 'more than 1 hour' are mutually exclusive because they cannot occur together.

(iii) $P(\text{woman}) = \dfrac{800}{2000} = 0.4$

$$P(\text{woman/commutes for 30 min to 1 hour}) = \frac{250}{700} = 0.3571$$

Since these two probabilities are not equal, the events 'woman' and 'commute for 30 minutes to 1 hour' are not independent.

2.17 SUMMARY

The term probability is used to describe the uncertainty *probability*, which measures the likelihood that an event will occur, is an important part of statistics. Probability theory is used extensively in analysing decision-making situations that involve risk or incomplete information.

PROBLEMS

P2.1 In Fig. P2.1, *E*, *T*, and *N* are the events that an automobile brought to a service garage needs an engine oil-change, tune-up, or new tires. Express in words the events represented by

(a) region 1

(b) region 3

(c) region 7

(d) regions 1 and 4 together

(e) regions 2 and 5 together

(f) regions 3, 5, 6, and 8 together.

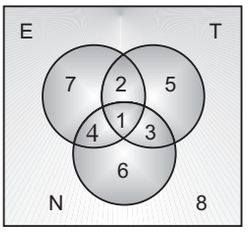


Fig. P2.1: Venn diagram

P2.2 In Problem 2.1, list the region or combinations of regions representing the events that an automobile brought to the garage needs

(a) tune-up, but neither engine oil-change nor new tires

(b) an engine oil-change and tune-up

(c) tune-up or new tires, but not an engine oil-change

(d) new tires.

P2.3 An electronic security system may be used in any one of the three modes, in any two of the three modes, or in all three modes: manual (*M*), semi-automatic (*S*) and automatic (*A*). In tests on 200 such systems, the following uses were found.

Number	Use
30	M∩S∩A
40	M∩S
20	A-only
40	S∩A
All remaining	M-only

(a) define M, S, and A as sets of systems used. Draw a Venn diagram, and show the number of systems in each use

(b) how many systems are in M? In S? In $(M \cup S)$?

P2.4 An age distribution of 100 employees at a manufacturing company are given as follows:

Age (years)	Number of employees
Under 40	1
40–49	14
50–59	41
60–69	27
70 and over	17
Total	100

For an employee selected at random, let

A : event the employee is under 40

B : event the employees is in his or here 40's

C : event the employee is in his or her 50's

S : event the employee is under 60

(a) use the tables and the *f*/*n* rule to find $P(S)$

(b) express event S in terms of events A, B and C

(c) determine $P(A)$, $P(B)$ and $P(C)$

(d) complete $P(S)$ using special addition rule and write the answers from parts (b) and (c). Also compare the answer in part (a).

P2.5 Refer to Problem P2.4. Find the probability that a randomly selected employee is

(a) 40 years old or under

(b) under 60 years old.

P2.6 In a particular company, 50% of the adult employees are female, 10% are divorced, and 5% are divorced females. For an employee selected at random, let

F = event the person is female

D = event the person is divorced.

(a) find $P(F)$, $P(D)$ and $P(F \text{ and } D)$

(b) determine $P(F \text{ or } D)$ and interpret your answer in terms of percentages

(c) find the probability that a randomly selected adult is male.

P2.7 A fair coin is tossed five times. Find the probability of getting at least one tail.

P2.8 Suppose we randomly select two students from a typical co-ed class and observe whether the student selected each time is a male or a female. Write all the outcomes for this experiment. Draw the Venn and tree diagrams for the experiment.

P2.9 In a group of freshman engineering class, some are in favour of mechanical engineering and others are against it. Two students are selected at random from this class and asked whether they are in favour of or against mechanical engineering. How many distinct outcomes are possible? Draw a Venn diagram and a tree diagram for this experiment. List all the outcomes included in each of the following events and state whether they are simple or compound events.

(a) both students are in favour of mechanical engineering

(b) at most one student is against mechanical engineering

(c) exactly one student in favour of mechanical engineering.

P2.10 Determine the conditional probability $P(\text{in favour/male})$ for the response data in Table P2.10 for the 100 employees. They were asked whether they are in favour or against paying higher salaries to CEOs of corporations.

Table P2.10

	In favour	Against	Total
Male	20	40	60
Female	4	36	40
Total	24	76	100

Draw the tree diagram to illustrate the computations.

P2.11 Refer to Problem P2.10. Find the conditional probability that a randomly selected employee is a female given that this employee is in favour of paying high salaries to the CEOs. Draw the tree diagram for this computation.

P2.12 *(a)* Determine the probability of getting 3 tails in tossing 3 coins.

(b) Two dice are thrown. Draw the sample space. Determine the probability of getting 6 in both the dice.

(c) Determine the probability of getting *(i)* exactly two tails (event A) and *(ii)* at least two tails (event B) in tossing 3 coins.

P2.13 *(a)* Determine the probability that the sum of the numbers shown on the two faces, when two dice are thrown

(i) is 7

(ii) is 10

(b) A die is thrown. Determine the probability of getting either an even number or a number greater than 4 or both.

P2.14 Table P2.14 gives the contingency table for age and rank of employees in a particular company. If an employee is selected at random,

(a) find the unconditional probability that the employee selected is in his/her 50's

(b) find the unconditional probability that the employee selected is in his or her 50's given that Grade 1 employee is selected

(c) interpret the probabilities found in parts *(a)* and *(b)* in terms of percentages.

Table P2.14

		Rank				
		R_1	R_2	R_3	R_4	Total
Age	Under 30 A_1	5	8	57	7	77
	30–39 A_2	50	170	160	20	400
	40–49 A_3	150	120	60	6	336
	50–59 A_4	150	60	30	5	245
	60 & over A_5	70	20	5	0	95
	Total	425	378	312	77	1192

P2.15 Refer to Problem P2.14. Compute the conditional probability $P(A_4/R_3)$ using unconditional probabilities.

P2.16 Data on the marital status of employees in a company are as given in the table below. The table provides a joint probability distribution for the marital status of adults by sex. The term "Single" is used for "Never Married".

Table P2.16

		Marital status				
		Single M_1	Married M_2	Widowed M_3	Divorced M_4	$P(S_1)$
Sex	Male S_1	0.10	0.30	0.01	0.04	0.45
	Female S_2	0.15	0.35	0.03	0.02	0.55
	$P(M_1)$	0.25	0.65	0.04	0.06	1.00

If an employee is selected at random,

(a) find the probability that the employee selected is divorced, given that the employee selected is male, given that the employee selected is divorced.

(b) use the conditional probability rule for *(a)*.

P2.17 In the research and development department of a company, the number of male and female employees are shown in the frequency distribution as given below in a tabular form.

Sex	Frequency
Male	30
Female	20
Total	50

Two employees selected at random from this department. The first employee obtained is not returned to the department for possible reselection; that is, the sampling is without replacement.

(a) determine the probability that the first employee obtained is female and the second is male

(b) draw the tree diagram.

P2.18 A bin contains 20 machine parts, 4 of which are defective. If 2 machine parts are selected at random (without replacement) from this box, what is the probability that both parts are defective? Draw the tree diagram.

P2.19 Table P2.19 gives the classification of all employees of a particular department in a large corporation by gender and college degree.

Table P2.19

	College graduate (G)	Not a college graduate (N)	Total
Male (M)	8	20	28
Female (F)	3	9	12
Total	11	29	40

If one of these employees is selected at random for membership on the planning committee, determine the probability that this employee is a female and a college graduate. Draw the tree diagram.

P2.20 A balanced coin is tossed five times. Determine the probability of getting at least one head.

P2.21 The probability that a patient is allergic to penicillin is 0.10. If the drug is administered to three individual patients, determine the probability that

(a) all three of them are allergic to it

(b) at least one of them is not allergic to it.

(c) draw the tree diagram.

P2.22 Draw the Venn and tree diagram for selecting three times two persons (a man or a woman) from the members of a club.

P2.23 The search committee of the school of engineering that plans to hire one new faculty member has prepared a final of five candidates, all of whom are equally qualified. Out of these five candidates two of them are woman. If the committee decides to select at random one person out of these five candidates, find the probability that this person will be a woman or a man? Also find the sum of these two probabilities and comment the result.

P2.24 In a random sample of 1000 companies surveyed showed that 35 of them have no design engineer, 210 have one design engineer each, 379 have two design engineers, 270 have three design engineers and 106 have four or more design engineers.

(a) prepare a frequency distribution table and calculate the relative frequencies for all categories

(b) if one company is selected randomly from these 1000 companies, find the probability that this company has to design engineers

(c) find the probability that the company selected has four or more design engineers.

P2.25 Of the 25 new graduates hired by a company, 14 are female and 9 are mechanical engineering majors. Of the 14 females, 6 are mechanical engineering majors.

(a) are these events "female" and "mechanical engineering major" independent?

(b) are these events "mutually exclusive"?

P2.26 An individual is picked at random from a group of 56 athletes. Suppose 26 of the athletes are female and 6 of them are swimmers. Also, there are 10 swimmers among the males.

(a) find the probability that she is a swimmer if that the individual picked is a female

(b) given that the individual picked is a swimmer, find the probability that the person is a male.

P2.27 *(a)* The probability that the stock market goes up on Monday is 0.7. Given that it goes up on Monday, the probability that it goes up on Tuesday is 0.3. Find the probability that the market goes up on both days.

(b) The probability that a student passes the midterm exam is 0.7. Given that a student fails the midterm exam, the probability that the student passes the final exam is 0.8. Find the probability that John (a student in that class) fails the midterm and passes the final.

(c) Past records show that the probability that the Chairman of the board attends a meeting is 0.7, that the President of the company attends a meeting is 0.8, and that they both attend a meeting is 0.6. Can we say that the Chairman and the President act independently regarding their attendance at the meeting of the board of directors?

P2.28 Two students are selected at random from a class without replacement whose frequency distribution of males and females is given in Table P2.28.

Table P2.28

Gender	Frequency
Male	17
Female	23
Total	40

(a) find the probability that the first student selected is female and the second is male

(b) draw a tree diagram.

P2.29 Given that A and B are two independent events, find their joint probability for the following data:

(a) $P(A) = 0.25$ and $P(B) = 0.78$

(b) $P(A) = 0.55$ and $P(B) = 0.23$

P2.30 The following data gives a two-way classification of all faculty members of a university based on gender and tenure.

	Tenured	Non-tenured
Male	75	48
Female	34	19

(a) if one of these faculty members is selected at random, find the following probabilities:

(i) P(male and non-tenured)

(ii) P(tenured and female)

(b) find P(tenured and non-tenured).

P2.31 In a group of 12 employees of a company, 5 are graduates and 7 are non-graduates. If 2 persons are selected at random from this group, find the probability that the first of them is a graduate and the second is a non-graduate. Draw a tree diagram for the solution.

P2.32 The probability that an employee at a university is a female is 0.3. The probability that an employee is a female and married is 0.2. Find the conditional probability that a randomly selected female employee from this university is a married person.

P2.33 For the data given in Problem P2.30, if one of these faculty members is selected at random, find the following probabilities.

(a) *P*(female or non-tenured)

(b) *P*(tenured or male).

P2.34 The probability that a new product introduced to the market is successful is 0.78. What is the probability that in two randomly selected new products introduced, at least one will be successful? Draw a tree diagram for the solution.

P2.35 Consider the following events for one roll of a die

A = an odd number is observed (1, 3, 5)

B = an even number is observed (2, 4, 6)

C = a number less than 5 is observed (1, 2, 3, 4)

(a) Are events A and B mutually exclusive?

(b) Are events B and C mutually exclusive?

P2.36 An experimental submarine is estimated to have a probability of 0.90 for a successful first flight. In case of failure there is a 0.03 probability of a catastrophic explosion, in which case the abort system cannot be used. The abort system has a reliability of 0.97. Calculate the probability of every possible outcome of the initial flight.

P2.37 With voltage fluctuations present, the probability of an electromechanical system going down is 75%; the probability of it going down during no voltage fluctuations condition is only 8%. The probability of voltage fluctuations occurring is 20%. What is the probability that voltage fluctuations are present given that the electro mechanical system is down?

P2.38 Out of a collection of 100 gaskets the following are defective as indicated:

Defective type	Number of gaskets
Type I defect	10
Type II defect	8
Type III defect	3
Type I and II defect	5
Type I, II and III defect	4
No defect	70

(a) if a gasket picked out the collection has a type II defect, what is the probability that it also has a type I defect?

(b) if the gasket picked has a type I defect, what is the probability that it also has a type II defect?

P2.39 Three persons: A, B and C are working independently to solve a mathematical puzzle. Based on their previous performance, the probabilities that they will succeed are 0.1, 0.2 and 0.3 respectively. What is the probability that the puzzle will be solved?

P2.40 A system consists of 25 different components, of which one; component A, has been identified as the critical component. The probability that the critical component is non-defective is 0.97. If the critical component is non-defective, then the system succeeds with at probability of 0.99. If not, then the probability of the system success drops to 0.93. What is the probability of system success?

P2.41 An electronic test set up has 97% probability of correctly identifying a defective component and a 3% probability of identifying a non-defective component as defective. A batch of 100 components of which 3 are known as defective is subjected to this electronic testing. If the test identifies a component as defective, what is the probability that it is truly defective?

P2.42 A mechanical system assembly plant purchases a certain component from three independent vendors: A, B and C, who supply 20%, 30% and 50% respectively of the total number of components needed by the assembly plant. On an average, the percentage of defective components supplied by vendors A, B and C are 5%, 3% and 2% respectively.

(a) if a component is selected at random, find the probability that it is defective

(b) if the selected component is defective, what is the probability that it was supplied by vendor C? by vendor A? by vendor B?

P2.43 A bowl contains five white balls, two red balls, and three green balls. What is the probability of getting either a white ball or a red ball in one draw from the bowl?

P2.44 One bowl contains five white balls, two red balls, and three green balls. Another bowl contains three yellow balls and seven black balls. Determine the probability of getting a red ball from the first bowl and a yellow ball from the second bowl in one draw from each bowl.

P2.45 The probability that a student is in favour of engineering ethics course is 0.60 and a student is against is 0.40. Two students are randomly selected, and it is observed whether they favour or oppose engineering ethics course.

(a) draw a tree diagram for this experiment

(b) determine the probability that at least one of the two students favours engineering ethics course.

P2.46 Customers who buy a certain make of an automobile can order an engine in any of three sizes. Of all automobiles sold, 50% have the smallest engine, 30% have the medium-size one, and 20% have the largest. Of automobiles with the smallest engine, 15% fail the emissions test within two years of purchase, while 10% of those with the medium size and 15% of those with the largest engine fail. Determine the probability that a randomly chosen automobile will fail on emissions test within two years.

P2.47 A company manufactures aluminium soft drink cans. It was found that the probability that a can has a flaw on its side is 0.03, the probability that a can has a flaw on the top is 0.02, and the probability that a can has a flaw on both sides and the top is 0.01.

(a) what is the probability that a randomly chosen can has a flaw?

(b) what is the probability that it has no flaw?

(c) what is the probability that a can will have a flaw on the side, given that it has a flaw on top?

(d) what is the probability that a can will have a flaw on the top, given that it has a flaw on the side?

P2.48 *(a)* Two coins are tossed. Determine the conditional probability of getting two heads (event B) given that at least one coin shows a head (event A)

(b) A box contains 5 red and 4 blue balls. Two balls are drawn one-by-one without replacement. Given that the first ball drawn is red, find the probability that both the balls drawn will be blue.

P2.49 Of the gas turbine engine blades manufactured by a certain process, 15% are defective. Five turbine blades are chosen at random. Assume they function independently. Find the probability that they all work.

P2.50 Refer Problem P2.49. What is the probability that at least one of the gas turbine engine blade manufactured works?

P2.51 In a certain city, 45% of the people consider themselves conservatives (C), 30% Consider themselves to be Liberals (L) and 25% consider themselves to be Independents (I). During the particular election, 50% of the Conservatives voted, 40% of the Liberals voted, and 60% of the Independents voted. If a randomly selected person voted, determine the probability that the voter is

(a) Conservative

(b) Liberal

(c) Independent.

P2.52 In a certain company employees, 5% of the men and 1% of the women are heavier than 100 kg. Furthermore, 70% of the employees are women. Suppose a randomly selected employee is heavier than 100 kg. Find the probability that the employee is a woman.

P2.53 In a certain engineering college hostel, 20% are freshmen of whom 10% own a car, 30% are sophomores of whom 20% own a car, 20% are juniors of whom 30% own a car, and 20% are seniors of whom 40% own a car. A student is randomly selected from the hostel.

(a) find the probability that the student owns a car

(b) if the student owns a car, find the probability that the student is a junior.

P2.54 At a certain grocery store, eggs come in cartons that hold a dozen eggs. The store experience indicates that 78% of the cartons have no broken eggs, 19% have one broken egg, 2% have two broken eggs, and 1% have three broken eggs, and that the percentage of cartons with four or more eggs broken eggs is negligible. An egg selected at random from a carton is found to be broken. What is the probability that this egg is the only broken one in the carton?

P2.55 The probability that a college student being male is 1/3 and that being female is 2/3. The probability that a male student completes the probability and statistics course successfully is 8/10 and that the female student does it is 4/5. A student is selected at random is found to have completed the course. Determine the probability that the student is a

(a) male

(b) female.

P2.56 In a car dealership, assume 55% of the cars are made in US and 15% of these are compact; 25% of the cars are made in Europe and 40% of these are compact; and finally, 20% are made in Japan and 60% of these are compact. If a car is picked at random from this dealership:

(a) find the probability that it is a compact

(b) draw a tree diagram for part *(a)*

(c) given that the car is a compact, find the probability that it is a European.

P2.57 In a survey of the evaluation of preliminary product designs by the customers, 90% of highly successful products received good reviews, 60% moderately successful products received good reviews, and 10% of poor products received good reviews. In addition, 40% of products have been highly successful, 35% have been moderately successful, and 25% have been poor products.

(a) find the probability that a product attains a good review

(b) if the new design attains a good review, what is the probability that it will be a highly successful product?

(c) if a product does not attain a good review, what is the probability that it will be a highly successful product?

P2.58 A cat is in a room and each of the four walls of the room has a door through which the cat could try to escape. However, there is a trap at each of the doors D_1, D_2, D_3 and D_4 and they work with probabilities 0.3, 0.2, 0.4 and 0.5 respectively. If the cat picks a door at random:

(a) find the probability that the cat will escape

(b) draw the tree diagram for the solution in part *(a)*

(c) given that the cat escapes, find the probability that the cat chose door D_3 to escape.

P2.59 According to a particular survey, 8% of the population has a lung disease. Of those having lung disease, 90% are smokers; of those not having lung disease, 25% are smokers. Find the probability that a randomly selected smoker has lung disease.

P2.60 Three different machines m_1, m_2 and m_3 were used for producing a large batch of similar manufactured items. Suppose that 30% of the items were produced by machine m_1, 30% by machine m_2, and 40% by machine m_3. Also, 1% of the items produced by machine m_1 are defective, 2% of the items produced by machine m_2 are defective, and 3% of the items produced by machine m_3 are defective. If one item is selected at random from the entire batch, and it is found to be defective. Determine the probability that this item was produced machine m_2.

REVIEW QUESTIONS

1. Define the following terms:
   (a) experiment
   (b) outcome
   (c) sample space
   (d) simple event
   (e) compound event.
2. Describe the properties of probability.
3. Describe an impossible event and a sure event. What is the probability of the occurrence of each of these two events?
4. Describe the three approaches to probability.
5. Explain the difference between the marginal and conditional probability of events.
6. What is meant by two mutually exclusive events?
7. Explain the meaning of independent and dependent events.
8. What is the complement of an event? What is the sum of the probabilities of two complementary events?
9. Explain the meaning of the following terms:
   (a) intersection of two events
   (b) the joint probability of two or more events
   (c) multiplication rule of probability
   (d) joint probability of two mutually exclusive events.
10. Explain the meaning of the following:
    (a) union of two events
    (b) addition rule of probability
    (c) classical probability rule
    (d) equally likely outcomes
    (e) sample point.

STATE TRUE OR FALSE

1. The probability of an event is always between 0 and 1, inclusive. (True/False)
2. The probability of an event that cannot occur is 1. (True/False)
3. The probability of an event that must occur is 1. (True/False)
4. An event that cannot occur is called a certain event. (True/False)
5. An event that must occur is called a certain event. (True/False)
6. An experiment is an action whose outcome cannot be predicted with certainty. (True/False)
7. An event is some specified result that may or may not occur when an experiment is performed. (True/False)
8. The experiment has a finite number of possible outcomes, all equally likely. (True/False)
9. The probability of an event equals the ratio of the number of ways that the event can occur to the total number possible outcomes. (True/False)
10. If a member is selected at random from a finite population, probabilities are identical to percentages (relative frequencies). (True/False)
11. The probability of an event is the proportion of times it occurs in a small number of repetitions of the experiment. (True/False)
12. Sample space is the collection of all possible outcomes for an experiment. (True/False)
13. An event is a collection of outcomes for the experiment, that is, any subset of the sample space. (True/False)
14. Two or more events are mutually exclusive events if no three of them have outcomes in common. (True/False)
15. If event A and event B are mutually exclusive, so are events A, B and C for every event C. (True/False)
16. If event A and event B are not mutually exclusive, neither are events A, B and C for every event C. (True/False)
17. If E is an event, then $P(E)$ represents the probability that event E occurs. (True/False)
18. If event A and event B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$. (True/False)
19. For any event E, $P(E) = 1 - P(\text{not } E)$. (True/False)
20. If A and B are any two events, then $P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$. (True/False)
21. The probability that event B occurs given that event A occurs is called a conditional probability. (True/False)
22. If A and B are any two events with $P(A) > 0$, then $P(B|A) = P(A \,\&\, B)/P(B)$. (True/False)
23. If A and B are any two events, then $P(A \,\&\, B) = P(A)/P(B|A)$. (True/False)
24. Event B is said to be independent of event A if $P(B|A) = P(B)$. (True/False)
25. If A and B are independent events, then $P(A \,\&\, B) = P(A) \cdot P(B)$. (True/False)
26. Two or more events are said to be mutually exclusive if at most one of them can occur when the experiment is performed, that is, if no two of them have outcomes in common. (True/False)

27. For any two events, the probability that one or the other of the events occurs equals the sum of the two individual probabilities. (True/False)
28. For any event, the probability that it occurs equals 1 minus the probability that it does not occur. (True/False)
29. Data obtained by observing values of one variable of a population are called univariate data. (True/False)
30. Data obtained by observing values of two variables of a population are called bivariate data. (True/False)
31. The joint probability equals the product of the marginal probabilities. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **False** 5. **True** 6. **True** 7. **True** 8. **True** 9. **True** 10. **True**
11. **False** 12. **True** 13. **True** 14. **False** 15. **False** 16. **True** 17. **True** 18. **True** 19. **True** 20. **True**
21. **True** 22. **False** 23. **False** 24. **True** 25. **True** 26. **True** 27. **False** 28. **True** 29. **True** 30. **True**
31. **True**

CHAPTER 3

Random Variables and Probability Distributions

The topic of random variables is fundamental to probability and statistics. These concepts are natural extensions of the ideas of variables and relative frequency distributions.

This chapter introduces the basic concepts and definitions of random variables (both discrete and continuous) and their mean and standard deviation, and probability distributions. Permutations and combinations are defined briefly. Discrete probability distributions (hypergeometric, binomial and Poisson distributions) and their mean and standard deviations are derived. Continuous probability distribution, namely, the normal distribution, its properties, mean and variance and the standard normal distribution are presented. Approximating probability is also presented. They include Binomial approximation to the hypergeometric, Poisson approximate to the binomial and normal approximation to the binomial and rule for distribution are presented.

3.1 RANDOM VARIABLES

This section introduces the important concept of a probability distribution, which gives the probability for each value of a variable that is determined by chance. This section also introduces procedures for finding the mean and standard deviation for a probability distribution. A *random variable* is a variable (typically represented by x) that has a single numerical value, determined by chance, for each outcome of a procedure. In other words, a *random variable* is a quantitative variable whose value depends on chance. In Chapter 1, we defined a variable as a characteristic that varies from one member of a population to another. When one or more members are selected at random from the population, the variable, in that context, is called a *random variable*. Therefore, mathematically, a random variable is a function defined on the outcome of the sample space.

A *probability distribution* is a description that gives the probability for each value of the random variable. It is often expressed in the format of a graph, table, or formula. A random variable is classified a discrete or continuous, depending upon the range of its values. (The range of a random variable is the set of values it can assume.)

A *discrete random variable* has either a finite number of values or a countable number of values, where *countable* refers to the fact that there might be infinitely many values, but they can be associated

with a counting process. Thus, we often say that a discrete random variable can assume at most a *countable infinite* number of values.

The number of eggs that a hen lays in a day, the salary of a new college graduate, the number of bacteria in a culture, the count of the number of students present in statistics class on a given day, the number of cavities that a child has, the price of gold on the exchange on a particular day, toss of a coin, the number of cars sold at a car dealership in a given month, color of a ball drawn from a collection of balls, etc., these are all examples of discrete random variables.

A *continuous random variable* has infinitely many values, and those values can be associated with measurements on a continuous scale without gaps or interruptions.

The maximum daily temperature, the amount of milk a cow produces in one day, the life of an electric bulb, the measure of voltage for a particular smoke detector battery, the distance by which a sharp shooter misses a target, the height of a person, the weight of a person, and the length of time a person has to wait at a bank counter, diameter of a manufactured shaft, time to failure of a machine component, repair time, the price of an automobile, the time taken to complete a medical examination, duration of a snow storm, etc. provide examples of continuous random variables.

3.1.1 Discrete Random Variables

Recall from Chapter 1 that the relative-frequency distribution or relative-frequency histogram of a discrete variable gives the possible values of the variable and the proportion of times each value occurs. We can extend the actions of relative-frequency distribution and relative-frequency histogram-concepts applying to variables of finite populations – to any discrete random variable. Here, we use the terms *probability distribution* and *probability histogram.*

The probability distribution of a discrete random variable, assuming a finite number of values, can be described by listing all the values that the random variable can assume, together with the corresponding probabilities. Such a listing is called the *probability function* of the random variable.

Probability histogram is a graph of the probability distribution that displays the possible values of a discrete random variable on the horizontal axis and the probabilities of those values on the vertical axis. The probability of each value is represented by a vertical bar whose height equals the probability.

In general, if a random variable X assumes the values $x_1, x_2, \ldots, x_k$, then we can represent the probability that X takes the values x_i by p_i. That is

$$P(X = x_i) = p\,(x_i)$$

The probability function can then be summarised in the form of a table, as shown in Table 3.1. By definition, a probability function gives the probabilities with which the values are assumed by the random variable. Hence, $0 \leq p\,(x_i) \leq 1$.

Table 3.1: The probability function of a random variable

Value x	Probability p(x)
x_1	$p(x_1)$
x_2	$p(x_2)$
⋮	⋮
x_k	$p(x_k)$
Sum	$\Sigma p(x_i) = 1$

Similarly, a random variable has to assume one of its possible values. Therefore, $\Sigma p(x_i) = 1$.

Properties of a Probability Function:

1. The probability that a random variable assumes a value x_i is always between 0 and 1.
   Hence, $0 \le p(x_i) \le 1$ (3.1)
2. The sum of all probabilities $p(x_i)$ is equal to 1.
   Hence, $\Sigma p(x_i) = 1$ (3.2)

Let X be a random variable. If the number of possible values of X is finite or countably infinite, X is called a *discrete random variable*. Each possible outcome is x_i. A number, $p(x_i) = p(X = x_i)$, gives the probability that the random variable X equals a value, x_i. The numbers, $p(x_i)$, for $i = 1, 2, \ldots$, must satisfy the following conditions:

1. $p(x_i) \ge 0$ for all i
2. $\sum_{i=1}^{\infty} p(x_i) = 1$ (3.3)
3. $0 \le p(x_i) \le 1$

The collection of pairs $[x_i, p(x_i)]$, $i = 1, 2, \ldots$ is called the *probability distribution* of X and $p(x_i)$ is called the *probability mass function* (or pmf) of X. The probability distribution for the variable X is characterised in Fig. 3.1. The *cumulative distribution function* of a discrete random variable is defined as

$$p(X = x_i) = \sum_{x_i \le X} p(x_i) \tag{3.4}$$

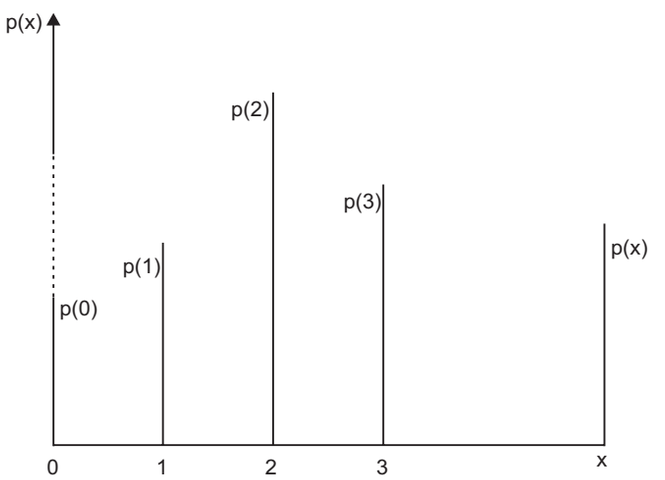


Fig. 3.1: Probability distribution of a discrete random variable x_i

The function is a step function that is constant over every interval not containing any of the points x_i, as shown in Fig. 3.2.

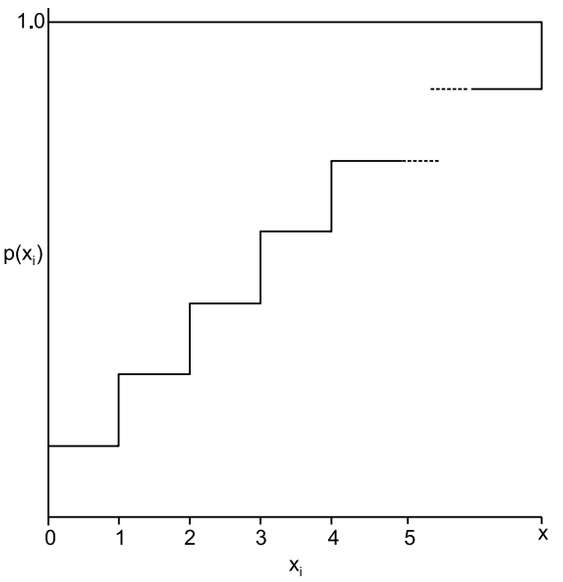


Fig. 3.2: Cumulative distribution function (CDF) of a discrete random variable x_i

3.1.2 Mean and Standard Deviation of a Discrete Random Variable

Consider a set of values, say $[x_i]$, $i = 1, 2, \ldots, n$, each of which occurs with frequency f_i for $i = 1, 2, \ldots, k$. The mean value is the sum of all the x's, hence $\sum_{i=1}^{k} x_i \cdot f_i$ divided by the total number of x's, $\sum_{i=1}^{k} f_i = n$, or

$$\frac{\sum_{i=1}^{k} x_i . f_i}{\sum_{i=1}^{k} f_i} = \sum_{i=1}^{k} \frac{x_i . f_i}{n} \tag{3.5}$$

But note that this could be written as

$$\frac{\sum_{i=1}^{k} x_i . f_i}{n} = \sum_{i=1}^{k} x_i . \frac{f_i}{n} \tag{3.6}$$

The relative frequencies f_i/n could be regarded as probabilities or values of $f(x)$. Hence,

If a random variable X has the discrete probability function $f(x)$, we define its *mean* or *expected value* or *expectation* as

$$E(X) = \mu = \Sigma x \cdot f(x) \text{ or } \Sigma x \cdot p(x) \tag{3.7}$$

where the summation extends over all the values of x and $p(x)$ are the probabilities.

The *variance* of a random variable X with probability distribution $f(x)$ is given by

$$\sigma^2 = \Sigma (x - \mu)^2 \cdot f(x) \tag{3.8}$$

where the summation extends over all the values of x.

The positive square root of σ^2, σ, is called the *standard deviation*.

We often denote the variance of X by $V(X)$.

It is often easier to use the following formula for σ^2:

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 \text{ or } \sum_x x^2 p(x) - \mu^2 \tag{3.9}$$

The variance of a random variable X is a measure of the dispersion or scatter in the possible values for X. The variance of X, denoted as σ^2 or $V(X)$ is

$$\sigma^2 = V(X) = \sum_x (x-\mu)^2 f(x) \tag{3.10}$$

$V(X)$ uses weight $f(x)$ as the multiplier of each possible squared deviation $(x - \mu)^2$.

Properties of summations and the definition of μ can be used to show that

$$V(X) = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x)$$

$$= \sum_x x^2 f(x) - 2\mu^2 + \mu^2 = \sum_x x^2 f(x) - \mu^2 \tag{3.11}$$

Therefore, an alternative formula for $V(X)$ can be used.

Summarising, we have

The *mean* or *expected value* of the discrete random variable X, denoted as μ or $E(x)$ is

$$\mu = E(X) = \sum_x x f(x) \tag{3.12}$$

The *variance* of X, denoted as σ^2 or $V(X)$ is

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2 \tag{3.13}$$

The *standard deviation* of X is

$$\sigma = \sqrt{V(X)} \tag{3.13a}$$

Example E3.1

The number of defects in a gas turbine blade manufactured by a certain process varies from blade to blade. Overall, 50% of the blades produced have no defects, 40% have 1 defect, 8% have 2 defects and 2% have 3 defects. Let x be the number of defects in a randomly selected turbine blade manufactured by the process.

(a) find $p(x = 0)$, $p(x = 1)$, $p(x = 2)$ and $p(x = 3)$

(b) plot the probability mass function of the random variable x

(c) find $F(2)$ and $F(1.5)$

(d) plot the cumulative distribution function $F(x)$ of the random variable x that represent the number of defects in a randomly chosen turbine blade

(e) population mean of the sample random variable, x

SOLUTION:

(a) $P(x = 0) = 0.50$, $P(x = 1) = 0.40$, $P(x = 2) = 0.08$ and $P(x = 3) = 0.02$

(b) The plot of the probability mass function of the random variable, x is shown in Fig. E3.1(a).

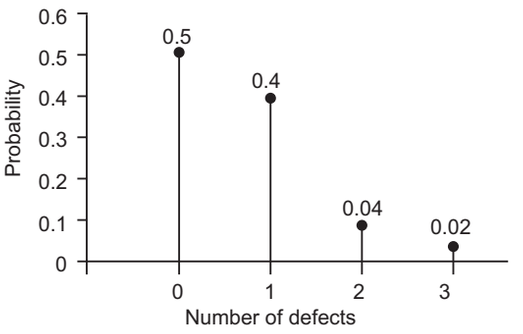


Fig. E3.1(a)

(c) Since $F(2) = P(x \leq 2)$, we need to find $P(x \leq 2)$. We can do this by summing the probabilities for the value of x that are less than or equal to 2, namely 0, 1, and 2. Hence

$F(2) = P(x \leq 2) = P(x = 0) + P(x = 1) + (x = 2) = 0.50 + 0.40 + 0.08 = 0.98$

Now $F(1.5) = P(x \leq 1.5)$. Hence, to compute $F(1.5)$, we must sum the probability for the values of x that are less than or equal to 1.5, which are 0 and 1. Therefore,

$F(1.5) = P(x \leq 1.5) = p(x = 0) + P(x = 1) = 0.50 + 0.40 = 0.90$

(d) First, we find $F(x)$ for each of the possible values of x, which are 0, 1, 2, and 3.

$F(0) = P(x \leq 0) = 0.50$

$F(1) = P(x \leq 1) = 0.50 + 0.40 = 0.90$

$F(2) = P(x \leq 2) = 0.50 + 0.40 + 0.08 = 0.98$

$F(3) = P(x \leq 3) = 0.50 + 0.40 + 0.08 + 0.02 = 1$

The plot of $F(x)$ is presented in Fig. E3.1(b).

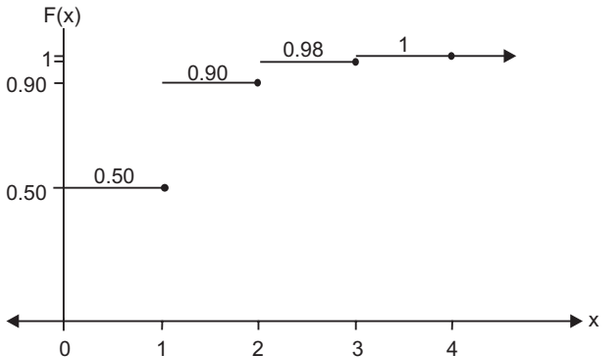


Fig. E3.1(b)

(e) The sample mean is the total number of defects divided by 100.

$$\text{mean} = \frac{0(50) + 1(40) + 2(8) + 3(2)}{100} = 0.62$$

This can be written as

$$\text{mean} = 0(0.50) + 1(0.40) + 2(0.08) + 3(0.02) = 0.62$$

The above calculation for the mean shows that the mean of a perfect sample can be obtained by multiplying each possible value of x by its probability, and summing the products, which is the definition of the population mean of a discrete random variable.

Example E3.2

Table E3.2 gives the probability distribution of a discrete random variable x.

Table E3.2

x	0	1	2	3	4	5
P(x)	0.03	0.13	0.22	0.30	0.20	0.12

Find the following probabilities:

(a) $P(x = 1)$, $P(x \leq 1)$, $P(x \leq 3)$ and $P(0 \leq x \leq 2)$

(b) probability that x assumes a value less than 3

(c) probability that x assumes a value greater than 3

(d) probability that x assumes a value in the interval 2 and 4.

SOLUTION:

(a) $P(x = 1) = 0.13$

$P(x \leq 1) = P(0) + P(1) = 0.03 + 0.13 = 0.16$

$P(x \leq 3) = P(3) + P(4) + P(5) = 0.30 + 0.20 + 0.12 = 0.62$

$P(0 \leq x \geq 2) = P(0) + P(1) + P(2) = 0.03 + 0.13 + 0.22 = 0.38$

(b) $P(x < 3) = P(0) + P(1) + P(2) = 0.03 + 0.13 + 0.22 = 0.38$

(c) $P(x > 3) = P(4) + P(5) = 0.20 + 0.12 = 0.32$

(d) $P(2 \leq x \leq 4) = P(2) + P(3) + P(4) = 0.22 + 0.30 + 0.20 = 0.72$

Example E3.3

Table E3.3 lists the probability distribution of x, where x is the number of car accidents that occur in a city during a week.

Table E3.3

x	0	1	2	3	4	5	6
P(x)	0.12	0.16	0.22	0.17	0.15	0.12	0.06

(a) draw a line graph for the probability distribution

(b) find the probability that the number of car accidents that will occur during a given week in the city is
(i) exactly 4 *(ii)* at least 3 *(iii)* less than 3 *(iv)* 3 to 5

SOLUTION:

(a) The line graph is shown in Fig. E3.3.

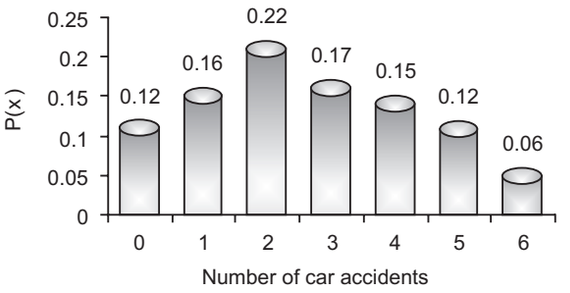


Fig. E3.3

(b) *(i)* $P(\text{exactly } 4) = P(4) = 0.15$

(ii) $P(\text{at least } 3) = P(x \geq 3) = P(3) + P(4) + P(5) + P(6) = 0.17 + 0.15 + 0.12 + 0.06 = 0.50$

(iii) $P(\text{less than } 3) = P(x < 3) = P(0) + P(1) + P(2) = 0.12 + 0.16 + 0.22 = 0.50$

(iv) $P(3 \text{ to } 5) = P(3) + P(4) + P(5) = 0.17 + 0.15 + 0.12 = 0.44$

Example E3.4

According to a survey, 75% of all elementary school teachers in a city in the year 2005 were women. Assume that this result holds true for the current population of all teachers in that city. Suppose 2 teachers are randomly selected from the population of all teachers in that city. Denoting x as the number of women in that sample, construct the probability distribution table of x. Draw a tree diagram for this problem.

SOLUTION:

Let M = teacher selected is a man

W = teacher selected is a woman

Then $P(W) = 0.75$ and $P(M) = 1 - P(W) = 1 - 0.75 = 0.25$

The tree diagram is shown in Fig. E3.4.

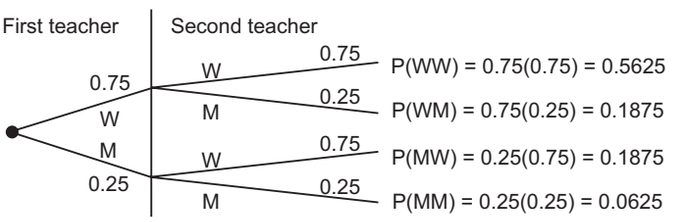


Fig. E3.4

Now, x = the number of women in a sample of 2 teachers. Table E3.4 lists the probability distribution of x, where $x = 0$ represents if neither teacher is a woman, $x = 1$ if the teacher is a woman and one is a man, and

$x = 2$ if the both teachers are woman. The probability in the table are obtained from the tree diagram in Fig. E3.4. The probability of $x = 1$ is obtained by adding the probabilities of *WM* and *MW*.

Table E3.4

Outcomes	x	P(x)
MM	0	0.0625
WM or MW	1	0.3750
WW	2	0.5625

Example E3.5

A machine design class has 20 students and 11 of them are female. Suppose 2 students are randomly selected from the class and x denotes the number of females in the sample.

(a) draw a tree diagram

(b) find the probability distribution of x.

Assume here that the draws are made without replacement from a small population and the probabilities of outcomes do not remain constant for each draw.

SOLUTION:

Let A = first student selected is a male

B = first student selected is a female

C = second student selected is a male

D = second student selected is a female

(a) The tree diagram is shown in Fig. E3.5.

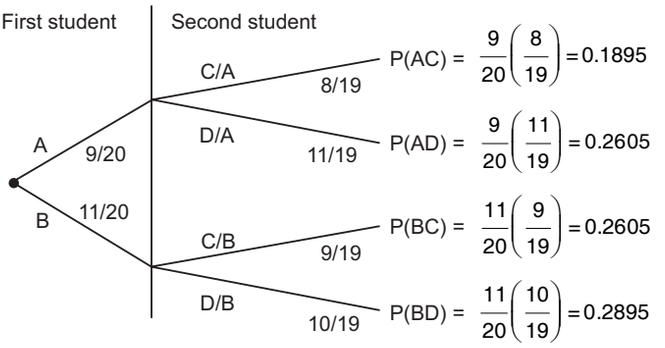


Fig. E3.5

(b) Table E3.5 lists the probability distribution of x, where x is the number of females in a sample of 2 students.

Table E3.5

Outcomes	x	P(x)
AC	0	0.1895
AD or BC	1	0.5210
BD	2	0.2895

Example E3.6

Table E3.6 lists the probability distribution of the number of breakdowns per week for a machine based on past data.

Table E3.6

Breakdowns per week	0	1	2	3
Probability	0.15	0.25	0.40	0.20

(a) present this probability distribution graphically

(b) find the probability that the number of breakdowns for this machine during a given week is

(i) exactly 2 *(ii)* 0 to 2 *(iii)* more than 1 *(iv)* at most 1.

SOLUTION:

Let x denotes the number of breakdowns for this machine during a given week. The Table E3.6 lists the probability distribution of x.

Table E3.6: Probability distribution of breakdowns

x	P(x)
0	0.15
1	0.25
2	0.40
3	0.20
	ΣP(x) = 1.00

(a) Figure E3.6 shows the line graph of the probability distribution.

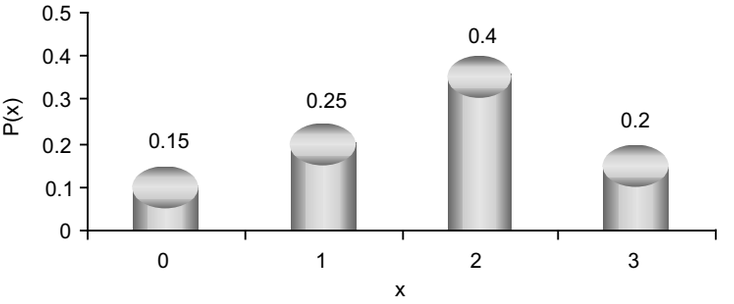


Fig. E3.6: Graphical representation of the probability distribution

(b) *(i)* The probability of exactly, two breakdowns is

$P(\text{exactly 2 breakdowns}) = P(x = 2) = 0.40$

(ii) The probability of 0 to 2 breakdowns is given by the sum of the probabilities of 0, 1 and 2 breakdowns

$P(0 \text{ to } 2 \text{ breakdowns}) = P(0 \le x \le 2)$

$= P(x = 0) + P(x = 1) + P(x = 2) = 0.15 + 0.25 + 0.40 = 0.80$

(iii) The probability of more than 1 breakdown is obtained by adding the probabilities of 2 and 3 breakdowns

$P(\text{more than 1 breakdown}) = P(x > 1)$

$= P(x = 2) + P(x = 3) = 0.40 + 0.20 = 0.60$

(iv) The probability of at most 1 breakdown is given by the sum of the probabilities of 0 and 1 breakdowns.

P(at most 1 breakdown) = $P(x \leq 1)$

$= P(x = 0) + P(x = 1) = 0.15 + 0.25 = 0.40$

Example E3.7

Table E3.7 lists the probability distribution of x, where x is the number of defects contained in a randomly selected manufactured product. Find the mean and standard deviation of x.

Table E3.7

x	0	1	2	3	4
P(x)	0.41	0.12	0.05	0.03	0.01

SOLUTION:

Refer to Table E3.7(a)

Table E3.7(a)

x	P(x)	x P(x)	x^2 P(x)
0	0.41	0	0
1	0.12	0.12	0.12
2	0.05	0.10	0.20
3	0.03	0.09	0.27
4	0.01	0.04	0.16
Total		0.35	0.75

Mean $\mu = \Sigma xP(x) = 0.35$ defects (from Table E3.7(a)).

Standard deviation $\sigma = \sqrt{x^2 P(x) - \mu^2} = \sqrt{0.75 - 0.35^2} = 0.79$ defects.

Example E3.8

Find the mean, the variance, and the standard deviation of x, where x denotes the number that shows up when a fair die is rolled.

SOLUTION:

The probability function of x and the necessary calculations are summarised in Table E3.8.

Table E3.8

x	p(x)	x p(x)	x^2p(x)
1	1/6	1/6	1/6
2	1/6	2/6	4/6
3	1/6	3/6	9/6
4	1/6	4/6	16/6
5	1/6	5/6	25/6
6	1/6	6/6	36/6
Σ	1	21/6 = 3.5	91/6 = 15.1667

The mean given by

$$\mu = \Sigma x p(x) = \frac{21}{6} = 3.5$$

Variance is given by

$$\sigma^2 = \Sigma x^2 p(x) - \mu^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = 2.92$$

The standard deviation = $\sqrt{\sigma^2} = \sqrt{2.92} = 1.71$

Hence, mean = 3.5, variance = 2.92 and standard deviation = 1.71.

3.1.3 Continuous Random Variables

A random variable is *continuous* if its probabilities are given by areas under a curve. The curve is called a *probability density function* for the random variable. The probability density function is sometimes called the *probability distribution*.

If the range space of the random variable X is an interval or a collection of intervals, X is called a *continuous random variable*. If a and b are real numbers and $a < b$, then the probability that X lies in the interval $[a, b]$ is defined as

$$P[a \le X \le b] = \int_a^b f(x)\, dx \tag{3.14}$$

The function $f(x)$ is known as the *probability density function* (or pdf) of the random variable X. The pdf must satisfy the following conditions:

1. $f(x) \ge 0, \qquad -\infty < x < \infty$
2. $\int f(x)\, dx = 1$ (3.15)

The probability density function (pdf) is shown in Fig. 3.3. The shaded area in Fig. 3.3 represents the probability that X lies in the interval $[a, b]$.

From Eq. (3.14), for any specified value of x, $P(X = x) = 0$, since

$$\int_x^x f(t)\, dt = 0 \tag{3.16}$$

Hence, we can also write

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b) \tag{3.17}$$

Since x is continuous

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.18}$$

The following conditions are satisfied

1. F is a non-decreasing function. If $a < b$, then $F(a) \le F(b)$.
2. $\lim_{x \to \infty} F(x) = 1$

3. $\lim_{x \to -\infty} F(x) = 0$ (3.19)

All probability questions about X can be expressed in terms of the cdf. For instance

$$P(a < X \leq b) = F(b) - F(a) \text{ for } a < b \tag{3.20}$$

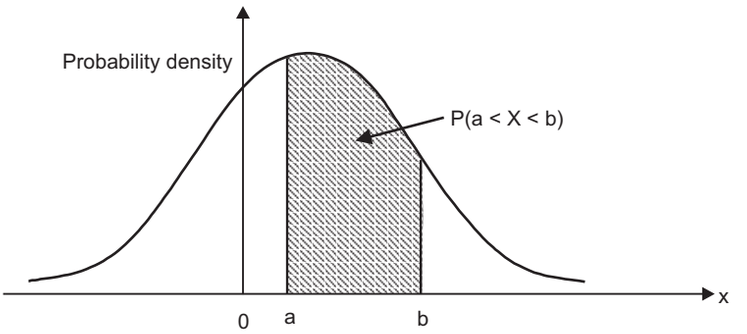


Fig. 3.3: The probability density function (pdf)
(The shaded area under the curve is P(a < X < b))

The cumulative distribution function of a continuous random variable X is $F(X) = P(X \leq x)$, just like for a discrete random variable. For a continuous variable, the value of $F(x)$ is obtained by integrating the probability density function.

Since $F(x) = P(X \leq x)$, we have

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.21}$$

where $f(t)$ is the probability density function.

Hence, the continuous distribution function (cdf) of X is the function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.22}$$

3.1.4 Mean and Variance for Continuous Random Variables

The population mean and variance of a continuous random variable are defined in the same way as those of a discrete random variable, except that the probability density function is used instead of the probability mass function. Specifically, if X is a continuous random variable, its population mean is defined to be the center of mass of its probability density function and its population variance is the moment of inertia around a vertical axis through the population mean.

Let X be a continuous random variable with probability density function $f(x)$. Then, the mean is given by

$$\mu_x = \int_{-\infty}^{\infty} x\, f(x)\, dx \tag{3.23}$$

The *mean* of X is sometimes called the *expectation* or the *expected* value of X and denoted by $E(X)$ or by μ.

The variance of X is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\, dx \tag{3.24}$$

An alternative formula for the variance is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu_X^2 \tag{3.25}$$

The variance of X is also denoted by $V(X)$ or by σ^2.

The standard deviation is the square root of the variance

$$\sigma_X = \sqrt{\sigma_X^2} \tag{3.26}$$

Alternate Formula for the Variance

Starting with the equation

$$\sigma_X^2 = \sum_x (x-\mu_x)^2 P(X=x) \tag{3.27}$$

Expanding the above equation, we obtain

$$\sigma_X^2 = \sum_x (x^2 - 2x\mu_X + \mu_X^2)\, P(X=x) \tag{3.28}$$

or

$$\sigma_X^2 = \sum_x [x^2 P(X=x) - 2x\mu_X P(X=x) + \mu_X^2 P(X=x)] \tag{3.29}$$

Summing the terms separately, we have

$$\sigma_X^2 = \sum_x x^2 P(X=x) - \sum_x 2x\mu_X P(X=x) + \sum_x \mu_X^2 P(X=x) \tag{3.30}$$

Noting that

$$\sum_x 2x\mu_X P(X=x) = 2\mu_X \sum_x xP(X=x) = 2\mu_X\mu_X = 2\mu_X^2 \tag{3.31}$$

and

$$\sum_x \mu_X^2 P(X=x) = \mu_X^2 \sum_x P(X=x) = \mu_X^2 (1) = \mu_X^2 \tag{3.32}$$

Substituting Eqs. (3.31) and (3.32) in Eq. (3.30), we get

$$\sigma_X^2 = \sum_x x^2 P(X=x) - 2\mu_X^2 + \mu_X^2 \tag{3.33}$$

or

$$\sigma_X^2 = \sum_x x^2 P(X=x) - \mu_X^2 \tag{3.34}$$

3.1.5 Expectation

A brief summary is given here:

If X is a random variable, the expected value of X, denoted by $E(X)$, for discrete and continuous variable is defined as

$$E(X) = \sum_{all\ i} x_i\, p(x_i) \qquad \text{if } X \text{ is discrete} \tag{3.35}$$

and $$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{if } X \text{ is continuous} \tag{3.36}$$

the expected value $E(X)$ of a random variable X is also referred to as the *mean* of X, μ_X (or μ), or the *first moment* of X. The quantity, $E(X^n)$, $n \geq 1$, is called the n^{th} *moment* of X and is defined as

$$E(X'') = \sum_{all\ i} x_i^n\, p(x_i) \qquad \text{if } X \text{ is discrete} \tag{3.37}$$

and $$E(X'') = \int_{-\infty}^{\infty} x^n f(x)\, dx \qquad \text{if } X \text{ is continuous} \tag{3.38}$$

The variance of a random variable, X, denoted by $V(X)$, Var(X), or σ^2, is defined by

$V(X) = E([X - E(X)]^2)$

A useful identity in computing $V(X)$ is given by

$$V(X) = E(X^2) - [E(X)]^2 \tag{3.39}$$

The mean $E(X)$ is a measure of the central tendency of a random variable. The variance, $V(X)$, is a measure of the spread or dispersion of the possible values of X around the mean $E(X)$. The standard deviation, σ, is the square root of the variance.

Example E3.9

The life of an electronic component is given by X, a continuous random variable assuming all values in the range $x \geq 0$. The pdf of the lifetime, in years is given as

$$f(x) = \begin{cases} \frac{1}{4} e^{-\frac{x}{4}}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

(a) determine the probability that the life of this component is between 2 and 4 years.

(b) determine cdf of the components

(c) determine the probability that this electronic component will last for 4 years

(d) find the mean and variance of the life of the component.

SOLUTION:

(a) $$P(2 \leq X \leq 4) = \frac{1}{4}\int_2^4 e^{-\frac{x}{4}} dx = -e^{-1} + e^{-\frac{1}{2}} = 0.2386$$

The probability that the laser will last between 2 and 4 years is obtained also from

$P(2 \leq X \leq 4) = F(4) - F(2) = (1 - e^{-4/4}) - (1 - e^{-2/4}) = e^{-1/2} - e^{-1} = 0.6065 - 0.3679 = 0.2386$

(b) The cdf of this component is given by

$$F(x) = \frac{1}{4}\int_0^x e^{-\frac{x}{4}}dx = 1 - e^{-\frac{x}{4}}$$

(c) The probability that the component will last for 4 years or less is given by

$$P(0 \le X \le 4) = F(4) - F(0) = F(4) = 1 - e^{4/4} = 0.6321$$

(d) The mean and variance of the life of the electronic component are determined as follows:

$$E(X) = \frac{1}{4}\int_0^\infty xe^{-\frac{x}{4}}dx = xe^{-\frac{x}{4}}\Big|_0^\infty + \int_0^\infty e^{-\frac{x}{4}}dx = 0 + \frac{1}{1/4}e^{-\frac{x}{4}}\Big|_0^\infty = 4 \text{ years}$$

To compute $V(X)$, first compute $E(X^2)$ as follows:

$$E(X^2) = \frac{1}{4}\int_0^\infty x^2 e^{-\frac{x}{4}}dx = x^2 e^{-\frac{x}{4}}\Big|_0^\infty + 8\int_0^\infty xe^{-\frac{x}{4}}dx = 32$$

Hence, variance is given by

$$V(X) = 32 - 4^2 = 16 \text{ years}$$

Example E3.10

Given $f(x) = e^{-x}$ for $0 < x$. Find the following probabilities:

(a) $P(1 < X)$

(b) $P(1 < X < 2.5)$

(c) $P(X = 3)$

(d) $P(X < 4)$

(e) $P(3 \le X)$

SOLUTION:

(a) $$P(1 < X) = \int_1^\infty e^{-x}dx = (-e^{-x})\Big|_1^\infty = e^{-1} = 0.3679$$

(b) $$P(1 < X < 2.5) = \int_1^{2.5} e^{-x}dx = (-e^{-x})\Big|_1^{2.5} = e^{-1} - e^{-2.5} = 0.2858$$

(c) $$P(X = 3) = \int_3^3 e^{-x}dx = 0$$

(d) $$P(X < 4) = \int_0^4 e^{-x}dx = (-e^{-x})\Big|_0^4 = 1 - e^{-4} = 0.9817$$

(e) $P(3 \le X) = \int_3^{\infty} e^{-x}\,dx = (-e^{-x})\Big|_3^{\infty} = e^{-3} = 0.0498$

Example E3.11

The probability density of the continuous random variable X is given by

$$f(x) = \begin{cases} 1/5 & \text{for } 2 < x < 7 \\ 0 & \text{elsewhere} \end{cases};$$

Find *(a)* $P(3 < X < 5)$

(b) area under the curve.

SOLUTION:

(a) $P(3 < X < 5) = \int_3^5 \frac{1}{5}\,dx = \frac{1}{5}\int_3^5 1\,dx = \frac{1}{5}(5-3) = \frac{2}{5}$

(b) Area under the curve $= \int_{-\infty}^{\infty} f(x)\,dx = \int_2^7 \frac{1}{5}\,dx = \frac{1}{5}x\Big|_2^7 = \frac{1}{5}(7-2) = 1$

Example E3.12

Given $f(x) = 1.5x^2$ for $-1 < x < 1$. Find the mean and variance of x.

SOLUTION:

$$E(X) = \int_{-1}^{1} x(1.5x^2)\,dx = \int_{-1}^{1} 1.5x^3\,dx = 1.5\frac{x^4}{4}\Big|_{-1}^{1} = 0$$

$$V(X) = \int_{-1}^{1} 1.5x^2(x-0)^2\,dx = 1.5\int_{-1}^{1} x^4\,dx = 1.5\frac{x^5}{5}\Big|_{-1}^{1} = 0.6$$

Example E3.13

The pdf of the random variable X is given by

$$f(x) = \begin{cases} \dfrac{a}{\sqrt{x}} & \text{for } 0 < x < 4 \\ 0 & \text{elsewhere} \end{cases};$$

Find *(a)* the value of a

(b) $P\left(X < \frac{1}{4}\right)$ and $P(X > 1)$.

SOLUTION:

(a) $\text{Area} = 1 = \int_0^4 \frac{a}{\sqrt{x}}\,dx = a\int_0^4 x^{-1/2}\,dx = a\left.\frac{x^{1/2}}{1/2}\right|_0^4 = 2a\cdot 2 = 4a$

Hence, $a = 1/4$.

(b) $P\left(X < \frac{1}{4}\right) = \int_0^{1/4} \frac{1}{4\sqrt{x}}\,dx = \frac{1}{4}\int_0^{1/4} x^{-1/2}\,dx = \frac{1}{4}\left.\frac{\sqrt{x}}{1/2}\right|_0^{1/4} = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$

$$P(X > 1) = 1 - \int_0^1 \frac{1}{4\sqrt{x}}\,dx = 1 - \left.\frac{1}{2}\sqrt{x}\right|_0^1 = \frac{1}{2}$$

Example E3.14

The thickness of a conductive coating in micrometers has a density function of $f(x) = 600\,x^{-2}$ for $100\ \mu\text{m} < x < 120\ \mu\text{m}$.

(a) find the mean and variance of the coating thickness

(b) if the cost of coating is \$0.25 per micrometer of thickness on each component, find the average cost of the coating per component.

SOLUTION:

(a) $E(X) = \int_{100}^{120} x\frac{600}{x^2}\,dx = 600\ln x\Big|_{100}^{120} = 109.39$

$$V(X) = \int_{100}^{120} (x - 109.39)^2\,\frac{600}{x^2}\,dx = 600\int_{100}^{120} 1 - \frac{2(109.39)}{x} + \frac{(109.39)^2}{x^2}\,dx$$

$$= 600(x - 218.75\ln x - 109.39^2\,x^{-1})\Big|_{100}^{120} = 33.19$$

(b) Average cost per component = \$0.25(109.39) = \$27.35.

Example E3.15

The distribution function of the random variable X is given by

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

Find *(a)* $P(X \le 2)$

(b) $P(1 < X < 3)$

(c) $P(X > 4)$

SOLUTION:

(a) $P(X \le 2) = F(2) = 1 - 3e^{-2} = 1 - 3(0.1353) = 1 - 0.4074 = 0.5926$

(b) $P(1 < X < 3) = F(3) - F(1) = 1 - 4e^{-3} - 1 + 2e^{-1} = 2e^{-1} - 4e^{-3} = 2(0.3679) - 4(0.0498) = 0.5366$

(c) $P(X > 4) = 1 - F(4) = 5e^{-4} = 5(0.0183) = 0.0915$

Example E3.16

The pdf of a random variable X is given by

$$f(x) = \begin{cases} \frac{4}{3}(1-x^3) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Sketch the pdf and find the values of the following:

(a) $P\left(X < \frac{1}{2}\right)$

(b) $P\left(\frac{1}{4} < X < \frac{3}{4}\right)$

(c) $P\left(X > \frac{1}{3}\right)$

SOLUTION:

The plot is shown in Fig. E3.16.

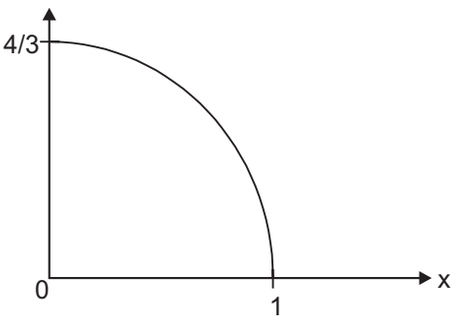


Fig. E3.16 pdf of the function f(x)

(a) $P\left(X < \frac{1}{2}\right) = \int_0^{1/2} 4(1-x^3)\frac{dx}{3} = 0.6458$

(b) $P\left(\frac{1}{4} < X < \frac{3}{4}\right) = \int_{1/4}^{3/4} 4(1-x^3)\frac{dx}{3} = 0.5625$

(c) $P\left(X > \frac{1}{3}\right) = \int_{1/3}^{1} 4(1-x^3)\frac{dx}{3} = 0.5597$

Example E3.17

The pdf of a random variable X is given by

$$f(x) = \begin{cases} ax^2 & \text{for } 0 \le x \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

(a) find the value of the constant a and sketch the pdf

(b) find $P(X > 3/2)$.

SOLUTION:

(a) We must have

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_1^2 a\,x^2\,dx = \frac{7}{3}a = 1$$

Hence, $a = 3/7$. The pdf is shown in Fig. E3.17.

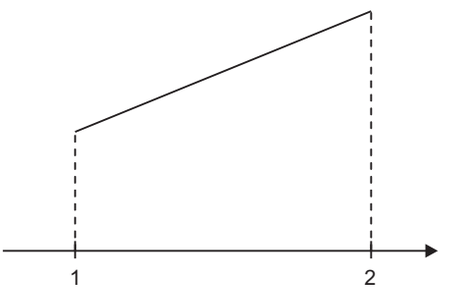


Fig. E3.17

(b) $P(X > 3/2) = \int_{3/2}^{2} f(x)\,dx = \int_{3/2}^{2} \frac{3}{7}x^2 dx = 37/56$

3.2 PERMUTATIONS AND COMBINATIONS

Factorial Notation: The product of the positive integers from 1 to n inclusive is denoted by $n!$ (read as "n factorial"). The value of the factorial of a number is obtained by multiplying all integers from that number to 1.

The symbol $n!$, read as "n factorial" represents the product of all integers from n to 1. In other words,

$$n! = n(n-1)(n-2)(n-3)\cdots 3\cdot 2\cdot 1 \tag{3.40}$$

It is also convenient to define

$$0! = 1,\ 1! = 1 \text{ and } n! = n\cdot((n-1)!) \tag{3.41}$$

Example E3.18

Evaluate the following:

(a) 2!, 3!, 4!, 5!, 6! and 7!

(b) (7 – 3)!, (14 – 12)!, (7 – 2)!, (13 – 4)!, and (6 – 6)!

(c) $\frac{8!}{6!}, \frac{12!}{9!}$ and $\frac{12!}{9!3!}$

SOLUTION:

(a) $2! = 2\cdot 1 = 2$

$3! = 3\cdot 2\cdot 1 = 6$

$4! = 4\cdot 3\cdot 2\cdot 1 = 24$

$5! = 5\cdot 4\cdot 3\cdot 2\cdot 1 = 120$

$6! = 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1 = 720$

$7! = 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1 = 5040$

(b) $(7-3)! = 4! = 4\cdot 3\cdot 2\cdot 1 = 24$

$(14-12)! = 2! = 2.1 = 2$

$(7-2)! = 5! = 5\cdot4\cdot3\cdot2\cdot1 = 120$

$(13-4)! = 9! = 9\cdot8\cdot7\cdot6\cdot5\cdot4\cdot3\cdot.2\cdot1 = 362{,}880$

$(6-6)! = 0! = 1$

(c) $\dfrac{8!}{6!} = \dfrac{8\cdot7\cdot6!}{6!} = 8.7 = 56$

$\dfrac{12!}{9!} = \dfrac{12\cdot11\cdot10\cdot9!}{9!} = 12\cdot11\cdot10 = 1320$

$\dfrac{12!}{9!3!} = \dfrac{12\cdot11\cdot10\cdot9!}{9!3!} = \dfrac{12\cdot11\cdot10}{3.2} = 220$

Example E3.19

Find the value of 9! by using the table in Appendix-*A*.

SOLUTION:

We locate 9 in the column labeled *n*. Then we read the value in the column for *n*! entered next to 9. Hence,

$9! = 362, 880$

3.2.1 Permutations

Any arrangement of a set of *n* objects in a given order is called a *permutation* of the objects (taken all at a time). Any arrangement of any $r \le n$ of those objects in a given order is called an *r-permutation* or *a permutation of the n objects taken r at a time*.

Ordered arrangement of a set of objects are called *permutations*. When the order in which they are arranged is disregarded, then the arrangements are known as *combinations*. The number of permutations of *n* different items is the number of different arrangements in which these items can be placed. A permutation of *n* different objects taken *r* at a time is an arrangement of *r* out of the *n* objects with attention given to the order of arrangement. The number of permutations of *n* objects taken *r* at a time is denoted by nP_r, $P(n, r)$ P_r^n, $(n)_r$ or $P_{n,r}$ and is written as

$$nP_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} \tag{3.42}$$

Derivation of the Formula nP_r

The first element in an *r*-permutation of *n* objects can be chosen in *n* different ways. Following this way, the second element in the permutation can be chosen in $n-1$ ways, and the 3rd element in the permutation can be chosen in $n-2$ ways etc. Finally, the r^{th} (last) element in the *r*-permutation can be chosen in $n-(r-1) = n-r+1$ ways. Hence, from the fundamental principle of counting, we have

$$nP_r = P(n, r) = n(n-1)\,(n-2)\cdots(n-r+1) \tag{3.43}$$

or $$n(n-1)\,(n-2)\cdots(n-r+1) = \frac{n(n-1)(n-2)\cdots(n-r+1).(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!} \tag{3.44}$$

in which $n!$ is the symbol for factorial $n = n(n-1)\cdots 3 \times 2 \times 1$, when $r = n$, $n\,P_n = n!$ (note by definition $0! = 1$). Tables of the values of factorial n are given in Appendix-*A*.

The permutation of n items taken all at a time, when the n items consists of r_1 alike, r_2 alike, … r_k alike, so that by definition $r_1 + r_2 + \cdots + r_k = n$.

The number of permutation is then

$$P = \frac{n!}{r_1!r_2!\cdots r_k!} \tag{3.45}$$

Also, we have $nP_n = n(n-1)\cdots 3\cdot 2\cdot 1 = n!$

Example E3.20

How many five-digit numbers can be formed from the numbers 1 to 8?

SOLUTION:

Here $n = 8$ and $r = 5$.

$$nP_r = \frac{n!}{(n-r)!} = \frac{8!}{(8-5)!} = \frac{8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}{3\times 2\times 1} = 1680 \text{ numbers.}$$

Example E3.21

How many numbers of permutations of letters are there in the word *quality*?

SOLUTION:

Here $n = 7$ (7 letters in word *quality*)

$r_1 = 1(q)$

$r_2 = 1(u)$

$r_3 = 1(a)$

$r_4 = 1(l)$

$r_5 = 1(i)$

$r_6 = 1(t)$

$r_7 = 1(y)$

Therefore,

$$P = \frac{n!}{r_1!r_2!\cdots r_{k!}} = \frac{7!}{1!1!1!\cdots 1!} = 7\times 6\times 5\times 4\times 3\times 2\times 1 = 5040$$

Example E3.22

(a) Evaluate $7P_3$, $6P_4$, $16P_1$, $3P_3$ and $6P_3$

(b) five persons arrive at the bank teller window at the same time. In how many different ways can these people line up?

(c) find the number of ways of arranging the letters in the word *subject*

(d) find the number of ways if each arrangement in part *(c)* starts with the letter *j*.

SOLUTION:

(a) $7P_3 = 7 \cdot 6 \cdot 5 = 210$

$6P_4 = 6 \cdot 5 \cdot 4 \cdot 3 = 360$

$16P_1 = 16$

$3P_3 = 3 \cdot 2 \cdot 1 = 6$

$6P_3 = 6 \cdot 5 \cdot 4 = 120$

(b) These paper can line up in different ways

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ ways.

(c) There are 7 distinct letters in the word *subject*. Hence, the number of arrangements is 7!. That is

$7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$

(d) Since the first letter is fixed as *j*, the total number of possibilities are found by arranging the remaining six letters in all possible ways. This can be done by 6! ways.

$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$ ways.

3.2.2 Combinations

As noted earlier, the order in which objects are arranged matters in *permutations*. However, order does not matter in *combinations*.

The number of combinations of n different objects taken r at a time is a selection of r out of the n objects with no attention given to the order of arrangement. The number of combinations of n objects taken r at a time is denoted by nC_r $C(n, r)$, $C_{n,r}$ or $\left(\frac{n}{r}\right)$ is $r!$ times smaller than the number of permutations.

Therefore,

$$nC_r = \binom{n}{r} = \frac{nP_r}{r!} = \frac{n!}{(n-r)!r!} \tag{3.46}$$

It should be noted from symmetry that the identity is valid

$$nC_r = nC_{n-r} \tag{3.47}$$

Also $\binom{n}{r} = \frac{nP_r}{r!}$

The arrangement for dividing the number of permutations by $r!$ to get the number of combinations is that each combination give rise to $r!$ possibilities when arranged in all possible ways, thereby giving all the permutations.

Tables of binomial coefficients $\left\{\frac{n}{r}\right\}$ are given in Appendix-*B*.

Note that in combinations, n is always greater than or equal to r. If n is smaller than r, then we cannot select r distinct elements from n.

Example E3.23

(a) Evaluate the following:

$$\binom{8}{4}, \binom{6}{6}, \binom{9}{3}, \binom{4}{0}, \binom{5}{3}, \binom{15}{2} \text{ and } \binom{20}{8}$$

(b) Four cards are chosen in succession from a deck of 52 cards. Find the number of ways this can be done *(i)* with replacement, *(ii)* without replacement.

(c) find the number m of groups of 3 that can be formed from 8 students in a class.

SOLUTION:

(a)

$$\binom{8}{4} = \frac{8\cdot7\cdot6\cdot5}{4!} = \frac{8\cdot7\cdot6\cdot5}{4\cdot3\cdot2\cdot1} = 70$$

$$\binom{6}{6} = \frac{6!}{6!(6-6)!} = \frac{1}{0!} = \frac{1}{1} = 1$$

$$\binom{9}{3} = \frac{9!}{(9-3)!3!} = \frac{9!}{6!3!} = \frac{9\cdot8\cdot7\cdot6\cdot5\cdot4\cdot3\cdot2\cdot1}{6\cdot5\cdot4\cdot3\cdot2\cdot1\cdot3\cdot2\cdot1} = 84$$

$$\binom{4}{0} = \frac{4!}{0!(4-0)!} = \frac{4!}{1\cdot4!} = 1$$

$$\binom{5}{3} = \frac{5!}{(5-3)!3!} = \frac{5!}{2!3!} = \frac{5.4.3.2.1}{2.1.3.2.1} = 10$$

$$\binom{15}{2} = \frac{15!}{(15-2)!2!} = \frac{15\cdot14\cdot13\cdots2\cdot1}{13\cdot12\cdots2\cdot1\cdot2\cdot1} = 105$$

$$\binom{20}{8} = \frac{20!}{(20-8)!8!} = 125,970$$

(b) *(i)* Since each card is replaced before the next card is chosen, each card can be chosen by 52 ways. Hence, there are

$(52)(52)(52)(52) = 52^4 = 7,311,616$

different ordered samples of size $n = 4$ with replacement.

(ii) Since there is no replacement, the first card can be chosen in 52 ways, the second card in 51 ways, 3rd card in 50 ways and the last card in 49 ways. Thus there are

$P(52, 4) = 52(51)(50)(49) = 6,497,400$

different ordered samples of size $r = 4$ without replacement.

(c)

$$m = C(8,3) = \binom{8}{3} = \frac{8\cdot7\cdot6}{3\cdot2\cdot1} = 56$$

Since each group is essentially a combination of 8 students taken 3 at a time.

Example E3.24

The president of a university has 12 departments under his administration. The president plans to visit 3 of these departments during the next week. If he randomly selects 3 departments from these 12, how many total selections are possible? Verify the answer using the table in Appendix-*B*.

SOLUTION:

The total possible selections for selecting 3 departments from 12 are

$$\binom{12}{3} = \frac{12!}{(12-3)!3!} = 220$$

From the table in Appendix-*B*, we read for $n = 12$ and $r = 3$, the answer is 220.

Example E3.25

Out of 5 men and 7 women, a committee consisting of 2 men and 3 women is to be formed. In how many ways can this be done if

(a) any men and any women can be included

(b) one particular women must be included on the committee

(c) two particular men cannot be included on the committee.

SOLUTION:

(a) 2 men out of 5 can be selected in $5C_2$ ways.
3 women out of 7 can be selected in $7C_3$ ways.
Hence, the number of possible selections = $(5C_2)\ (7C_3) = (10)\ (35) = 350$

(b) 2 men out of 5 can be selected $5C_2$ ways.
2 additional women out of 6 can be selected in $6\ C_2$ ways.
Hence, the number of possible selections = $(5C_2)\ (6C_3) = (10)\ (15) = 150$

(c) 2 men out of 3 can be selected $3C_2$ ways.
3 women out of 7 can be selected in $7C_2$ ways.
Hence, the number of possible selections = $(3C_2)\ (7C_3) = (3)\ (35) = 105$

3.3 DISCRETE DISTRIBUTIONS

Discrete random variables are used to describe random phenomena in which only integer values can occur.

3.3.1 Hypergeometric Distribution

In this section we consider dependent Bernoulli random variables. A common source of dependent Bernoulli random variables is sampling without replacement from a finite population. Suppose that a finite population consists of a known number of successes and failures. If we sample a fixed number of units from that population, the number of successes in our sample will have a distribution that is a member of the family of hypergeometric distributions.

Definition of the Hypergeometric Distribution

A set of N objects contains K objects classified as successes and $N - K$ objects classified as failures. A sample of size n objects is selected randomly (without replacement) from the N objects, where $K \le N$ and $n \le N$.

Let the random variable X denotes the number of successes in the sample. Then X is a *hypergeometric random variable* and

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \qquad x = \max\{0, n + K - N\} \text{ to } \min\{K, n\} \tag{3.48}$$

The expression $\{K, n\}$ is used in the definition of the range of X because the maximum number of successes that can occur in the sample is the smaller of the sample size, n, and the number of successes available, K.

Also, if $n + K > N$, at least $n + K - N$ successes must occur in the sample.

It should be noted here that in Eq. (3.48)

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \tag{3.49}$$

is the number of a parts taken b at a time. The hypergeometric distribution is the appropriate possibility model for sampling without replacement. Other distributions are often employed when the ratio of n/N becomes small, say 0.10 or less.

The mean and variance of a hypergoemetric random variable can be determined from the trials that comprise the experiment. However, the trials are not independent, and so the calculations are more difficult than for a binomial distribution. The results are stated as follows:

Mean and Variance of a Hypergeometric Distribution

If X is a hypergoemetric random variable with parameters N, K, and n, then the mean

$$\mu = E(X) = np \tag{3.49a}$$

and the variance

$$\sigma^2 = V(X) = np(1-p)\left(\frac{N-n}{N-1}\right) \tag{3.49b}$$

where

$$p = \frac{K}{N}. \tag{3.49c}$$

Here, p is interpreted as the proportion of successes in the set of N objects. For a hypergeometric random variable, $E(X)$ is similar to the mean of a binomial random variable. Also, $V(X)$ differs from the result for a binomial random variable only by the term shown below:

The term in the variance of a hypergeometric random variable $\frac{N-n}{N-1}$ is called the *finite population correlation factor*.

Example E3.26

Given that X has a hypergeometric distribution with $N = 100$, $n = 4$ and $K = 20$. Determine the following:

(a) $P(X = 1)$

(b) $P(X = 6)$

(c) $P(X = 4)$

(d) the mean and variance of X.

SOLUTION:

Here, $K = 20$, $X = 1$, $N = 100$ and $n = 4$.

(a) $$P(X=1) = \frac{\binom{K}{n}\binom{N-K}{n-X}}{\binom{N}{n}} = \frac{\binom{20}{4}\binom{80}{3}}{\binom{100}{4}} = \frac{20(82160)}{3921225} = 0.4191$$

(b) $P(X = 6) = 0$, since the sample size is only 4.

(c) $$P(X=4) = \frac{\binom{K}{n}\binom{N-K}{n-X}}{\binom{N}{n}} = \frac{\binom{20}{4}\binom{80}{0}}{\binom{100}{4}} = \frac{4845(1)}{3921225} = 0.001236$$

(d) $$\text{Mean} = E(X) = np = n\left(\frac{K}{N}\right) = 4\left(\frac{20}{100}\right) = 0.8$$

$$\text{Variance} = V(X) = np(1-p)\left(\frac{N-n}{N-1}\right) = 40(0.2)(0.8)\left(\frac{96}{99}\right) = 0.6206$$

Example E3.27

Suppose a box contains five red balls and ten blue balls. If seven balls are selected at random without replacement, find the probability that at least 3 red balls will be obtained.

SOLUTION:

Let X denotes the number of red balls that are obtained. Then, X has a hypergeometric distribution, $D = 5$, $A + B = 15$, $n = 7$, $B = 10$ and $A = 5$. The maximum value of X is $\min\{n, A\} = 5$. Hence,

$$P(X \geq 3) = \sum_{x=3}^{5} \frac{\binom{5}{x}\binom{10}{7-x}}{\binom{15}{7}} = \frac{2745}{6435} = 0.4266$$

Example E3.28

A lot of 75 gaskets contains five in which the variability in thickness around the circumference of the gasket is unacceptable. A sample of 10 gaskets is selected at random, without replacement. What is the probability that

(a) none of the unacceptable gaskets is in the sample

(b) at least one unacceptable gasket is in the sample

(c) exactly one unacceptable gasket is in the sample

(d) the mean number of unacceptable gaskets in the sample.

SOLUTION:

Let X denotes the number of unacceptable gaskets in the sample of 10.

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

(a) $$P(X=0) = \frac{\binom{5}{0}\binom{70}{10}}{\binom{75}{10}} = \frac{\frac{70!}{10!60!}}{\frac{75!}{10!65!}} = \frac{65\times 64\times 63\times 62\times 61}{75\times 74\times 73\times 72\times 71} = 0.4786$$

(b) $$P(X\geq 1) = 1 - P(X=0) = 1 - 0.4786 = 0.5214$$

(c) $$P(X=0) = \frac{\binom{5}{1}\binom{70}{9}}{\binom{75}{10}} = \frac{\frac{5!70!}{9!6!}}{\frac{75!}{10!65!}} = \frac{5\times 65\times 64\times 63\times 62\times 10}{75\times 74\times 73\times 72\times 71} = 0.3923$$

(d) $$E(X) = 10\left(\frac{5}{75}\right) = \frac{2}{3}$$

Example E3.29

Of 50 manufactured steel rods in a production process by a company, 12 have defects. If 10 steel rods are selected at random for inspection,

(a) find the probability that exactly 3 of the 10 have defects

(b) find the mean and variance of X.

SOLUTION:

(a) Let X represents the number of sampled steel rods that have defects. Then, $K = 12$, $N = 50$ and $n = 10$.

$$P(X=3)=\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}=\frac{\binom{12}{3}\binom{38}{7}}{\binom{50}{10}}=\frac{(220)(12,620,256)}{10,272,278,170}=0.2703$$

(b) Mean $= \mu_X = np = \dfrac{(10)(12)}{50} = 2.4$

where $p = \dfrac{12}{50}$

Variance $= \sigma_X^2 = np(1-p)\left(\dfrac{N-n}{N-1}\right) = 10\left(\dfrac{12}{50}\right)\left(1-\dfrac{12}{50}\right)\left(\dfrac{50-10}{50-1}\right) = 1.4890$

Example E3.30

A large bin contains 80 balls of which 32 are red balls and 48 are blue balls. Suppose 15 balls are picked at random. Find the probability of getting 4 red balls, the mean number of red balls, and the standard deviation of the number of red balls if the sample is picked

(a) with replacement

(b) without replacement.

SOLUTION:

We will identify success with "picking a red ball" and let X = number of red balls in the sample.

(a) If the sampling is with replacement, we have a binomial distribution with 15 trials and probability of success $p = \dfrac{32}{80} = 0.4$. Hence,

$P(X=4)$ = Binomial(4; 15, 0.4) = 0.127 (from binomial distribution in Appendix-*C*)

$\mu = np = 15(0.4)$

$\sigma = \sqrt{np(1-p)} = \sqrt{15(0.4)(0.6)} = 1.9$

(b) If the sampling is without replacement, we have a hypergeometric distribution and

$$P(X=4)=\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}=\frac{\binom{32}{4}\binom{48}{11}}{\binom{80}{15}}=0.122$$

Note that this answer is not much different from the one obtained in part *(a)* using the binomial formula. The reason is that N (= 80), compared to n (= 15), is fairly large.

Mean, $\mu = np = 15(0.4) = 6$

Standard deviation

$$\sigma = \sqrt{np(1-p)\left(\frac{N-n}{n-1}\right)} = \sqrt{15\left(\frac{32}{80}\right)\left(1-\frac{32}{80}\right)\left(\frac{80-15}{80-1}\right)} = 1.72$$

3.3.2 The Binomial Probability Distribution

The binomial probability distribution is one of the most widely used discrete probability distribution. It is used to find the probability that an outcome will occur x times in n performances of an experiment. The binomial distribution is applied to experiments that satisfy the four conditions of a *binomial experiment*. Each repetition of a binomial experiment is called a *trial* or a *Bernoulli trial* (after Jacob Bernoulli).

A trial with only two possible outcomes is used so frequently as a building block of a random experiment that is called a *Bernoulli trial*. It is usually assumed that the trials that constitute the random experiment are *independent*. This implies that the outcome from one trial has no effect on the outcome to be obtained from any other trial. In addition, it is often reasonable to assume that the *probability of a success in each trial is constant*.

3.3.3 The Binomial Experiment

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.

One of the two outcomes of a trial is called a *success* and the other a *failure*.

The random variable X that equals the number of trials that result in a success has a *binomial random variable* with parameters $0 < p < 1$ and $n = 1, 2, \ldots$. The probability mass function (pmf) of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n \tag{3.50}$$

3.3.4 The Binomial Formula

3.3.4.1 Binomial Theorem

The n^{th} power of $(p + q)$ can be expressed in terms of binomial coefficients as

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{2!}p^{n-2}q^2 + \cdots + \frac{n!}{r!(n-r)!}p^{n-r}q^r + \cdots + q^n$$

$$= \sum_{r=0}^{n}\binom{n}{r}p^{n-r}q^r = \sum_{r=0}^{n}{}_nC_r\,p^{n-r}q^r \tag{3.50a}$$

Since $(p + q) = (q + p)$, we can also write

$$(p+q)^n = \sum_{r=0}^{n} {}_nC_r\, p^r q^{n-r} \tag{3.51}$$

Consider n trials in each of which the probability of success is p. Then the probability of failure is $1 - p = q$. To find the probability of r successes, we observe that:

the probability of 1 success in 1 try is p,

the probability of 2 successes in 2 tries is $p \times p$ or p^2,

the probability of 3 successes in 3 tries is $p \times p \times p$ or p^3,

$\vdots$

the probability of r successes in r tries is p^r and

the probability of subsequent $(n - r)$ failures in $(n - r)$ tries is $(1 - p)^{n-r} = q^{n-r}$.

The probability of r successes followed by $(n - r)$ failures is $p^r(1 - p)^{n-r}$. Here, we have considered only one particular group or combination of r events; i.e., we have started with r successes and finished with $(n - r)$ failures; every other possible ordering of r successes and $(n - r)$ failures will also have the same probability.

The number of possible orderings or the number of selections for r successes and $(n - r)$ failures in n trials is $n!/[r!(n - r)!]$.

Therefore, the probability P_r of an event succeeding r times is

$$P_r = \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r} \tag{3.52}$$

or
$$P_r = {}_nC_r\, p^r q^{n-r} \tag{3.53}$$

This term is similar to the r^{th} term of the binomial expansion $(q + p)^n$ [see Eq. (3.51)], which can be written

$$(q+p)^n = \sum_{r=0}^{n} {}_nC_r\, p^r q^{n-r} \tag{3.54}$$

The successive terms of the expansion give the probability P_r of an event succeeding r times in n trials for values of r varying in steps of one of 0 to n.

3.3.4.2 Cumulative Terms for Binomial Distribution

The probability of an event succeeding at least r times in n trials.

The probability P_r of an event succeeding exactly r times in n trials is given by Eq. (3.53)

$$P_r = {}_nC_r p^r q^{n-r}$$

The probability of an event succeeding at least r' times in n trials is given by

$$\sum_{r=r'}^{r=n} P_r = \sum_{r=r'}^{r=n} {}_nC_r\, p^r q^{n-r} \tag{3.55}$$

The values of the summation of Eq. (3.55) are given in Appendix-*D* for p ranging from 0.05 to 0.50; with n between 2 and 20, and r' between 1 and 20.

For $p > 0.5$, we can utilise the fact that the probability is

$$P_r = 1 - \sum_{r=n-r'+1}^{r=n} {}_nC_r q^r p^{n-r} \tag{3.56}$$

3.3.4.3 Mean and Standard Deviation of Binomial Distribution

If p is the proportion of successes in the population, then the mean number of successes in n trials is

$$\mu = np \tag{3.57}$$

This is clear, as the mean number of successes in n trials is equal to the probability of success in one trial times the number of trials.

The standard deviation for a binomial frequency distribution is

$$\sigma = \sqrt{npq} \text{ or } \sqrt{np(1-p)} \tag{3.58}$$

q is not independent but is equal to $(1 - p)$. The binomial distribution can be expressed in terms of two parameters, n and p.

Equations (3.57) and (3.58) will now be derived using the definition of expectation for the mean and variance. We have

$$E(r) = \mu = \sum_{r=0}^{n} rP_r = \sum_{r=0}^{n} r \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

$$= np \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!(n-r)!} p^{r-1}(1-p)^{n-r} = np[p+q]^{n-1}$$

or
$$\mu = np \tag{3.59}$$

For the variance σ^2

$$E(r-\mu)^2 = \sigma^2 = \sum_{r=0}^{n} (r-\mu)^2 P_r = \sum_{r=0}^{n} r^2 P_r - 2\mu \sum_{r=0}^{n} rP_r + \mu^2 \sum_{r=0}^{n} P_r$$

$$= \sum_{r=0}^{n} r^2 P_r - 2\mu(\mu) + \mu^2$$

Since
$$\sum_{r=0}^{n} P_r = 1$$

and
$$\sum_{r=0}^{n} rP_r = \mu = np$$

Thus
$$\sigma^2 \sum_{r=0}^{n} r^2 P_r - (np)^2 = \sum_{r=0}^{n} [r(r-1)+r]P_r - (np)^2$$

$$= \sum_{r=2}^{n} r(r-1) \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} + \sum_{r=0}^{n} rP_r - (np)^2$$

$$= n(n-1)p^2 \sum_{r=2}^{n} \frac{(n-2)!}{(r-2)!(n-r)!} p^{r-2}(1-p)^{n-r} + np - (np)^2$$

$$= n(n-1)p^2(p+q)^{n-2} + np - (np)^2$$

or
$$\sigma^2 = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq \tag{3.60}$$

Hence
$$\sigma = \sqrt{npq} \tag{3.61}$$

Summarizing, if X is a binomial random variable with parameters p and n, then the mean,

$$\mu = E(X) = np \tag{3.61a}$$

and the variance,

$$\sigma^2 = V(X) = np(1-p) \tag{3.61b}$$

A binomial distribution with $n = 20$ and $p = 0.10$ is shown in Fig. 3.4(a).

Cumulative binomial distribution are given in Appendix-*D*.

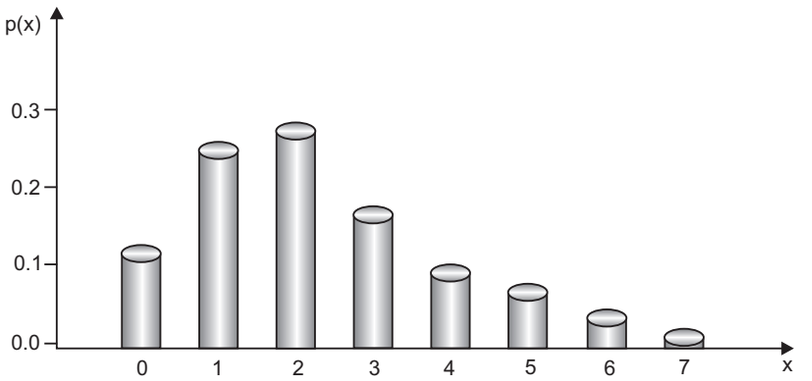


Fig. 3.4(a): Binomial pmf for n = 20, p = 0.10

Example E3.31

Let X denotes the number of mechanical components that are defective in a testing process and assume that X is a binomial random variable with $p = 0.001$. If 1000 of these components are tested, find the following:

(a) $P(X = 1)$

(b) $P(X \geq 1)$

(c) $P(X \leq 2)$

(d) mean and variance of X

SOLUTION:

We have from Eq. (3.50)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

(a) $P(X = 1) = \binom{1000}{1} (0.001)^1 (0.999)^{999} = 0.368$

(b) $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{1000}{1} (0.001)^1(0.999)^{999} = 0.632$

(c) $P(X \leq 2) = \binom{1000}{0}(0.001)^0(0.999)^{1000} + \binom{1000}{1}(0.001)^1(0.999)^{999}$

$$+\binom{1000}{2}(0.001)^2(0.999)^{998} = 0.920$$

(d) $E(X) = 1000(0.001) = 1$

$V(X) = 1000(0.001)\,(0.999) = 0.999$

Example E3.32

A professional basket player makes 80% of the free throws he tries. Assuming this percentage will hold true for future attempts, find the probability that in the next eight tries the number of free throws he will make is

(a) exactly 8

(b) exactly 5.

SOLUTION:

Here $n = 8$, $p = 0.80$, $q = 1 - p = 1 - 0.80 = 0.20$

(a) $P(\text{exactly } 8) = P(8) = \binom{n}{x} p^x(1-p)^{n-x} = \binom{8}{8} (0.80)^8(0.20)^0 = 1(0.167772)(1) = 0.1678$

(b) $P(\text{exactly } 5) = P(5) = \binom{n}{x} p^x(1-p)^{n-x} = \binom{8}{5} (0.80)^5(0.20)^3 = (56)(0.32768)(0.008) = 0.1468$

Example E3.33

A soft drink company conducted a taste survey before marketing a new soft drink. The results of the survey showed that 80% of the people like this soft drink. On a certain day, 8 customers bought it.

(a) let x denotes the number of customers in this sample of 8 who will like this soft drink. Using the binomial probabilities table, find the probability distribution of x and draw a graph of the probability distribution. Find the mean and standard deviation.

(b) Using the binomial distribution of part *(a)*, find the probability that exactly three of the eight customers will like the soft drink.

SOLUTION:

(a) Here $n = 8$, $p = 0.80$. From Appendix-*C*, for $n = 8$, and $p = 0.88$ we have x, $P(x)$ tabulated as in Table E3.33.

Table E3.33

x	P(x)
0	0.0000
1	0.0001
2	0.0011
3	0.0092
4	0.0459
5	0.1468
6	0.2936
7	0.3355
8	0.1678

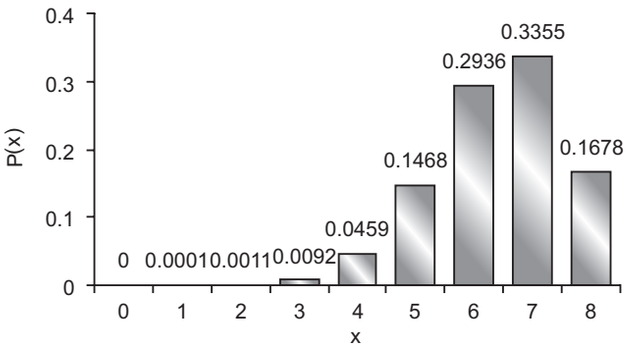


Fig. E3.33

The mean and standard deviation of x are (from Eqs. (3.61a) and (3.61b)

Mean

$$\mu = np = 8(0.80) = 6.4$$

Standard deviation

$$\sigma = \sqrt{npq} = \sqrt{8(0.80)(0.20)} = 1.1317$$

(b) P(exactly 3 customers like the soft drink) = $P(3) = 0.0092$

Example E3.34

A certain mechanical system contains 10 components. Assuming that the probability of each individual component will fail is 0.2 and the components fail independently of each other. Given that at least one of the components has failed, what is the probability that at least 2 of the components have failed?

SOLUTION:

The number X of components that fail will have a binomial distribution with parameters $n = 10$ and $p = 0.2$. Hence,

$$P(X \geq 2 \mid X \geq 1|) = \frac{P(X \geq 2)}{P(X \geq 1)} = \frac{1 - P(X = 0) - P(X = 1)}{1 - P(X = 0)}$$

From Appendix-*C*, for $n = 10$, $p = 0.2$, we have

$$P(X \geq 2 \mid X \geq 1|) = \frac{1 - 0.1074 - 0.2684}{1 - 0.1074} = \frac{0.6242}{0.8926} = 0.6993$$

Example E3.35

(a) If a fair coin is (probability of heads equals 1/2) is tossed independently 10 times. Use the table in Appendix-*C* of the binomial distribution to find the probability that strictly more heads are obtained than tails.

(b) Suppose that the probability that a certain experiment will be successful is 0.3, and let X denotes the number of successes that are obtained in 15 independent performances of the experiment. Use the table in Appendix-*C* for the binomial distribution to determine the value of $P(6 \leq X \leq 9)$.

SOLUTION:

(a) Let X be the number of heads obtained. More heads than tails are obtained if $X \in \{6, 7, 8, 9, 10\}$. The probability of this event is the sum of the numbers in the binomial table in Appendix-*C*. Corresponding to

$P = 0.5$

and $n = 10$

for $x = 6, 7, 8, 9, 10 = 0.2051 + 0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.37695.$

By the symmetry of this binomial distribution, we can also compute the sum as

$$(1 - P(X = 5))/2 = (1 - 0.2461)/2 = 0.37695.$$

(b) From the table in Appendix-*C* for the binomial distribution with parameters $n = 15$ and $p = 0.3$, that

$$P(6 \leq X \leq 9) = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9)$$
$$= 0.1472 + 0.0811 + 0.0348 + 0.0116 = 0.2747$$

Example E3.36

According to a particular survey, the probability is 0.40 that a traffic fatality involves an intoxicated or alcohol-impaired driver or non-occupant. In 8 traffic fatalities, find the probability that the number, A, which involve an intoxicated or alcohol-impaired driver or non-occupant is

(a) exactly 3; at least 3; at most 3

(b) between 2 and 4, inclusive

(c) find and interpret the mean of the random variable A

(d) obtain the standard deviation of A.

SOLUTION:

Here, $n = 8$ and $p = 0.40$. Thus $q = 1 - p = 1 - 0.40 = 0.60$

From Eq. (3.50), we have

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Refer to the table in Appendix-*C* for $n = 8$ and $p = 0.40$

(a) $P(\text{exactly } 3) = \binom{8}{3} (0.4)^3(0.6)^5 = 0.0241$

$$P(3 \text{ or more}) = 1 - P(0 \text{ or } 1 \text{ or } 2) = 1 - \binom{8}{0} (0.4)^0(0.6)^8 - \binom{8}{1}(0.4)^1(0.6)^7 - \binom{8}{2} (0.4)^2(0.6)^6$$

$$= 1 - 0.0168 - 0.0896 - 0.2090 = 0.6846$$

$$P(\text{at most } 3) = P(0) + P(1) + P(2) + P(3)$$

$$= \binom{8}{0} (0.4)^0(0.6)^8 - \binom{8}{1} (0.4)^1(0.6)^7 - \binom{8}{2}(0.4)^2(0.6)^6 + \binom{8}{3} (0.4)^3(0.6)^5$$

$$= 0.0168 + 0.0896 + 0.2090 + 0.2787 = 0.5941$$

(b) $P(2 \le A \le 4) = P(2) + P(3) + P(4) = \binom{8}{2} (0.4)^2(0.6)^6 - \binom{8}{3} (0.4)^3(0.6)^5 - \binom{8}{4} (0.4)^4(0.6)^4$

$$= 0.2092 + 0.2787 + 0.2322 = 0.7201 \approx 0.72$$

(c) The mean of A is $\mu = np = 8(0.4) = 3.2$. On average, of 8 traffic fatalities, 3.2 will involve an intoxicated or alcohol-impaired driver or non-occupant.

(d) Standard deviation $= \sqrt{\sigma^2} = \sqrt{np(1-p)} = \sqrt{8(0.4)(0.6)} = \sqrt{19.2} = 4.3418$

Example E3.37

According to a particular survey, 14.9% of those who have received a doctor's degree in engineering are blacks. Suppose that 6 people who have received their doctor's degree in engineering are randomly selected. Find the probability that

(a) exactly 2 are black

(b) exactly 4 are black

(c) at least 2 are black

(d) find the probability distribution of the number of blacks in a sample of 6 persons who have received their doctor's degree in engineering

(e) why is the probability distribution obtained in part *(d)* only approximately correct? What is the exact distribution called?

SOLUTION:

Here, $n = 6$ and $p = 0.149$ and $q = 1 - p = 1 - 0.149 = 0.851$

From Eq. (3.50), we have

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

(a) $P(2) = \binom{6}{2} (0.149)^2(0.851)^4 = 0.1747$

(b) $P(4) = \binom{6}{4} (0.149)^4(0.851)^2 = 0.0054$

(c) $P(\text{at least 2}) = 1 - P(0 \text{ or } 1) = 1 - \binom{6}{0} (0.149)^0(0.851)^6 - \binom{6}{1} (0.149)^1(0.851)^5$

$$= 1 - 0.3798 - 0.3990 = 0.2212$$

(d) $P(X = x) = \binom{6}{x} (0.149)^x(0.851)^{6-x}$ for $x = 0, 1, 2, \ldots, 6$.

Applying this formula for each value of x gives the results in the Table E3.37.

Table E3.37

x	P(X = x)
0	0.3798
1	0.3990
2	0.1747
3	0.0408
4	0.0054
5	0.0004
6	0.0000

(e) The sampling was actually done without replacement, so the trials are not independent and the success probability changes very slightly from trial to trial. The exact probability distribution is called a hypogeometric distribution.

Example E3.38

Suppose the probability is 0.67 that the favorite in a horse race will finish in the money (first, second, or third place). In the next 5 races, what is the probability that the favorite finishes in the money

(a) exactly 2 times

(b) exactly 4 times

(c) at least 4 times

(d) between 2 and 4 times, inclusive

(e) find the probability distribution of the random variable, X, the number of times the favorite finishes in the money in the next 5 races

(f) identify the probability distribution of X as right skewed, symmetric, or left skewed without checking its probability distribution or its histogram

(g) Draw a probability histogram for X

(h) find the mean and standard deviation of the random variable X.

SOLUTION:

Refer to the table in Appendix-*C*.

Here $n = 5$, $p = 0.67$ and $q = 1 - p = 1 - 0.67 = 0.33$

From Eq. (3.50) we have

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$P(0) = \binom{5}{0} (0.67)^0 (0.33)^5 = 0.004$$

$$P(1) = \binom{5}{1} (0.67)^1 (0.33)^4 = 0.040$$

$$P(2) = \binom{5}{2} (0.67)^2 (0.33)^3 = 0.161$$

$$P(3) = \binom{5}{3} (0.67)^3 (0.33)^2 = 0.328$$

$$P(4) = \binom{5}{4} (0.67)^4 (0.33)^1 = 0.332$$

$$P(5) = \binom{5}{5} (0.67)^5 (0.33)^0 = 0.135$$

(a) $P(2) = 0.161$

(b) $P(4) = 0.332$

(c) $P(X \geq 4) = P(4) + P(5) = 0.332 + 0.135 = 0.467$

(d) $P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = 0.161 + 0.328 + 0.332 = 0.821$

(e)

x	P(X = x)
0	0.004
1	0.040
2	0.161
3	0.328
4	0.332
5	0.135

(f) Left-skewed

(g) See Fig. E3.38.

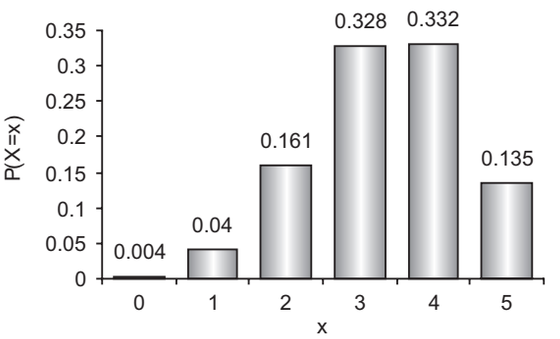


Fig. E3.38

x	P(X = x)	xP(X = x)	x^2	$x^2P(X = x)$
0	0.004	0.000	0	0.000
1	0.040	0.040	1	0.040
2	0.161	0.322	4	0.644
3	0.328	0.984	9	2.952
4	0.332	1.328	16	5.312
5	0.135	0.675	25	3.375
Σ		3.349		12.323

(h) Mean = $\mu = 3.349$ (from the above table)

Variance,

$$\sigma^2 = 12.323 - 3.349^2 = 1.107$$

Hence, standard deviation = $\sqrt{1.107} = 1.052$

Alternate calculation:

$$\mu = np = 5(0.67) = 3.35$$

$$\sigma^2 = np(1 - p) = 5(0.67)(0.33) = 1.1055$$

or $$\sigma = \sqrt{1.1055} = 1.051$$

3.3.5 Poisson Distribution

The Poisson distribution, named after the French mathematician Simeon D. Poisson is another important probability distribution of a discrete random variable that has many applications. The Poisson distribution is applied to experiments with random and independent occurrences.

Given an interval of real numbers, assume events occur at random throughout the interval. If the interval can be partitioned into subintervals of small enough length such that

1. the probability of more than one event in a subinterval is zero
2. the probability of one event in a subinterval is the same for all subintervals and proportional to the length of the subinterval, and
3. the event in each subinterval is independent of other subintervals, the random experiment is called a *Poisson process*.

Independence of occurrences means that one occurrence (or non-occurrence) of an event does not influence the successive occurrence or nonoccurrences of that event. The occurrences are always considered with respect to an interval. The interval may be a time interval, a space interval, or a volume interval. The actual number of occurrences within an interval is random and independent. If the average number of occurrences for a given interval is known, then by using the Poisson probability distribution we can determine the probability of a certain number of occurrences, x, in that interval. Note that the number of actual occurrences in an interval is denoted by x.

Conditions to apply Poisson probability distribution

The following three conditions must be satisfied to apply the Poisson probability distribution:

1. x is a discrete random variable
2. The occurrences are random
3. The occurrences are independent

The following are a few examples of discrete random variables for which the occurrences are random and independent.

Some of the phenomena that follow the Poisson distribution are:

1. counts of flaws in castings
2. the number of vehicles on a highway
3. the number of customers visiting a bank
4. the number of accidents that occur on a given highway during a period of time
5. the number of telephone calls
6. counts of power outages
7. counts of atomic particles emitted from a specimen.

The random variable X that equals the number of events in the interval is a *Poisson random variable* with parameter $0 < \lambda$, and the probability mass function (pmf) of X is

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}; \qquad x = 0, 1, 2, \ldots \tag{3.62}$$

The Poisson distribution is made up of a series of terms:

$$e^{-1}, e^{-1}\lambda, \frac{e^{-\lambda}\lambda^{2}}{2!}, \frac{e^{-\lambda}\lambda^{3}}{3!}, \ldots.$$

representing respectively, the probabilities of occurrence of 0, 1, 2, 3, 4, etc. events, where e is the base of the natural logarithm and λ is the mean frequency of occurrence. The sum of the probabilities is one (1) because

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^{k}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k}}{k!} \tag{3.62a}$$

and the summation on the right-hand side of the equation (3.62*a*) is recognized to be Taylor's expansion of e^{x} evaluated at λ. Hence, the summation equals e^{λ} and the right-hand side equals $e^{-\lambda}e^{\lambda} = 1$. Poisson distribution for $\lambda = 3$ is shown in Fig. 3.4(*b*).

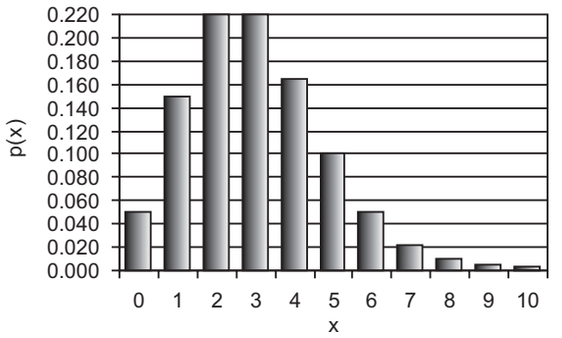


Fig. 3.4(b)

3.3.5.1 Derivation from Binomial Distribution

The Poisson distribution can also be deduced from the binomial distribution, provided that n is large ($\to\infty$), p is very small ($\to 0$), and np is finite. It was shown earlier that in n trials the probability of an event succeeding r times is

$$P_r = \frac{n!}{r!(n-r)!} p^r q^{n-r} \tag{3.63}$$

When n is large compared with r,

$$\frac{n!}{(n-r)!} = n(n-1)(n-2)...(n-r+1) \simeq n^r$$

Therefore, the probability of r successes becomes

$$P_r = \frac{n^r}{r!} p^r q^{n-r} \tag{3.64}$$

Now if p is very small and r is not large,

$$q^r = (1-p)r \simeq 1$$

and $$q^{n-r} \simeq q^n = (1-p)^n$$

Hence $$P_r = \frac{(np)^r}{r!}(1-p)^n = \frac{(np)^r}{r!}\left[1 - np + \frac{n(n-1)(-p)^2}{2!} + \frac{n(n-1)(n-2)(-p)^3}{3!} + \cdots\right]$$

$$\simeq \frac{(np)^r}{r!}\left[1 - np + \frac{(np)^2}{2!} - \frac{(np)^3}{3!} + \cdots\right]$$

Thus $$P_r = \frac{(np)^r}{r!} e^{-np} \tag{3.65}$$

This, then, is the probability of r successes in n trials.

3.3.5.2 Mean and Standard Deviation

The mean number of occurrences of an event per unit of time (or space) is

$$\mu = np \tag{3.66}$$

and the standard deviation of the numbers of events is

$$\sigma = \sqrt{np} \tag{3.67}$$

Thus the mean and variance are equal to one another:

$$\mu = \sigma^2 = np \tag{3.68}$$

Equations (3.66) and (3.68) will now be derived.

$$\text{mean} = E(r) = \sum_{r=0}^{\infty} rP_r = \sum_{r=0}^{\infty} \frac{re^{-\lambda}\lambda^r}{r!} = 0 + \lambda e^{-\lambda} + 2\frac{\lambda^2 e^{-\lambda}}{2!} + 3\frac{\lambda^3 e^{-\lambda}}{3!} + \cdots$$

$$= \lambda e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right] = \lambda e^{-\lambda} e^{\lambda} = \lambda \tag{3.69}$$

The variance σ^2 is given by

$$\sigma^2 = E(r-\lambda)^2 = \sum_{r=0}^{\infty}(r-\lambda)^2 P_r$$

$$\sigma^2 = \sum_{r=0}^{\infty} r^2 P_r - 2\lambda \sum_{r=0}^{\infty} rP_r + \lambda^2 \sum_{r=0}^{\infty} P_r$$

Now $$-2\lambda \sum_{r=0}^{\infty} rP_r = -2\lambda(\lambda) = -2\lambda^2$$

$$\lambda^2 \sum_{r=0}^{\infty} P_r = \lambda^2(1) = \lambda^2$$

and $$\sum_{r=0}^{\infty} r^2 P_r = \sum_{r=0}^{\infty}[r(r-1)+r]P_r = \sum_{r=0}^{\infty} r(r-1)P_r + \lambda = \left[0+0+2\frac{\lambda^2 e^{-\lambda}}{2!}+6\frac{\lambda^3 e^{-\lambda}}{3!}+\cdots\right]+\lambda$$

$$= \lambda^2 e^{-\lambda}\left[1+\lambda+\frac{\lambda^2}{2!}+\cdots\right]+\lambda = \lambda^2 e^{-\lambda}(e^{\lambda})+\lambda = \lambda^2+\lambda$$

Thus, collecting the terms, we have

$$\sigma^2 = \lambda^2 + \lambda - 2\lambda^2 + \lambda^2$$

or $$\sigma^2 = \lambda \tag{3.70}$$

Hence, if X is a Poisson random variable with parameter λ, then

the mean $\mu = E(X) = \lambda$, and

the variance $\sigma^2 = V(X) = \lambda^2$

Note that the Poisson distribution contains only one parameter, *np*, the mean occurrence of an event, and we do not know the value of *n*. In the binomial distribution we know the number of times an event occurs and the number of times an event does not occur.

Cumulative Poisson probabilities are given in Appendix-*D*. A plot of the Poisson distribution for $\lambda = 3$ is shown in Fig. 3.5. This Poisson distribution has a very long tail to the right (the distribution is skewed).

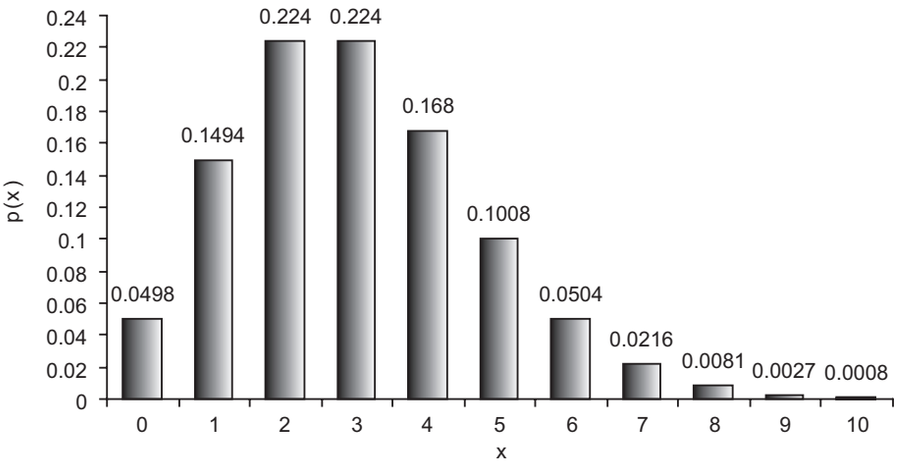


Fig. 3.5: Poisson distribution for $\lambda = 3$

Example E3.39

A survey found that 1.5% of occupied housing units have 7 or more people living within. Use the Poisson distribution to determine the approximate probability that, of 200 randomly selected occupied housing units, there are

(a) none with 7 or more persons

(b) 3 or more with 7 or more persons

SOLUTION:

We use $\lambda = np = 200(0.15) = 3.0$. We also note that $n \geq 100$ and $np \leq 10$. Hence, apply Eq. (3.62),

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

(a) $P(X = 0) = \dfrac{e^{-3.0}(3.0)^0}{0!} = 0.0498$

(b) P(at least 3) = 1 – $P(X = 0$ or 1 or 2)

$$= 1 - \left[e^3 \frac{3^0}{0!} + e^{-3} \frac{3^1}{1!} + e^{-3} \frac{3^2}{2!} \right] = 1 - [0.0498 + 0.1494 + 0.2240] = 0.5768$$

Example E3.40

Assume the number of errors along a magnetic recording surface is a Poisson random variable with a mean of one error every 10^5 bits. A sector of data consists of 5000 eight-bit bytes. Find

(a) the probability of more than one error in a sector

(b) the mean number of sectors until an error is found.

SOLUTION:

Let x denotes the number of errors in a sector. Then, X is a Poisson random variable with $\lambda = 0.4$.

(a) $P(X > 1) = 1 - P(X < 1) = 1 - e^{-0.4} - e^{-0.4}(0.4) = 1 - 0.6703 - 0.2681 = 0.06157$

(b) Let Y denotes the number of sectors until an error is found. Then, Y is a geometric random variable and

$P = P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-0.4} = 1 - 0.6703 = 0.3297$

$$E(Y) = \frac{1}{p} = \frac{1}{0.3297} = 3.03306$$

Example E3.41

The probability that an individual recovers from an illness in a 2-week time period without medical treatment is 0.1. If 20 independent individuals suffering from this illness are treated with a medicine and 4 recover in a 2-week period. If the medicine has no effect, what is the probability that 4 or more people recover in a 2-week time period?

SOLUTION:

Let x denotes the number of individuals that recover in 2-week time period. Assume the individuals are independent, then, X is binomial random variable with $n = 20$ and $p = 0.1$.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - [P(X = 0) + P(x = 1) + P(x = 2) + P(X = 3)]$$

$$= 1 - \left[\binom{20}{0} (0.1)^0 (0.9)^{20} + \binom{20}{1} (0.1)^1 (0.9)^{19} + \binom{20}{2} (0.1)^2 (0.9)^{18} + \binom{20}{3} (0.1)^3 (0.9)^{17} \right]$$

$$= 1 - [0.1216 + 0.2702 + 0.2852 + 0.1901] = 0.1330$$

Example E3.42

A survey found that the traffic flowing through an intersection with an average of 3 cars per 30 seconds. Assume the traffic flow can be modeled as a Poisson distribution.

(a) find the probability of no cars through the intersection within 30 seconds

(b) find the probability of 3 or more cars through the intersection within 30 seconds

(c) find the minimum number of cars through the intersection so that the probability of this number or fewer cars in 30 seconds is at least 90%

(d) if the variance of the number of cars through the intersection per minute is 20, is the Poisson distribution appropriate?

SOLUTION:

(a) $\lambda = 3$ cars/30 seconds

we have $f(x) = \dfrac{e^{-\lambda} \lambda^x}{x!}$

$$P(X = 0) = \frac{e^{-3} 3^0}{0!} = 0.0498$$

(b) $P(X \geq 3) = 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$

$$= 1 - \left[\frac{e^{-3} 3^0}{0!} + \frac{e^{-3} 3^1}{1!} + \frac{e^{-3} 3^2}{2!} \right] = 1 - [0.0498 + 0.1494 + 0.2240] = 0.5768$$

(c) $P(X \leq x) \geq 0.9$ or $x = 5$

(d) $\sigma^2 = \lambda = 6$. This is not appropriate.

Example E3.43

A student's campus newspaper in a particular university contains an average of 1.2 typographical errors per page.

(a) find the probability that a randomly selected page of this newspaper will contain exactly 4 typographical errors using the Poisson formula

(b) find the probability that the number of typographical errors on a randomly selected page will be

(i) more than 3

(ii) less than 4

Use the Poisson probabilities table in Appendix-*D*.

SOLUTION:

Let x be the number of typographical errors on a randomly selected page of this newspaper. Since it contains an average of 1.2 typographical errors per page, $\lambda = 1.2$. We have Eq. (3.62)

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

(a) $P(\text{exactly } 4) = P(x = 4) = \dfrac{\lambda^{x}e^{-\lambda}}{x!} = \dfrac{(1.2)^{4}e^{-1.2}}{4!} = 0.0260$

(b) *(i)* $P(\text{more than } 3) = P(x > 3) = P(4) + P(5) + P(6) + P(7) = 0.0260 + 0.0062 + 0.0012 + 0.0002 = 0.0336$

(ii) $P(\text{less than } 4) = P(x < 4) = P(0) + P(1) + P(2) + P(3) = 0.3012 + 0.3614 + 0.2169 + 0.0867 = 0.9662$

Example E3.44

In a statistics course final examination, 15% of the students fail.

(a) find the probability that in a random sample of 100 students in that statistics class who took the final examination exactly 20 will fail. Use the Poisson formula.

(b) find the probability that the number of students who fail this statistics final examination in a randomly selected 100 students is

(i) at most 9

(ii) 10 to 16

(iii) at least 20

Use the Poisson probabilities table in Appendix-*D*.

SOLUTION:

Let x be the number of students in a random sample of 100 who fail the final examination in statistics. Since, an average, 15% of the students fail the examination, $\lambda = 0.15(100) = 15$.

(a) $P(\text{exactly 20 fail}) = P(x = 20) = \dfrac{e^{-\lambda}\lambda^{x}}{x!} = \dfrac{(e^{-15})(15)^{20}}{20!} = 0.0418$

(b) *(i)* $P(\text{at most 9 fail}) = P(x \le 9) = P(0) + P(1) + P(2) + P(3) + \cdots + P(9)$

$= 0.0000 + 0.0000 + 0.0000 + 0.0002 + 0.0006 + 0.0019 + 0.0048 + 0.0104 + 0.0194 + 0.0324$

$= 0.0697$

(ii) $P(10 \text{ to } 16 \text{ fail}) = P(10 \le x \le 16) = P(10) + P(11) + P(12) + P(13) + P(14) + P(15) + P(16)$

$= 0.0486 + 0.0663 + 0.0829 + 0.0956 + 0.1024 + 0.1024 + 0.0960 = 0.5942$

(iii) $P(\text{at least 20}) = P(x \ge 20) = P(20) + P(21) + P(22) + \cdots + P(39)$

$= 0.0418 + 0.0299 + 0.0204 + 0.0133 + 0.0083 + 0.0050 + 0.0029 + 0.0016 + 0.0009 + 0.0004$
$+ 0.0002 + 0.0001 + 0.0001 + 0.0000 + \cdots + 0.0000 = 0.1249$

Example E3.45

A hardware store in a big city receives an average of 9.8 telephone calls per hour.

(a) find the probability that exactly 6 telephone calls will be received at this store during a certain hour. Use the Poisson formula.

(b) find the probability that the number of telephone calls received at this store during a certain hour will be

(i) less than 8

(ii) more than 12

(iii) 5 to 8

Use the Poisson distribution table in Appendix-*D*.

SOLUTION:

Let x be the number of telephone calls received at this hardware store during a certain hour. Since the average number of telephone calls per hour is 9.8, $\lambda = 9.8$.

(a) $P(\text{exactly } 6) = P(x = 6) = \dfrac{e^{-\lambda}\lambda^{x}}{x!} = \dfrac{(e^{-9.8})(9.8)^{6}}{6!} = 0.0677$

(b) *(i)* $P(\text{less than } 8) = P(x < 8) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7)$

$= 0.0001 + 0.0005 + 0.0027 + 0.0087 + 0.0213 + 0.0418 + 0.0682 + 0.0955 = 0.2388$

(ii) $P(\text{more than } 12) = P(x > 12) = P(13) + P(14) + \text{P}(15) + \cdots$

$= 0.0685 + 0.0479 + 0.0313 + 0.0192 + 0.01111 + 0.0060 + 0.0031 + 0.0015 + 0.0007 + 0.0003$

$+ 0.0001 + 0.0001 = 0.1898$

(iii) $P(5 \text{ to } 8) = P(5 \le x \le 8) = P(5) + P(6) + P(7) + P(8) = 0.0418 + 0.0682 + 0.0955 + 0.1170 = 0.3225$

Example E3.46

An average of 0.7 accidents occur per day in a large city.

(a) find the probability that no accidents will occur in that city on a given day

(b) write the probability distribution of x, where x denotes the number of accidents that will occur in that city on a given day

(c) find the mean, variance and standard deviation of the probability distribution developed in part *(b)*.

SOLUTION:

x = number of accidents on a given day in that city.

Since the average number of accidents per day is 0.7, $\lambda = 0.7$.

(a) $P(\text{no accidents}) = P(x = 0) = \dfrac{e^{-\lambda}\lambda^{x}}{x!} = \dfrac{e^{-0.7}(0.7)^{0}}{0!} = 0.4966$

(b) From the table in Appendix-*D*, we have

x	P(x)
0	0.4966
1	0.3476
2	0.1217
3	0.0284
4	0.0050
5	0.0007
6	0.0001
7	0.0000

(c) The mean, variance and standard deviation are:

mean $\mu = \lambda = 1.7$

variance $\sigma^2 = \lambda = 0.7$ and

standard deviation, $\sigma = \sqrt{\lambda} = \sqrt{0.7} = 0.8366$

Example E3.47

The number of male mates of a queen bee was found to have a Poisson distribution with parameter $\lambda = 2.7$. Find the probability that the number, *N*, of male mates of a queen bee is

(a) exactly 2

(b) at most 2

(c) between 1 and 3, inclusive

(d) on average, how many male mates does a queen bee have?

(e) develop a table of probabilities for the random variable, *N*. Compute the probabilities until they are zero to 4 decimal places

(f) draw a histogram of the probabilities in part *(c)*.

SOLUTION:

(a) $P(N = 2) = e^{-2.7}\dfrac{2.7^2}{2!} = 0.2450$

(b) $P(N \le 2) = P(0 \text{ or } 1 \text{ or } 2) = e^{-2.7}\dfrac{2.7^0}{0!} + e^{-2.7}\dfrac{2.7^1}{1!} + e^{-2.7}\dfrac{2.7^2}{2!} = 0.0672 + 0.1815 + 0.2450 = 0.4937$

(c) $P(1 \le N \le 3) = P(1 \text{ or } 2 \text{ or } 3) = e^{-2.7}\dfrac{2.7^1}{1!} + e^{-2.7}\dfrac{2.7^2}{2!} + e^{-2.7}\dfrac{2.7^3}{3!} = 0.1815 + 0.2450 + 0.2205 = 0.6470$

(d) On average, the number of mates of a queen bee is

$\mu = \lambda = 2.7$

(e) Refer to Appendix-*D*. The results are shown in Table E3.47.

Table E3.47

N	P(N = n)
0	0.0672
1	0.1815
2	0.2450
3	0.2205
4	0.1488
5	0.0804
6	0.0362
7	0.0139
8	0.0047
9	0.0014
10	0.0004
11	0.0001
12	0.0000

(f) See Fig. E3.47

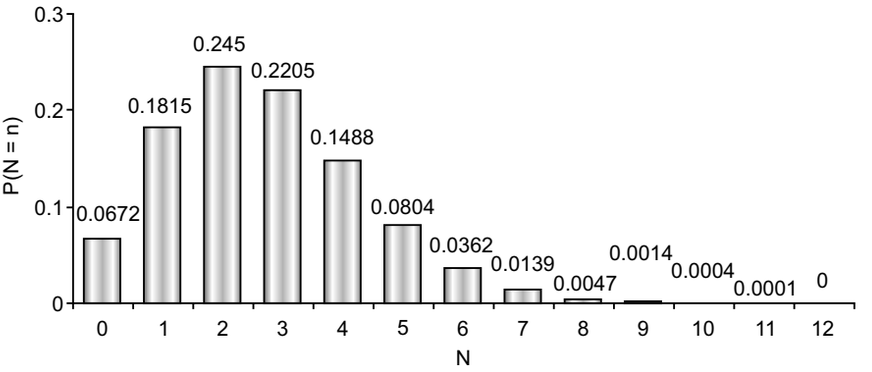


Fig. E3.47

3.4 CONTINUOUS PROBABILITY DISTRIBUTIONS

3.4.1 The Normal Distribution

In everyday life, people deal with and use a wide variety of variables. Some of these variables — such as heights of people, scores in final examinations, aptitude test scores, TOEFL (Test of English as a Foreign Language), and GRE (Graduate Record Examination) share an important characteristic: their distributions have roughly the shape of a normal curve, that is, a special type of bell-shaped curve like one shown in Fig. 3.6.

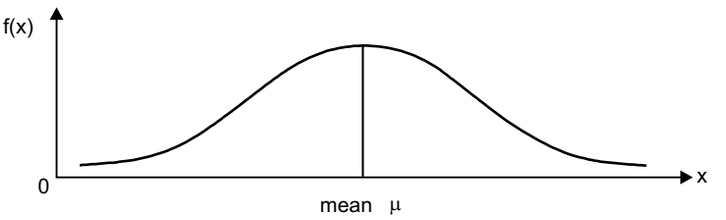


Fig. 3.6 Normal curve

The normal distribution is in many respects the cornerstone of statistics. A random variable X is said to have a normal distribution with mean $\mu(-\infty < \mu < \infty)$ and variance $\sigma^2 > 0$ if it has the density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)\sigma]^2} \qquad -\infty < X < \infty \tag{3.71}$$

The distribution is illustrated graphically in Fig. 3.6. The normal distribution is used so extensively that the shorthand notation $X \sim N(\mu, \sigma^2)$ is often used to indicate that the random variable X is normally distributed with mean μ and variance σ^2.

A variable is said to be a normally distributed variable or to have a *normal distribution* if its distribution has the shape of a normal curve.

If a variable of a population is normally distributed and is the only variable under consideration, the general practice is to assume that the population is *normally distributed* or that it is a *normally distributed population*. In reality, a distribution is unlikely to have exactly the shape of a normal distribution curve. If a

variable's distribution is shaped roughly like a normal curve, we can consider that the variable is an *approximately normally distributed variable* or that it has *approximately a normal distribution*.

A normal distribution (and hence a normal curve) is completely determined by the mean and standard deviation. Hence, two normally distributed variables having the same mean and standard deviation must have the same distribution. We identify a normal curve by stating the corresponding mean and standard deviation and calling those to *parameters* of the normal curve.

A normal distribution is symmetric about and centred at the mean of the variable, and its spread depends on the standard deviation of the variable. The larger the standard deviation, the flatter and more spread out is the distribution.

In summary, the normal curve associated with a normal distribution is bell shaped, centered at μ, and close to the horizontal axis outside the range from $\mu - 3\sigma$ to $\mu + 3\sigma$ as shown in Figs. 3.6 and 3.6(*a*).

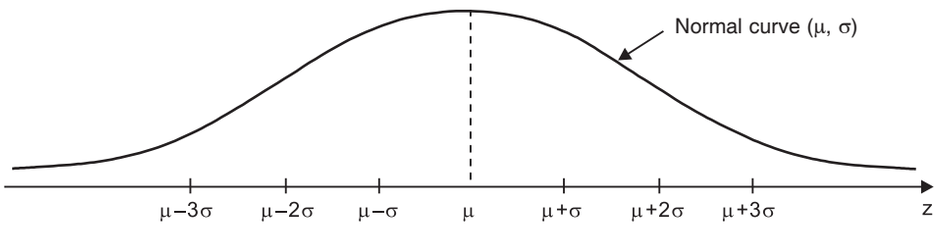


Fig. 3.6(a)

Random variables with different means and variances can be modeled by normal probability functions with appropriate choices of the centre and width of the curve. The value of $E(X) = \mu$ determines the centre of the probability density function and the value of $V(X) = \sigma^2$ determines the width. Figure 3.6(*c*) illustrates several normal probability density functions with selected values of m and σ^2. Each one has the characteristic symmetric bell-shaped curve, but the centers and dispersions differ.

Definition of Normal Distribution

A random variable X with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x-\mu}{2\sigma^2}\right)} \qquad -\infty < x < \infty$$

is a *normal random variable* with parameters μ, where $-\infty < x < \infty$, and $\sigma > 0$.

Also $E(X) = \mu$ and $V(X) = \sigma^2$

and the notation $N(\mu, \sigma^2)$ is used to denote the distribution. The mean and variance of X are equal to μ and σ^2, respectively.

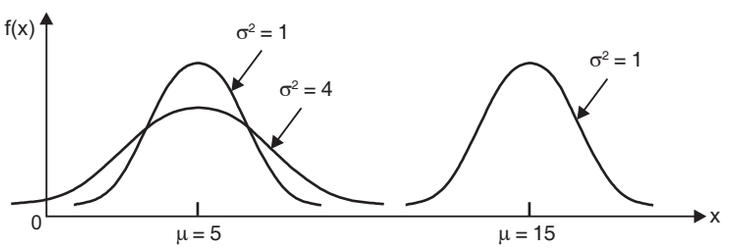


Fig. 3.6(b): Normal probability density functions for selected values of the parameters μ and σ^2

3.4.1.1 Properties of the Normal Distribution

The normal distribution has several important properties:

1. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
2. $f(x) \geq 0$ for all x
3. $\lim_{x\to\infty} f(x) = 0$ and $\lim_{x\to\infty} f(x) = 0$
4. $f[(x + \mu)] = f[-(x - \mu)]$. The density is symmetric about μ.
5. The maximum value of $f(x)$ occurs at $x = \mu$.
6. The points of inflection of $f(x)$ are at $x = \mu \pm \sigma$.

Also, $P(a \leq x \leq b) = \int_a^b f(x)\,dx$ = area under $f(x)$ from a to b for any a and b.

Property 1 may be demonstrated as follows. Let $y = (x - \mu)/\sigma$ in Eq. (3.71) and denote the integral as I. That is,

$$I = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(1/2)y^2}\,dy \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(1/2)x^2}\,dx \tag{3.72}$$

on changing to polar coordinates with the transformation of variables $y = r \sin\theta$ and $z = r \cos\theta$, the integral becomes

$$I^2 = \frac{1}{2\pi}\int_0^{\infty}\int_0^{2\pi} re^{-(1/2)r^2}\,d\theta\,dr = \int_0^{\infty} re^{-(1/2)r^2}\,dr = 1 \tag{3.73}$$

3.4.1.2 Mean and Variance of the Normal Distribution

The mean of the normal distribution may be determined easily. Since

$$E(X) = \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma^2]}\,dx \tag{3.74}$$

and if we let $z = (x - \mu)/\sigma$, we obtain

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}(\mu + \sigma z)e^{-z^2/2}\,dz \tag{3.75}$$

$$= \mu = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\,dz + \sigma\frac{1}{\sqrt{2\pi}} ze^{-z^2/2}\,dz \tag{3.76}$$

Since the integrand of the first integral is that of a normal density with $\mu = 0$ and $\sigma^2 = 1$, the value of the first integral is one. The second integral has value zero, that is,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} ze^{-z^2/2}\,dz = \frac{1}{\sqrt{2\pi}} ze^{-z^2/2}\Big|_{-\infty}^{\infty} = 0$$

and thus

$$E(X) = \mu[1] = \sigma[0] = \mu \tag{3.77}$$

To find the variance we must evaluate

$$V(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2} dx$$

And letting $z = (x - \mu)/\sigma$, we obtain

$$V(X) = \int_{-\infty}^{\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \sigma^2 \left[\int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz \right]$$

$$= \sigma^2 \left[\frac{-ze^{-z^2/2}}{\sqrt{2\pi}} \Bigg|_{-\infty}^{\infty} \right] + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \sigma^2 (0+1)$$

so that $$V(X) = \sigma^2 \tag{3.78}$$

In summary the mean and variance of the normal density given in Eq. (3.78) are μ and σ^2, respectively.

3.4.1.3 The Cumulative Normal Distribution

The distribution function F is

$$F(x) = P(X \le x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2} du \tag{3.79}$$

It is impossible to evaluate this integral without resorting to numerical methods, and even then the evaluation would have to be accomplished for each pair (μ, σ^2). However, a simple transformation of variables, $z = (X - \mu)/\sigma$, allows the evaluation to be independent of μ and σ.

That is,

$$F(X) = P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma} \right) = \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_{-\infty}^{(x-\mu)/\sigma} \varphi(z)\, dz = \phi\left(\frac{x-\mu}{\sigma} \right) \tag{3.80}$$

3.4.1.4 The Standard Normal Distribution

The probability distribution in Eq. (3.80) above,

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < x < \infty$$

is a normal distribution with mean 0 and variance 1; that is, $Z \sim N(0, 1)$ and we say that Z has a *standard normal distribution*. A graph of the probability density function is shown in Fig. 3.7.

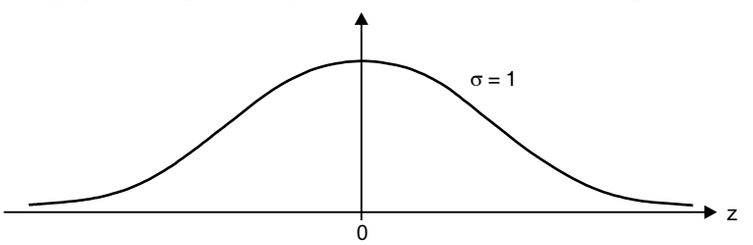


Fig. 3.7: The standard normal distribution

The corresponding distribution function is ϕ, where

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} du \tag{3.81}$$

where $$z = \frac{x - \mu}{\sigma}$$

and this function has been well tabulated.

A few useful results for a normal distribution are summarised below and shown in Fig. 3.7(*a*). For random variable,

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$

In other words, any normally distributed variable has the following properties:

1. 68.27% of all possible observations lie within one standard deviation to either side of the mean, that is, between $\mu - \sigma$ and $\mu + \sigma$.
2. 95.45% of all possible observations lie within two standard deviation to either side of the mean, that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$.
3. 99.73% of all possible observations lie within three standard deviation to either side of the mean, that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$.

These properties are shown in Fig. 3.7(*b*).

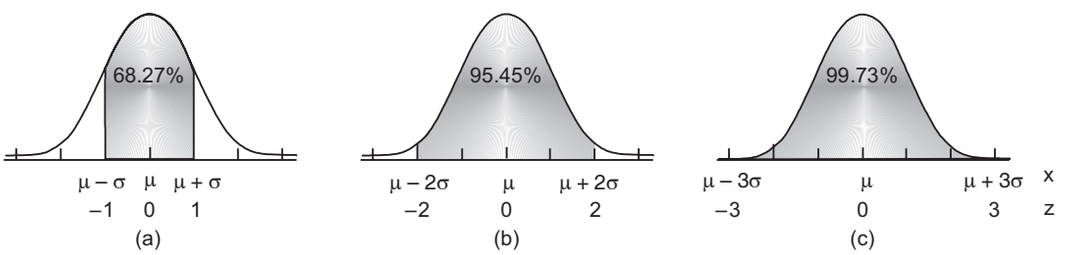


Fig. 3.7(a)

In addition, from the symmetry of $f(x)$, $P(X > \mu) = P(X < \mu) = 0.5$. The above model assigns some probability to each interval of the real line since $f(x)$ is positive for all x. The probability density function decreases as x moves further from μ. As a result, the probability that a measurement falls far from μ is small, and at some distance from μ the probability of an interval can be approximated as zero.

For a normally distributed variable, the percentage of all possible observations that (i.e., within any specified range) equals the corresponding area under its associated normal curve, expressed as a percentage. This result holds approximately for a variable that is approximately normally distributed.

A normally distributed variable having mean 0 and standard deviation 1 is said to have the *standard normal distribution*. Its associated normal curve is called the *standard normal curve*, which is shown in Fig. 3.7(*c*).

The area under a normal probability density function beyond 3σ from the mean is quite small. Since more than 0.9973 of the probability of a normal distribution is within the interval $(\mu - 3\sigma, \mu + 3\sigma)$, 6σ is often referred to as the *width* of a normal distribution. Advanced integration methods can be used to show that the area under the normal probability density function from $-\infty < x < \infty$ is 1.

3.4.1.5 Problem-Solving Procedure

If X is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, the random variable

$$Z = \frac{X - \mu}{\sigma}$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is, Z is a standard normal random variable.

Suppose X is a normal random variable with mean μ and variable σ^2. Then

$$P(X \le x) = P\left(\frac{X - \mu}{\sigma} \le \frac{x - \mu}{\sigma}\right) = P(Z \le z)$$

where Z is a *standard normal variable*, and $z = \left(\frac{x - \mu}{\sigma}\right)$ is the z-value obtained by standardising X.

The probability is obtained by using the table in Appendix-*E* with $z = (x - \mu)/\sigma$.

Basic Properties of the Standard Normal Curve

1. The total area under the standard normal curve is 1.
2. The standard normal curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.
3. The standard normal curve is symmetric about 0. That is, the part of the curve to the left of the vertical line at the centre in Fig. 3.7(*b*) is the mirror image of the curve to the right of it.
4. Almost all the area under the standard normal curve lies between –3 and 3.

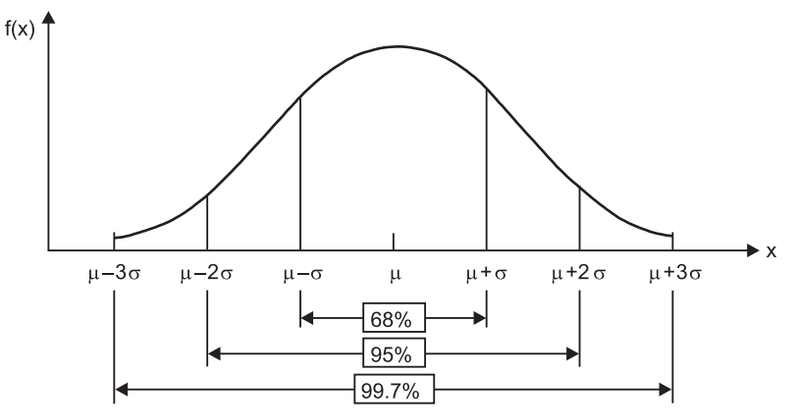


Fig. 3.7(b): Probabilities associated with a normal distribution

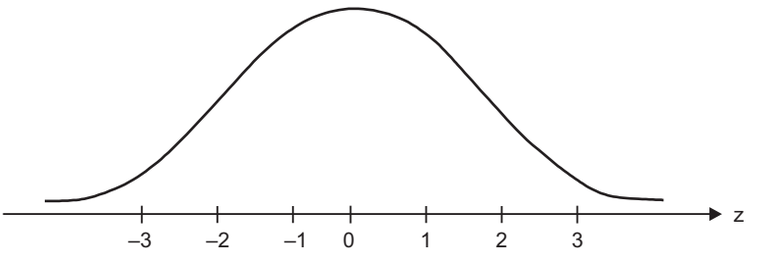


Fig. 3.7(c): Standard normal distribution

A normal random variable with $\mu = 0$ and $\sigma^2 = 1$ is called a *standard normal variable* and is denoted as Z. The cumulative distribution function of a standard normal random variable is denoted as $\phi(z) = P(Z \leq z)$.

The tables in Appendix-*E* provides cumulative probabilities for a standard normal random variable.

The procedure in solving problems involving calculating the cumulative normal probabilities is simple. For example, suppose that $X \sim N(50, 4)$, and we wish to find the probability that x is less than or equal to 54; that is $P(x \leq 54) = F(54)$. Since that standard normal random variable is

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

we can *standardize* the point of interest $x = 104$ to obtain

$$z = \frac{x - \mu}{\sigma} = \frac{54 - 50}{2} = 2$$

Now the probability that the *standard* normal random variable z is less than or equal to 2 is equal to the probability that the *original* normal random variable x is less than or equal to 54. Expressed mathematically,

$$F(x) = \phi\left(\frac{x - \mu}{\sigma}\right) = \phi(z)$$

$$F(54) = \phi(2)$$

The table in Appendix-*E* contains cumulative standard normal probabilities for various values of z. From this table, we can read

$$\phi(2) = 0.9772$$

Note that in the relationship $z = (x - \mu)/\sigma$, the variable z measures the departure of x from the mean μ in standard deviation (σ) units. In our example $F(54) = f(2)$ indicates that 54 is *two* standard deviations ($\sigma = 2$) above the mean. In general, $x = \mu + \sigma z$. In solving problems, we sometimes need to use the symmetry property of φ in addition to the tables.

In order to find the percentages of all possible observations of a normally distributed variable that lie within any specified range, we express the range in terms of z-scores and then determine the corresponding area under the standard normal curve. The stepwise procedure to determine a percentage or probability for a normally distributed variable is presented below:

1. Sketch the normal curve associated with the variable as shown in Fig. 3.7(*d*).
2. Shade the region of interest and mark its delimiting x-value(s).

3. Compute the z-scores for the delimiting x-value(s) found in step 2.
4. Use the table in Appendix-E to find the area under the standard normal curve delimited by the z-score(s) found in step 3.

The probability distributions presented in this chapter are summarised in Table 3.1.

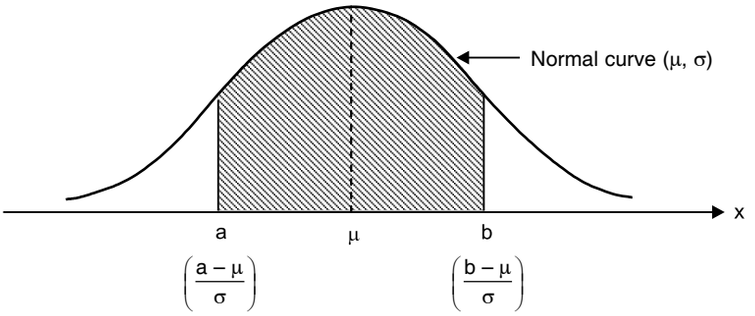


Fig. 3.7(d)

Table 3.1: Summary of probability distributions

Name	Probability Density Function	Mean	Variance
Discrete			
Hypergeometric	$\dfrac{\binom{K}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$ $x = \max(0, n-N+K), 1, \ldots \min(K, n), K \le N, n \le N$	np, where $p = \dfrac{K}{N}$	$np(1-p)\left(\dfrac{N-m}{N-1}\right)$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n, 0 \le p \le 1$	np	$np(1-p)$
Poisson	$\dfrac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 2, \ldots, \lambda > 0$	λ	λ
Continuous			
Normal	$\dfrac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ $-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$	μ	σ^2

Example E3.48

(a) Determine the area under the standard normal curve that lies to the left of 1.24, as shown in Fig. E3.48(*a*).

(b) Determine the area under the standard normal curve that lies to the right of 0.77, as shown in Fig. E3.48(*b*).

(c) Determine the area under the standard normal curve that lies between –0.69 and 1.83, as shown in Fig. E3.48(*c*).

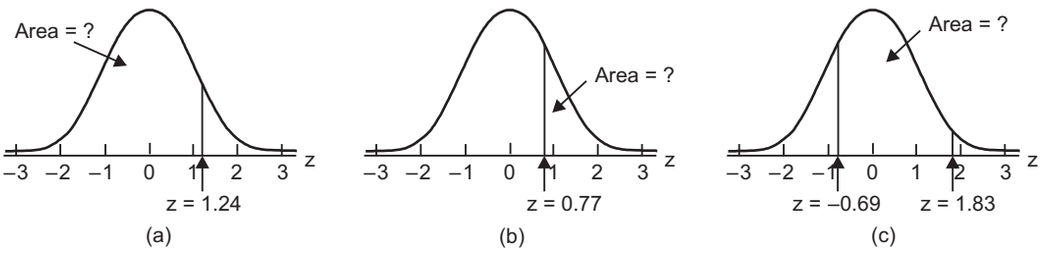


Fig. E3.48

SOLUTION:

(a) Refer to the table in Appendix-*E*. First, we go down the left hand column, labeled *z*, to "1.2". Then, going across that row to the column labeled "0.04", we reach 0.892512. This number is the area under the standard normal curve that lies to the left of 1.24, as shown in Fig. E3.48(*d*).

(b) Because the total area under the standard normal curve is 1, the area to the right of 0.77 equals 1 minus the area of the left of 0.77. We find this by first giving down the *z*-column to "0.7". Then, going across that row to the column, labeled "0.07", we reach 0.779350, which is the area under the standard normal curve that lies to the left of 0.77. Thus, the area under the standard normal curve that lies to the right of 0.77 is 1 – 0.779350 = 0.220650 as shown in Fig. E3.48(*e*).

(c) The area under the standard normal curve that lies between –0.69 and 1.83 equals the area to the left of 1.83 minus the area to the left of –0.69. The table in Appendix-*E* shows that these latter two areas are 0.966375 and 0.245097, respectively. Hence the area we seek is 0.966375 – 0.245097 = 0.721278, as shown in Fig. E3.48(*f*).

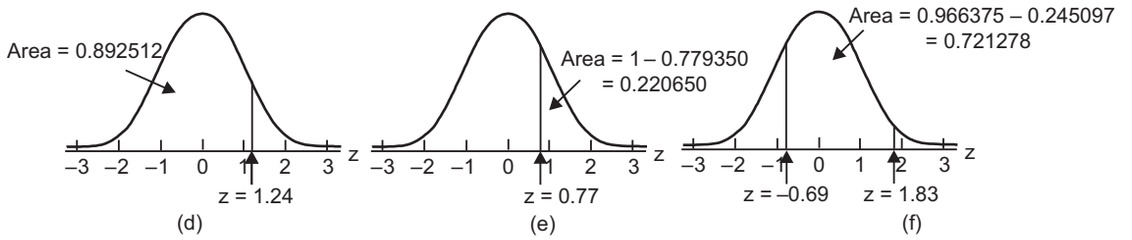


Fig. E3.48

Example E3.49

Determine the *z*-score having an area of 0.06 to its left under the standard normal curve, as shown in Fig. E3.49.

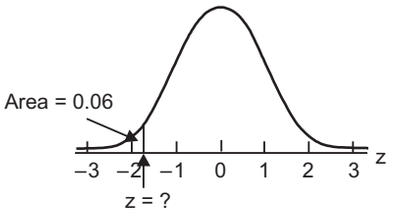


Fig. E3.49

SOLUTION:

See Fig. E3.48(*a*).

Search the body of the table in Appendix-*E* for the area 0.06. There is no such area, so use the area closest to 0.06, which is 0.060571. The *z*-score corresponding to that area is –1.55. Hence, the *z*-score having area 0.06 to its left under the standard normal curve is roughly –1.55, as shown in Fig. E3.49(*a*).

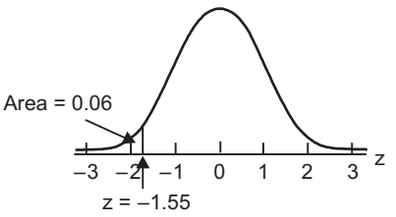


Fig. E3.49(a)

Example E3.50

Use the table in Appendix-*E* to find

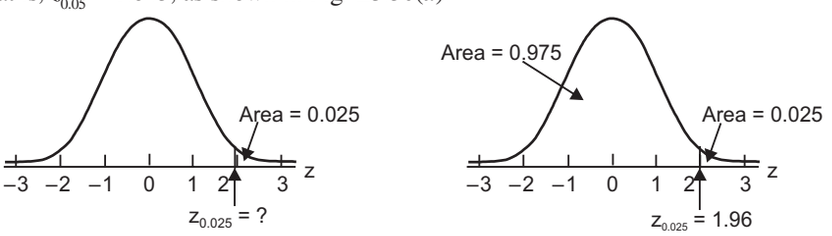
(a) $z_{0.025}$

(b) $z_{0.05}$

SOLUTION:

(a) $z_{0.025}$ is the *z*-score that has an area of 0.025 to its right under the standard normal curve, as shown in Fig. E3.50(*a*). Because the area to its right is 0.025, the area to its left is 1 – 0.025 = 0.975, as shown in Fig. E3.50(*b*). Table in Appendix-*E* contains an entry for the area 0.975002, its corresponding *z*-score is 1.96. Thus, $z_{0.025} = 1.96$, as shown in Fig. E3.50(*b*).

(b) $z_{0.05}$ is the *z*-score that has an area of 0.05 to its right under the standard normal curve, as shown in Fig. E3.50(*c*). Because the area to its right is 0.05, the area to its left is 1 – 0.05 = 0.95, as shown in Fig. E3.50(*d*). Table in Appendix-*E* does not contain any entry for the area 0.95 and has two area entries equally closest to 0.95 – namely, 0.949497 and 0.950529. The *z*-scores corresponding to those two areas are 1.64 and 1.65 respectively. So our approximation of $z_{0.05}$ is the mean of 1.64 and 1.65, that is, $z_{0.05} = 1.645$, as shown in Fig. E3.50(*d*).



(a, c) (b, d)

Fig. E3.50

Example E3.51

Find the two z-scores that divide the area under the standard normal curve into a middle 0.95 area and two outside 0.025 areas, as shown in Fig. E3.51(*a*).

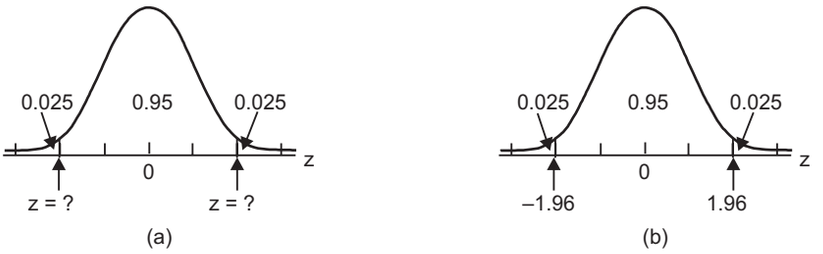


Fig. E3.51

SOLUTION:

The area of the shaded region on the right in Fig. E3.51(*a*) is 0.025. The corresponding z-score, $z_{0.025}$, is 1.96. Because the standard normal curve is symmetric about 0, the z-score on the left is –1.96. Therefore the two required z-scores are ±1.96, as shown in Fig. E3.51(*b*).

Example E3.52

In a long distance run of 10 km in a city the times of finishers are normally distributed with mean 71 minutes and standard deviation 9 minutes.

(a) determine the percentage of finishers with times between 60 and 80 minutes

(b) find the percentage of finishers with times less than 85 minutes

(c) obtain and interpret the 40th percentile for the finishing times

(d) find and interpret the 8th decile for the finishing times.

SOLUTION:

(a) For finishers with times of 60 and 80 minutes, the z-values are

$$z = \frac{x-\mu}{\sigma} = \frac{60-71}{9} = -1.22 \text{ and } z = \frac{x-\mu}{\sigma} = \frac{80-71}{9} = 1.00$$

The area to the left of $z = -1.225$ is 0.111233 and the area to the left of $z = 1.00$ is 0.841345. Hence, the area between $z = -1.22$ and $z = 1.00$ is 0.841345 – 0.111233 = 0.730112. Thus, the percentage of finishers with times between 60 minutes and 80 minutes in that city 10 km run is 73.01%.

(b) For finishing time of 85 minutes, the z-value is

$$z = \frac{x-\mu}{\sigma} = \frac{85-71}{9} = 1.56$$

The area of the left of $z = 1.56$ is 0.9406. Thus, the percentage of finishers with times less than 85 minutes is 94.06%.

(c) Using the table in Appendix-*E*, we find that an area of 0.40 lies to the left of $z = -0.25$. We convert this z-value to an x-value using $x = \mu + z\sigma$. Hence 40% of the finishers had times less than 40th percentile, $x = 71 + (-0.25)(9) = 68.75$ minutes.

(d) The 8th decile is the same as the 80th percentile. Using the table in Appendix-*E*, we find that an area of 0.80 lies to the left of $z = 0.84$. We convert the z-value to an x-value using $x = \mu + z\sigma$. Thus, 80% of the finishing times were less than $71 + (0.84)9 = 78.56$ minutes.

Example E3.53

If X has a normal distribution for which the mean is 1 and standard deviation is 2, find the value of each of the following:

(a) $P(X \leq 3)$

(b) $P(X > 1.5)$

(c) $P(X = 1)$

(d) $P(2 < X < 5)$

(e) $P(X \geq 0)$

(f) $P(-1 < X < 0.5)$

(g) $P(|X| \leq 2)$

(h) $P(1 \leq -2X + 3 \leq 8)$

SOLUTION:

Given $\mu = 1$ and standard deviation = 2. Refer to the table in Appendix-*E* for cumulative standard normal standard distribution.

$$z = \frac{x - \mu}{\sigma} = \frac{x - 1}{2}$$

then z has a standard normal distribution.

(a) $P(X \leq 3) = P(Z \leq 1) = \Phi(1) = 0.841345$

(b) $P(X > 1.5) = P(Z > 0.25) = 1 - \Phi(0.25) = 0.4013$

(c) $P(X = 1) = 0$, because X has a continuous distribution.

(d) $P(2 < X < 5) = P(0.5 < Z < 2) = \Phi(2) - \Phi(0.5) = 0.2858$

(e) $P(X \geq 0) = P(Z \geq -0.5) = P(Z \leq 0.5) = \Phi(0.5) = 0.6915$

(f) $P(-1 < X < 0.5) = P(-1 < Z < -0.25) = P(0.25 < Z < 1) = \Phi(1) - \Phi(0.25) = 0.2426$

(g) $P(|X| \leq 2) = P(-2 \leq X \leq 2) = P(-1.5 \leq Z \leq 0.5) = P(Z \leq 0.5) - P(Z \leq -1.5) = P(Z \leq 0.5) - P(Z \leq 1.5)$
$= \Phi(0.5) - [1 - \Phi(1.5)] = 0.6247$

(h) $P(1 \leq -2X + 3 \leq 8) = P(-2 \leq -2X \leq 5) = P(-2.5 \leq X \leq 1) = P(-1.75 \leq Z \leq 0) = P(0 \leq Z \leq 1.75)$
$= \Phi(1.75) - \Phi(0) = 0.4599$

Example E3.54

Assume X is normally distributed with a mean of 10 and a standard deviation of 2. Determine the following:

(a) $P(X < 13)$

(b) $P(X > 9)$

(c) $P(6 < X < 14)$

(d) $P(2 < X < 4)$

(e) $P(-2 < X < 8)$

SOLUTION:

Refer to the table in Appendix-*E*.

(a) $P(X < 13) = P\left(Z < \dfrac{13-10}{2}\right) = P(Z < 1.5) = 0.93319$

(b) $P(X > 9) = 1 - P(X < 9) = 1 - P(Z < -0.5) = 0.69146$

(c) $P(6 < X < 14) = P\left(\dfrac{6-10}{2} < Z < \dfrac{14-10}{2}\right) = P(-2 < Z < 2) = [P(Z < 2) - P(Z < -2)] = 0.9545$

(d) $P(2 < X < 4) = P\left[\dfrac{2-10}{2} < Z < \dfrac{4-10}{2}\right] = P(-4 < Z < -3) = P(Z < -3) - P(Z < -4) = 0.00132$

(e) $P(-2 < X < 8) = P(X < 8) - P(X < -2) = P\left(Z < \dfrac{8-10}{2}\right) - P\left(Z < \dfrac{-2-10}{2}\right)$

$= P(Z < -1) - P(Z < -6) = 0.15866$

Example E3.55

The compressive strength of samples of cement can be modeled by a normal distribution with a mean of 7000 kg/cm^2 and a standard deviation of 100 kg/cm^2.

(a) find the probability of a sample's strength that is less than 7250 kg/cm^2

(b) find the probability that a sample's strength is between 6800 and 6900 kg/cm^2

(c) what strength is exceeded by 95% of the samples?

SOLUTION:

Refer to the table in Appendix-*E*.

(a) $P(X < 7250) = P\left(Z < \dfrac{7250-7000}{100}\right) = P(Z < 2.5) = 0.99379$

(b) $P(6800 < X < 6900) = P\left[\dfrac{6800-7000}{100} < Z < \dfrac{6900-7000}{100}\right] = P(-2 < Z < -1)$

$= P(Z < -1) - P(Z < -2) = 0.13591$

(c) $P(X > x) = P\left(Z > \dfrac{x-7000}{100}\right) = 0.95$

Hence, $\dfrac{x-7000}{100} = -1.65$ and $x = 6835$.

Example E3.56

(a) A process manufactures ball bearings whose diameters are normally distributed with mean 3.505 cm and standard deviation 0.008 cm. Specifications call for the diameter to be in the interval 3.5 ± 0.01 cm. What proportion of the ball bearings will meet the specification?

(b) Suppose the process can be recalibrated so that the mean will equal to 3.5 cm, the centre of the specification interval. The standard deviation of the process remains 0.008 cm. What proportion of the diameter will meet the specifications?

SOLUTION:

Let X represents the diameter of a randomly chosen ball bearing. Then, $X \sim N(3.505, 0.008^2)$. Figure E3.56(*a*) shows the probability density function of the $N(3.505, 0.008^2)$ population. The shaded area represents $P(3.49 < X < 3.51)$, which is the proportion of ball bearings that meet the specification.

(a) We compute the z-scores of 3.49 and 3.51 as follows:

$$z = \frac{3.49 - 3.505}{0.008} = -1.88 \text{ and } z = \frac{3.51 - 3.505}{0.008} = 0.63$$

The area to the left of $z = -1.88$ is 0.0301. The area to the left of $z = 0.63$ is 0.7357. The area between $z = 0.63$ and $z = -1.88$ is 0.7357 – 0.0301 = 0.7056. Hence, approximately 70.56% of the diameters of the ball bearings will meet the specifications.

(b) The mean is 3.5000 rather than 3.505. The calculations are as follows: (see Fig. E3.56(*b*)).

$$z = \frac{3.49 - 3.50}{0.008} = -1.25 \text{ and } z = \frac{3.51 - 3.50}{0.008} = 1.25$$

The area to the left of $z = -1.25$ is 0.1056. The area of the left of $z = 1.25$ is 0.8944. The area between $z = 1.25$ and $z = -1.25$ is 0.8944 – 0.1056 = 0.7888. Hence, recalibrating will increase the proportion of diameters that meet the specification to 78.88%.

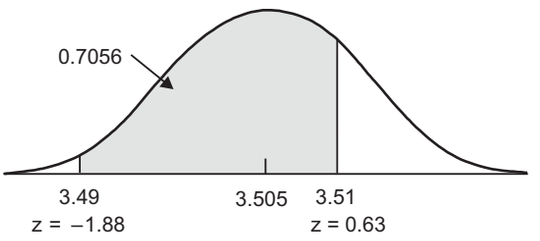


Fig. E3.56(a)

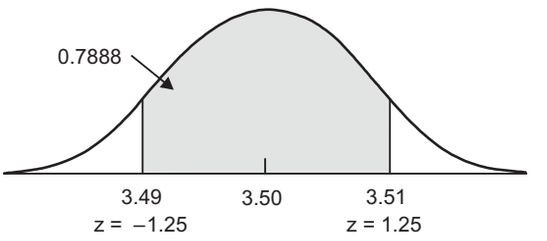


Fig. E3.56(b)

Example E3.57

Assume X is normally distributed with a mean of 5 and a standard deviation of 1. Determine the value of x that solves each of the following:

(a) $P(X > x) = 0.5$

(b) $P(X > x) = 0.95$

(c) $P(x < X < 5) = 0.2$

(d) $P[-x < (X - 5) < x] = 0.95$

(e) $P[-x < (X - 5) < x] = 0.99$

SOLUTION:

(a) $P(X < x) = P\left(Z > \dfrac{x-5}{1}\right) = 0.5$

Therefore, $\dfrac{x-5}{1} = 0$ and $x = 5$.

(b) $P(X < x) = P\left(Z > \dfrac{x-5}{1}\right) = 1 - P\left(Z < \dfrac{x-5}{1}\right) = 0.95$

Therefore, $P\left(Z < \dfrac{x-5}{1}\right) = 0.05$ and $\dfrac{x-5}{2} = -1.64$ and hence $x = 3.36$.

(c) $P(x < X < 5) = P\left(\dfrac{x-5}{1} < Z < 0\right) = 0.2$

Therefore, $P\left(Z < \dfrac{x-5}{1}\right) = 0.3$ and $\dfrac{x-5}{2} = -0.52$ and hence $x = 4.48$

(d) $P[5 - x < X < 5 + x] = P\left(\dfrac{-x}{2} < X < \dfrac{x}{2}\right) = 0.95$

Therefore, $x/2 = 0.98$

and $x = 1.96$

(e) $P[5 - x < X < 5 + x] = P[-x/2 < X < x/2] = 0.99$

Therefore, $x/2 = 1.79$

and $x = 2.58$.

Example E3.58

The diameter of a metal shaft for a precision instrument is assumed to be normally distributed with a mean of 0.5 mm and a standard deviation of 0.025 mm.

(a) what is the probability that shaft diameter is greater than 0.31 mm?

(b) what is the probability that shaft diameter is between 0.235 and 0.315 mm?

(c) the diameter of 90% of samples is below what value?

SOLUTION:

(a) $P(X > 0.31) = P\left(Z > \dfrac{0.315 - 0.25}{0.025}\right) = P(Z > 3) = 0.00135.$

(b) $P(0.235 < X < 0.315) = P(-0.6 < Z < 2.6) = P(Z < 2.6) - P(Z < -0.6) = 0.99534 - 0.27425 = 0.72109.$

(c) $P(X < x) = P\left(Z < \dfrac{x-0.25}{0.025}\right) = 0.90$

Hence, $\dfrac{x-0.25}{0.025} = 1.28$

and $x = 0.282$

Example E3.59

The length of a metal rod used in a machine system is normally distributed with a length of 45.10 mm and a standard deviation of 0.05 mm.

(a) what is the probability that a rod is longer than 45.15 mm or shorter than 49.85 mm?

(b) what should the process mean be set as to obtain the greatest number of rods between 44.85 and 45.15 mm?

(c) if the rods that are not between 44.85 and 45.15 mm are scrapped, what is the yield for the process mean that one would select in part *(b)*?

SOLUTION:

(a) $$P(45.15 < X) = P\left[\left(\frac{45.15 - 45.10}{0.05}\right) < Z\right] = P(1 < Z) = 1 - 0.084134 = 0.15866$$

$$P(X < 49.85) = P\left(Z < \frac{49.85 - 45.10}{0.05}\right) = P(Z < -5) = 0$$

Hence, the answer is 0.159.

(b) The process mean should set at the centre of the specifications, that is, at $\mu = 45.0$

(c) $$P(44.85 < X < 45.15) = P\left[\frac{44.85 - 45.0}{0.05}\right] < Z < \left[\frac{45.15 - 45.0}{0.05}\right] = P(-3 < Z < 3) = 0.9973$$

Example E3.60

Refer to Example E3.59. If the process is centred so that the mean is 45 mm and the standard deviation is 0.05 mm and that 10 rods are measured and are assumed to be independent, determine

(a) the probability that all 10 rods measured are between 45.85 and 45.15 mm

(b) the expected number of the 10 rods that are between 44.85 and 45.15 mm.

SOLUTION:

(a) $P(44.85 < X < 45.15) = 0.9973$ from Example E3.59. Therefore, by independence, the probability of 10 rods are within the given limit is $0.9973^{10} = 0.9733$.

(b) Let Y denotes the number of rods from the 10 selected that are within the given limits. Then, Y is binomial with $n = 10$ and $p = 0.9973$.

Hence, $E(Y) = 9.973$.

Example E3.61

The weight of a electronic component is normally distributed with a mean of 6 ounces and a standard deviation of 0.25 ounce.

(a) find the probability that the electronic component weighs more than 6.5 ounces.

(b) what must the standard deviation of weight be in order for the company to state that 99% of its electronic components are less than 6.5 ounces?

(c) if the standard deviation stays at 0.25 ounces, what must the mean weight be in order for the company to state that 99.9% of its electronic components are less than 6.5 ounces?

SOLUTION:

(a) $P(X > 6.5) = P\left(Z > \dfrac{6.5-6}{0.25}\right) = P(Z > 2) = 0.02275$

(b) If $P(X < 13) = 0.999$, then $P\left(Z < \dfrac{6.5-6}{\sigma}\right) = 0.999$

Hence, $\dfrac{0.5}{\sigma} = 3.09$

$\sigma = 0.1618$

(c) If $P(X < 6.5) = 0.999$, then $P\left(Z < \dfrac{6.5-\mu}{0.25}\right) = 0.999$

Therefore, $\dfrac{6.5-\mu}{0.25} = 3.09$ and $\mu = 5.7275$

Example E3.62

The weight of a mechanical component is normally distributed with a mean of 22 oz and a standard deviation of 0.5 oz.

(a) what is the probability that a component weighs more than 23 oz?

(b) what must the standard deviation of weight be in order for the company to state that 99.9% of its components are less than 23 oz?

(c) if the standard deviation remains at 0.5 oz, what must the mean weight be in order for the company to state that 99.9% of its components are less than 23 oz?

SOLUTION:

Refer to the table in Appendix-*E*.

(a) $P(X > 23) = P\left(Z > \dfrac{23-22}{0.5}\right) = P(Z > 2) = 1 - P(Z \le 2) = 1 - 0.977250 = 0.022750.$

(b) If $P(X < 23) = 0.999$, then $P\left[Z < \dfrac{23-22}{\sigma}\right] = 0.999$

Hence, $\dfrac{1}{\sigma} = 3.09$ and $\sigma = \dfrac{1}{3.09} = 0.324$

(c) If $P(X < 23) = 0.999$, then $P\left[Z < \dfrac{23-\mu}{0.5}\right] = 0.999$

Hence, $\dfrac{23-\mu}{0.5} = 3.09$

or $\mu = 21.455.$

Example E3.63

Refer to the Example E3.56. Assume that the process has been recalibrated so that mean diameter is now 3.5 cm. To what value must the standard deviation be lowered so that 95% of the diameters will meet the specifications?

SOLUTION:

The specification interval is 3.49 — 3.51 cm. We must find a value for σ so that this interval spans the middle 95% of the population of ball bearing diameters, as shown in Fig. E3.63. The z-score that has 2.5% of the area to its left is $z = -1.96$. The z-score that has 2.5% of the area to its right is $z = 1.96$ (from the symmetry of the curve). It follows that the lower specification limit, 3.49, has a z-score of –1.96, while the upper limit of 3.51 has a score of 1.96. Either of these facts may be used to find σ.

Now $$z = \frac{x - \mu}{\sigma}$$

or $$1.96 = \frac{3.51 - 3.50}{\sigma}$$

or $$\sigma = 0.005102 \text{ cm}$$

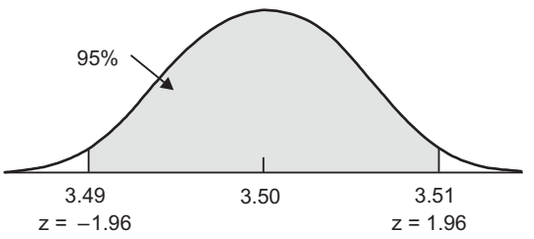


Fig. E3.63

Example E3.64

An experiment needs a 2.41 cm diameter steel rod. Suppose that the diameter of a steel rod has a normal distribution with a mean of 2.41 cm and a standard deviation of 0.01 cm

(a) determine the probability that a diameter is greater than 2.42 cm

(b) what diameter is exceeded by 95% of the samples?

(c) if the specifications require that the diameter is between 2.39 cm and 2.43 cm, what proportion of the samples meet specifications?

SOLUTION:

Refer to the table in Appendix-*E*.

(a) Let X denotes the diameter of the steel rod.

$X \sim N(2.41, 0.01^2)$

$$P(X > 2.42) = 1 - P(X \le 2.42) = 1 - \Phi\left[\frac{2.42 - 2.41}{0.01}\right] = 1 - \Phi(1) = 0.1587$$

$$(b)\quad P(X > x) = P\left[Z > \frac{x - 2.41}{0.01}\right] = 0.95$$

Hence, $$\frac{x - 2.41}{0.01} = 1.645 \text{ and } x = 2.3936 \text{ cm}$$

$$(c)\quad P(2.39 < X < 2.43) = \Phi\left[\frac{2.43 - 2.41}{0.01}\right] - \Phi\left[\frac{2.39 - 2.41}{0.01}\right] = \Phi(2) - \Phi(-2) = 0.9545$$

3.5 APPROXIMATING PROBABILITY DISTRIBUTIONS

In some engineering quality control applications, it is quite useful to approximate one probability distribution by another. Figure 3.8 shows the approximation guide, that is, the conditions under which one distribution may be approximated by another.

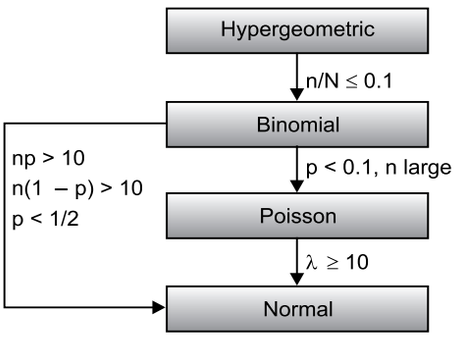


Fig. 3.8: Approximation guide

3.5.1 Binomial Approximation to the Hypergeometric

When the ratio n/N is small, that is, say $n/N \leq 0.1$, then the binomial distribution parameter $p = D/N$ can be used as a good approximation to the hypergeometric distribution. The smaller the function n/N, the better the approximation.

3.5.2 Poisson Approximation to the Binomial

The Poisson distribution can be used as an approximation to the binomial distribution when the sample size n is large and the probability of success p is small (the same applies when q is small, p and q being of course interchangeable), i.e., when the binomial distribution is highly skewed. As a guide, we can say that a good approximation is obtained when $n \geq 20$ and $p \leq 0.05$, and the approximation improves with a decrease in p.

The advantage of using the Poisson distribution to approximate the binomial distribution is that the Poisson distribution, having only one parameter, is well tabulated.

When the parameter is small, say, $p < 0.1$, and n is large, with $\lambda = np$ constant, the Poisson distribution is used as an approximation to the binomial distribution. The larger the value of n and the smaller the value of p, the better the approximation.

3.5.3 Normal Approximation to the Binomial

Recalling that if $X \sim \text{Bin}(n, p)$, then $X = Y_1 + Y_2 + \cdots + Y_n$, where $Y_1 + Y_2 + \cdots + Y_n$ is a sample from a Bernoulli (p) population. Hence, X is the sum of the sample observations.

The sample proportion is

$$\hat{p} = \frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

which is also the sample mean $\bar{Y}$. The Bernoulli population has mean $\mu = p$ and variance $\sigma^2 = p(1-p)$. It follows from the central limit theorem (see section 3.8) that if the number of trials n is large, then $X \sim N(np, np(1-p))$, and $\hat{p} \sim N(p, p(1-p/n))$.

In the binomial case, the accuracy of the normal approximation depends on the mean number of successes np and the mean number of failures $n(1-p)$. The larger the value of np and $n(1-p)$, the better the approximation. Usually, the normal distribution is used as an approximation to the binomial distribution when np and $n(1-p)$ are both greater than 5. A better and more conservative rule is to use the normal approximation whenever $np > 10$ and $n(1-p) > 10$.

Hence if $X \sim \text{Bin}(n, p)$ and if $np > 10$ and $n(1-p) > 10$, then

$X \sim N(np, np(1-p))$ approximately

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \quad \text{approximately}$$

Hence, if X is a binomial random variable with parameter n and p

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable.

If n is large, then one can justify the approximation of the binomial distribution by the normal distribution with mean np and variance $np(1-p)$. Noting that the binomial distribution is discrete and the normal distribution is continuous, we can make a continuity correction such that

$$p(X = a) = \Phi\left[\frac{(a + 1/2) - np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{(a - 1/2) - np}{\sqrt{np(1-p)}}\right] \tag{3.82}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. We also write probability statements as

$$P(a \le X \le b) = \Phi\left[\frac{(b + 1/2) - np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{(a - 1/2) - np}{\sqrt{np(1-p)}}\right] \tag{3.83}$$

Although p should be about 0.5, the approximation can be used without excessive loss of accuracy for $0.1 \le p \le 0.9$. If p is close to 0.5 and $n > 10$, the approximation is fairly good. For other values of p, the value of n should be larger.

In general, the approximation is good as long as $np > 5$ for $p \le 0.5$ or when $nq > 5$ when $p > 0.5$.

The normal distribution is also used to approximate the sample fraction nonconforming, $\hat{p}$. The random variable $\hat{p}$ is normally distributed with mean p and variance $p(1-p)/n$, so that we have

$$p\left[\frac{a}{n} \le \frac{X}{n} \le \frac{b}{n}\right] = P\left[\frac{a}{n} \le \hat{p} \le \frac{b}{n}\right]$$

which indicates that we divide all terms in Eq.(3.83) by n to obtain

$$P\left(\frac{a}{n} \le \hat{p} \le \frac{b}{n}\right) = \Phi\left[\frac{\frac{b+1/2}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right] - \Phi\left[\frac{\frac{a-1/2}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right] \tag{3.84}$$

Continuity Correction: The binomial distribution is discrete, while the normal distribution is continuous. The *continuous correction* is an adjustment, made when approximating a discrete distribution with a continuous one that can improve the accuracy of the approximation.

3.5.4 Normal Approximation to the Poisson

If X is a Poisson random variable with $E(X) = \lambda$ and $V(X) = \lambda$

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

is approximately a standard normal random variable. The approximation is good for $\lambda > 10$.

The normal distribution with mean $\mu = \sigma^2 = \lambda$ can be used as an approximation to the Poisson distribution if the mean λ of the distribution is large, say 10 or more.

The continuity correction used is given by

$$P(a \le X \le b) = \Phi\left[\frac{(b+1/2) - \lambda}{\sqrt{\lambda}}\right] - \Phi\left[\frac{(a-1/2) - \lambda}{\sqrt{\lambda}}\right] \tag{3.85}$$

Example E3.65

The number of widgets made per shift is equal to 1000. It is known that 8% of them are nonconforming. A sample of 20 widgets is taken. What is the probability that 2 or fewer widgets in the sample are nonconforming?

(a) set up the equation using the hypergoemetric distribution

(b) solve it using the binomial approximation

(c) solve it using the Poisson approximation.

SOLUTION:

$D = 80, N = 1000, n = 20, p = 0.08$

(a) $$P(X \le 2) = \sum_{x=0}^{2} \frac{\binom{80}{x}\binom{920}{20-x}}{\binom{1000}{20}}$$

$$(b)\quad P(X \le 2) = \sum_{x=0}^{2} \binom{20}{x} (0.08)^x (0.92)^{20-x}$$

$$= \binom{20}{0}(0.08)^0(0.92)^{20} + \binom{20}{1}(0.08)^1(0.92)^{19} + \binom{20}{2}(0.08)^2(0.92)^{18}$$

$$= 0.189 + 0.328 + 0.271 = 0.788$$

(c) $\lambda = np = 20(0.08) = 1.6$

$P(x \le 2) = 0.783$

Example E3.66

If 20% of the electronic components made in a company are nonconforming, find the probability that in a random sample of 100 such electronic components selected and using normal approximation.

(a) at most 16 will be nonconforming

(b) exactly 16 will be nonconforming

(c) between 16 and 22 will be nonconforming.

SOLUTION:

Normal approximation

$$np = 100(0.20) = 20 \cdot \sqrt{np(1-p)} = 4$$

$$(a)\quad P(x \le 16) = \Phi\left[\frac{16.5 - 20}{4}\right] = \Phi(-0.809) = 1 - 0.809 = 0.191$$

$$(b)\quad P(x = 16) = \Phi\left[\frac{16.5 - 20}{4}\right] - \Phi\left[\frac{15.5 - 20}{4}\right] = 0.191 - \phi(-1.125) = 0.191 - (1 - 0.855) = 0.046$$

$$(c)\quad P(16 \le x \le 22) = \Phi\left[\frac{22.5 - 20}{4}\right] - \Phi\left[\frac{15.5 - 20}{4}\right] = \phi(0.65) - 0.14 = 0.734 - 0.145 = 0.589$$

Example E3.67

Use the normal approximation to the binomial distribution to determine the probability of getting 7 heads and 7 tails in the 14 tosses of a balanced coin.

SOLUTION:

We must thus determine the area under the curve between 6.5 and 7.5. $n = 14$ and $x = 7$, $\mu = 14\left(\frac{1}{2}\right) = 7$ and

$$\sigma = \sqrt{14\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = \sqrt{3.5} = 1.871$$

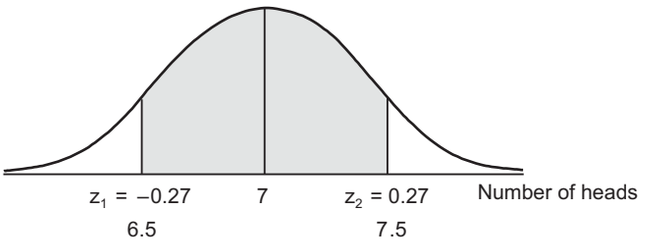


Fig. E3.67

$$z_1 = \frac{6.5-7}{1.871} = -0.27 \text{ and } z_2 = \frac{7.5-7}{1.871} = 0.27$$

The table for normal distribution in Appendix-*E* corresponding to $z = -0.27$ and 0.27 are 0.393580 and 0.60642 and we find that the normal approximation to the probability of "7 heads and 7 tails" is 0.60642 – 0.393580 = 0.2128.

The table for binomial approximation in Appendix-*C* gives 0.2095. Hence, the error of the approximation is

0.2095 – 0.2128 = –0.0033 and the percentage error is $\frac{0.0033}{0.2095}$ (100) = 1.5752% in the absolute value.

Example E3.68

Fifteen per cent of the parts produced in a manufacturing process fail a standardized quality test.

(a) using the Poisson formula, find the probability that in a random sample of 100 parts which went through this test exactly 20 will fail.

(b) using the Poisson probabilities table, find the probability that the number of parts which fail this test in a randomly selected 100 parts is

(i) at most 9

(ii) 10 to 16

(iii) at least 20.

SOLUTION:

Let x denotes the number of parts in a random sample of 100 which fail the test. Since, on average, 15% of the parts fail the test, $\lambda = 0.15 \times 100 = 15$.

(a) $P(\text{exactly 20 fail}) = P(x = 20) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(15)^{20} e^{-15}}{20!} = 0.0418$

(b) (i) $P(\text{at most 9 fail}) = P(x \le 9) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7) + P(8) + P(9)$

$= 0.0 + 0.0 + 0.0 + 0.0002 + 0.0006 + 0.0019 + 0.0048 + 0.0104 = 0.0194 + 0.0324 = 0.0697$

(ii) $P(10 \text{ to } 16 \text{ fail}) = P(10 \le x \le 16) = P(10) + P(11) + P(12) + P(13) + P(14) + P(15) + P(16)$

$= 0.0486 + 0.0663 + 0.0829 + 0.0956 + 0.1024 + 0.01024 + 0.0960 = 0.5942$

(iii) $P(\text{at least 20}) = P(x \ge 20) = P(20) + P(21) + P(22) + \cdots$

$= 0.0418 + 0.299 + 0.0204 + 0.0133 + 0.0083 + 0.0050 + 0.0029 + 0.0016 + 0.0009$

$+ 0.0004 + 0.0002 + 0.0001 + 0.0001 = 0.1249$

Example E3.69

If the probability is 0.10 that a certain mortgage company will refuse a loan application, use the normal approximation to determine the probability that the mortgage company will refuse at most 40 of 450 mortgage loan applications.

SOLUTION:

$$n = 450,\ p = 0.10,\ \mu = np = 450(0.1) = 45,\ \sigma = \sqrt{np(1-p)} = \sqrt{450(0.1)(0.9)} = \sqrt{40.5} = 6.3640$$

$$z = \frac{40.5 - 45}{6.3640} = -0.7071$$

Referring to the table for normal distribution in Appendix-*E* corresponding to $z = -0.7071$, we find that the normal approximation to the probability is 0.22065.

Hence, the probability that the mortgage company will refuse the load application is 0.22065.

Example E3.70

The number of nonconforming manufactured parts per shift is Poisson distributed, with a mean of 16. Find the probability that there are between 14 and 20 nonconforming parts on a shift that is 14, 15, 16, 17, 18, 19 or 20.

(a) use the Poisson distribution

(b) use the normal approximation.

SOLUTION:

(a) $\lambda = 15$

$P[14 \le x \le 20] = F(20) - F(13) = 0.917 - 0.363 = 0.554$

(b) $$P[14 \le x \le 20] = \phi\left[\frac{20.5 - 15}{4}\right] - \phi\left[\frac{13.5 - 15}{4}\right] = \phi(1.375) - \phi(-0.375) = 0.9154 - 0.3538 = 0.5616$$

Example E3.71

Suppose we want to use the normal approximation to the binomial distribution to determine $B(1; 150, 0.05)$

(a) is it justified in using the approximation?

(b) make the approximation.

SOLUTION:

(a) Here $n = 100$, $p = 0.05$ and mean $= 150(0.05) = 75$ and $n(1-p) = 150 - 7.5 = 142.5$

Yes. It is justified in using the approximation since np is > 5 and $n(1-p)$ is > 100.

(b) $\mu = 7.5$, $\sigma^2 = np(1-p) = 150(0.05)(0.95) = 7.125$ or $\sigma = 2.67$.

$$z_1 = \frac{0.5 - 7.5}{2.67} = -2.62 \text{ and } z_2 = \frac{1.5 - 7.5}{2.67} = -2.25$$

The entries in table for normal distribution in Appendix-*E* corresponding to $z_1 = -2.62$ and $z_2 = -2.25$, we find that the normal approximation to the probability is

$0.012224 - 0.004396 = 0.007828$

Example E3.72

Refer to Example E3.71. Make the Poisson approximation.

SOLUTION:

Here $\lambda = 7.5$ and $P(7.5) = \dfrac{e^{-\lambda}\lambda^{x}}{x!} = \dfrac{e^{-7.5}(7.5)^{1}}{1!} = (0.00055)7.5 = 0.0041$

Example E3.73

In a survey it was found that 1.5% of the adult population in a large city suffer from drug addition. Of 100 randomly selected adult population, let X denotes the number who have drug addiction problem.

(a) what are the parameters for the appropriate normal distribution?

(b) what are the parameters for approximating Poisson distribution?

(c) compute the individual probabilities for the binomial distribution in part *(a)*

(d) compute the individual probabilities for the Poisson distribution in part *(b)*. Find the probabilities until they are zero to 4 decimal places.

(e) compare the probabilities found in parts *(c)* and *(d)*.

(f) apply both the binomial probabilities and Poisson probabilities found in parts *(c)* and *(d)* to determine the probability that the number who suffer from drug addiction is exactly 3; between 2 and 5 (inclusive); less than 4% of those surveyed; more than 2. Compare the two results in each case.

SOLUTION:

(a) Here $n = 100$ and $p = 1.5/100 = 0.015$

(b) $\lambda = np = 100(0.015) = 1.5$

(c) The binomial probability function is

$$P(X = x) = \binom{100}{x} (0.015)^{x}(0.985)100^{-x} \qquad \text{for } x = 0, 1, 2, \ldots, 100$$

Starting at $x = 0$ and evaluating the function until it is zero to four decimal places, we obtain the following Table E3.73(a).

Table E3.73(a)

x	P(X = x)
0	0.2206
1	0.3360
2	0.2532
3	0.1260
4	0.0465
5	0.0136
6	0.0033
7	0.0007
8	0.0001
9	0.0000

(d) The Poisson probability function is

$$P(X = x) = e^{1.5}\frac{1.5^x}{x!} \qquad \text{for } x = 0, 1, 2, \ldots$$

Starting at $x = 0$ and evaluating this function until it is zero to four decimal places, we obtain the following Table E3.73(b).

Table E3.73(b)

x	P(X = x)
0	0.2231
1	0.3347
2	0.2510
3	0.1255
4	0.0471
5	0.0141
6	0.0035
7	0.0008
8	0.0001
9	0.0000

(e) All the probabilities for $x = 0$ to 9 agree to 2 decimal places and are usually at most a couple of thousandths apart. The agreement is really good.

(f) See Table E3.73(c).

Table E3.73(c)

Event	Binomial probability	Poisson probability
X = 3	0.1260	0.1255
2 ≤ X ≤ 5	0.4393	0.4377
X < 4	0.9358	0.9343
X > 2	0.1902	0.1912

Each pair of probabilities in the above Table E3.73(c) agree to within 0.0016 or less.

3.6 CHEBYSHEV'S THEOREM

The standard deviation σ of a random variable X measures the weighted spread of the values of X about the mean μ. For smaller σ, we would expect that X will be closer to μ. A more precise statement of this expectation is given by Chebyshev (1821–1894) which is stated as follows:

Let X be a random variable with mean μ and standard deviation σ. Then, for any integer k, the probability that a value of X lies in the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least $1 - \frac{1}{k^2}$.

Hence, $$P[\mu - k\sigma \le X \le \mu + k\sigma] \ge 1 - \frac{1}{k^2} \tag{3.86}$$

Proof: By definition

$$\sigma^2 = \text{Var}(X) = \Sigma(x_i - \mu)^2 p_i \tag{3.87}$$

Now, deleting all terms for which $|x_i - \mu| \le k\sigma$ and denoting the summation of the remaining terms by $\Sigma^*(x_i - \mu)^2 p_i$. then, we have

$$\sigma^2 \ge \Sigma^*(x_i - \mu)^2 p_i \ge \Sigma^* k^2\sigma^2 p_i = k^2\sigma^2\Sigma^* p_i = k^2\sigma^2 p(|X - \mu| > k\sigma) \tag{3.88}$$

or $$= k^2\sigma^2[1 - p(|X - \mu| \le k\sigma)] = k^2\sigma^2[1 - P(\mu - k\sigma \le X \le \mu + k\sigma)] \tag{3.89}$$

when $\sigma > 0$, and dividing Eq. (3.89) by $k^2\sigma^2$, we get

$$\frac{1}{k^2} \ge 1 - P(\mu - k\sigma \le X \le \mu + k\sigma) \tag{3.90}$$

or $$P(\mu - k\sigma \le X \le \mu + k\sigma) \le 1 - \frac{1}{k^2} \tag{3.91}$$

Equation (3.91) is the Chebyshev's inequality for $\sigma > 0$.

If $\sigma = 0$, then $x_i = \mu$ for all $p_i > 0$, and we have

$$P[\mu - k(0) \le X \le \mu + k(0)] = P(X = \mu) = 1 > 1 - \frac{1}{k^2} \tag{3.91a}$$

Chebyshev's theorem gives a lower bound for the area under a curve between two points that are an opposite sides of the mean and at the same distance from the mean. Chebyshev's theorem is stated as follows: For any number k greater than 1, at least $\left(1 - \frac{1}{k^2}\right)$ of the data values lie within k standard deviations of the mean. Chebyshev's theorem is illustrated in Fig. 3.9.

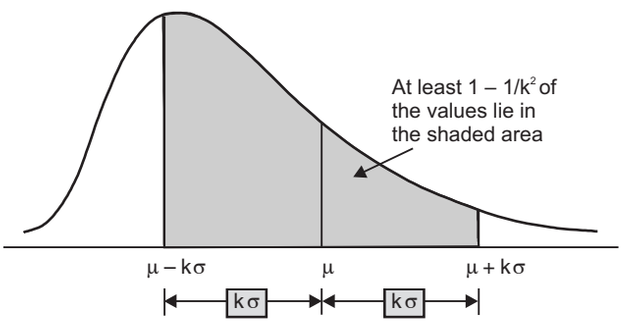


Fig. 3.9: Chebyshev's theorem

For $k = 2$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75 \text{ or } 75\%$$

That is, at least 75% of the values of a data set lie within two standard deviations of the mean as shown in Fig. 3.10.

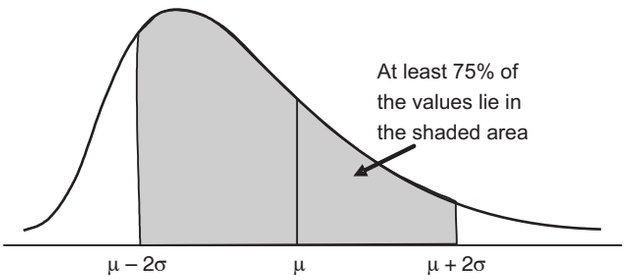


Fig. 3.10: Percentage of values within two standard deviations of the mean

For $k = 3$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 0.89 \text{ or } 89\%$$

When $k = 3$, this inequality shows that regardless of the assumed distribution of X, the probability is no more than $1/3^2 = 1/9$ that the random variable takes on a value more than three standard deviations away from its mean (see Fig. 3.11).

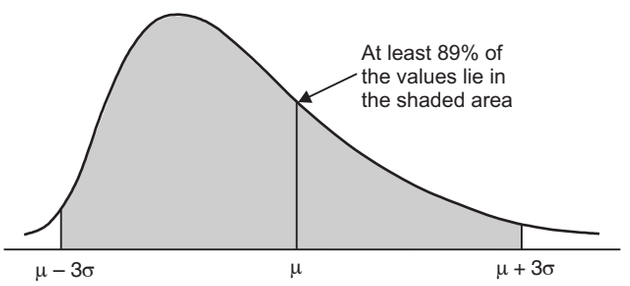


Fig. 3.11: Percentage of values within three standard deviations of the mean

Chebyshev's inequality is applicable to both sample and population data. Chebyshev's inequality is also applicable to distribution of any shape. It should be noted here that Chebyshev's inequality can be used only for $k > 1$ (since when $k = 1$, $1 - \frac{1}{k^2} = 0$ and when $k < 1$, $1 - \frac{1}{k^2}$ is negative).

Example E3.74

The mean and standard deviation for the final examination scores in *Economics* course are 80 and 7.5 respectively. Determine the percentage of students who scored between 65 and 95, using Chebyshev's theorem.

SOLUTION:

From the given data

Mean $\mu = 80$ and standard deviation $\sigma = 10$.

Each of the two points, 60 and 95, is 15 units away from the mean.

Hence, $$k = \frac{15}{\sigma} = \frac{15}{7.5} = 2$$

$$1-\frac{1}{k^2}=1-\frac{1}{2^2}=0.75 \text{ or } 75\%$$

That is, at least 75% of the students scored between 60 and 95 as shown in Fig. E3.74.

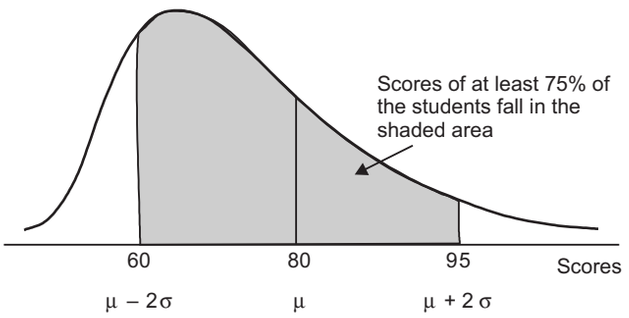


Fig. E3.74: Percentage of students who scores between 60 and 95

Example E3.75

A random variable X has mean $\mu = 35$ and standard deviation $\sigma = 2$. Apply Chebyshev's inequality to estimate

(a) $P(X \le 45)$

(b) $P(X \ge 30)$

SOLUTION:

Chebyshev's inequality states:

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \ge 1-\frac{1}{k^2}$$

(a) Substituting $\mu = 35$ and $\sigma = 2$ in $\mu + k\sigma$, we have

$$35 + 2k = 45 \text{ or } k = 5$$

Hence, $$1-\frac{1}{k^2}=1-\frac{1}{5^2}=0.96$$

Similarly, $\mu - k\sigma = 35 - 10 = 25$

Chebyshev's inequality gives

$P(25 \le X \le 45) \ge 0.96$.

The event corresponding to $X \le 45$ contains as a subset the event corresponding to $25 \le X \le 45$. Hence

$$P(X \le 45) \ge P(25 \le X \le 45) \ge 0.96$$

Therefore, the probability that X is less than or equal to 45 is at least 96%

(b) Substituting $\mu = 35$ and $\sigma = 2$ in $\mu + k\sigma$, we have

$35 - 2k = 30$ for k or $k = 2.5$

$$1-\frac{1}{k^2}=1-\frac{1}{2.5^2}=0.84$$

Since $\mu + k\sigma = 35 + 5 = 40$, Chebyshev's inequality gives

$$P(30 \le X \le 40) \ge 0.84$$

The event corresponding to $X \ge 30$ contains as a subset the event corresponding to $30 \le X \le 40$. Hence

$$P(X \ge 45) \ge P(30 \le X \le 40) \ge 0.84$$

Thus, the probability that X is greater than or equal to 30 is at least 84%.

Example E3.76

A random variable X has mean $\mu = 85$ and standard deviation $\sigma = 5$.

(a) what inferences can be made from Chebyshev's inequality for $k = 2$ and $k = 3$?

(b) estimate the probability that X lies between 65 and 105.

(c) determine an internal $[a, b]$ about the mean for which the probability that X lies in the interval is at least 99%.

SOLUTION:

(a) We set $k = 2$ and obtain $\mu - k\sigma = 85 - 2(5) = 75$ and $\mu + k\sigma = 85 + 2(5) = 95$.

We can therefore conclude from Chebyshev's inequality that the probability that a value of X lies between 75 and 95 is at least $1 - \frac{1}{2^2} = 0.75$.

$$P(75 \le X \le 95) \ge 0.75$$

If $k = 3$, we find that the probability that X lies between 70 and 100 is at least $1 - \frac{1}{3^2} = 8/9 = 0.889$

(b) Here $k\sigma = 20$ since $85 - 20 = 65$ and $85 + 20 = 105$ or $k(5) = 20$ or $k = 4$.

Hence, by Chebyshev's inequality

$$P(65 \le X \le 105) \ge 1 - \frac{1}{4^2} = 0.94$$

Thus, the probability X lies between 65 and 105 is at least 94%.

(c) Setting $1 - \frac{1}{k^2} = 0.99$, we get $1 - 0.99 = 1/k^2$ or $k = 10$.

Hence the interval is $[85 - 10(5), 85 + 10(5)] = [35, 135]$.

3.7 EMPIRICAL RULE

If the distribution of the data is approximately bell-shaped, then empirical rule applies. The empirical rule is stated as follows:

For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.75% of the observations lie within three standard deviations of the mean.

The empirical rule is illustrated in Fig. 3.12. The bell-shaped distribution, also known as the normal distribution has been described earlier. The empirical rule is applicable to population data as well as sample data.

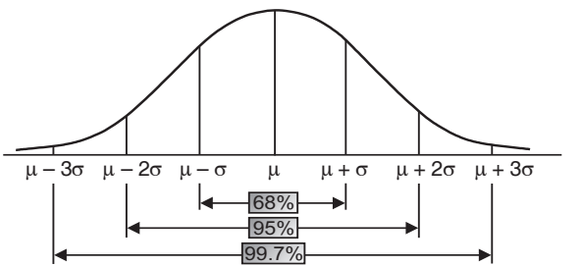


Fig. 3.12: Empirical rule

Example E3.77

The weight distribution of a sample of 10,000 male students in a particular university is bell-shaped with a mean of 40 kg and a standard deviation of 5 kg. Find the percentage of students who weigh approximately 30 to 50 kg.

SOLUTION:

From the given data, we have for the sample

Mean $\bar{x}$ = 40 kg and standard deviation s = 5 kg.

Each of the two points, 30 and 50, is 10 units away from the sample mean of 40 kg. Dividing 10 by the standard deviation (5 kg), we have $\frac{10}{\sigma} = \frac{10}{5} = 2$. The distance between 30 and 40 and between 40 and 50 kg is each equal to 2s as shown in Fig. E3.77, the area from 30 to 50 kg is the area from $\bar{x} - 2s$ to $\bar{x} + 2s$.

Applying the empirical rule, approximately 95% of the male students in the sample weigh between 30 and 50 kg.

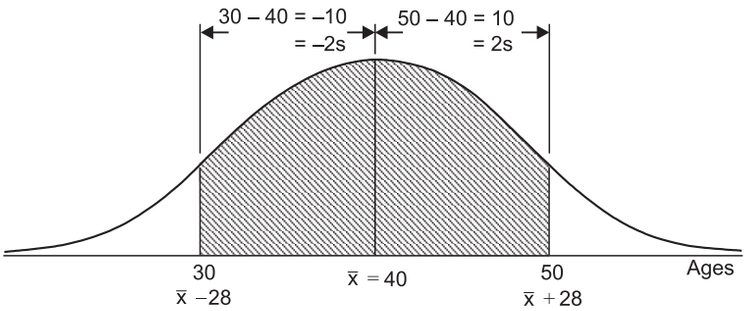


Fig. E3.77: Percentage of male students who weigh between 30 and 50 kg

3.8 THE CENTRAL LIMIT THEOREM

Many commonly used statistical methods depend on the central limit theorem. The central limit theorem is very important result in statistics. The central limit theorem states that if a large enough sample is drawn

from a population, then the distribution of the sample mean is approximately normal, no matter what population the sample was drawn from. If the variable is normally distributed, so is the variable mean, $\bar{x}$. This key fact also holds approximately if x is not normally distributed, provided only that the sample size is relatively large.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ.

Let $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of the sample observations.

Then, if n is sufficiently large,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ approximately}$$

and $S_n \sim N(n\mu, n\sigma^2)$ approximately

The statement of the *central limit theorem* specifies that $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}$, which hold for any sample mean. The sum of the sample items is equal to the mean multiplied by the sample size, that is, $S_n = n\bar{x}$. It follows that $\mu s_n = n\mu$ and $\sigma_{S_n}^2 = n^2\sigma^2/n = n\sigma^2$.

Hence, for a relatively large sample size, the variable is approximately normally distributed, regardless of the distribution of the variable under consideration. The approximation becomes better with increasing sample size.

The further the variable under consideration is from being normally distributed, the larger the sample size must be for a normal distribution to provide an adequate approximation to the distribution of $\bar{x}$. Generally, however, a sample size of 30 or more ($n \geq 30$) is large enough.

The proof of the central limit theorem is difficult and very complex to be included here.

Summarising, the central limit theorem says that if $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution. The normal approximation for $\bar{X}$ depends on the sample size n. In many cases of practical interest, if $n \geq 30$, the normal approximation will be satisfactory regardless of the shape of the population. If $n < 30$, the central limit theorem will work if the distribution of the population is not severely non normal.

Example E3.77

A mechanical component has tensile strength that is normally distributed with mean 75.5 MPa and standard deviation 3.5 MPa. Find the probability for a random sample of $n = 6$, these component specimens will have sample mean tensile strength that exceeds 75.75 MPa.

SOLUTION:

Given $\mu_{\bar{X}} = 75.5$ MPa

Therefore, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{3.5}{\sqrt{6}} = 1.429$

$$P(\bar{X} \geq 75.75) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) = P(Z \geq 0.175) = 1 - P(Z \leq 0.175)$$

$$= 1 - 0.56945 = 0.43055$$

Example E3.78

A normal population has mean 80 and variance 25. How longer must the random sample be if we want the standard error of the sample average to be 1.5?

SOLUTION:

Given $\sigma^2 = 25$ or $\sigma = 5$ and $\sigma_{\bar{X}} = 1.5$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{n}} = 1.5$$

or $$n = \left(\frac{\sigma}{\sigma_{\bar{X}}}\right)^2 = \left(\frac{5}{1.5}\right)^2 = 11.11 \text{ or } \approx 12$$

PROBLEMS

Section 3.1 Random Variables

P3.1 The frequency distribution of the number of orders received per day by a distribution company during the past 3 months (90 days)

Number of orders received per day	2	3	4	5	6
Number of days	11	20	23	22	14

(a) construct a probability distribution table for the number of orders received per day

(b) are the probabilities in *(a)* exact or approximate?

(c) if x denotes the number of orders received on any given day, calculate the following probabilities

(i) $P(x = 3)$

(ii) $P(x \geq 3)$

(iii) $P(2 \leq x \leq 4)$

(iv) $P(x < 4)$.

P3.2 The following table gives the probability distribution of a discrete random variable x.

x	0	1	2	3	4	5	6
P(x)	0.11	0.14	0.18	0.15	0.12	0.09	0.05

Find the following probabilities:

(a) $P(x = 4)$

(b) $P(x \geq 3)$

(c) $P(x < 3)$

(d) $P(3 \leq x \leq 5)$

P3.3 The following table gives the probability distribution of automobiles sold on a given day in an automobile dealership.

Automobile sold	0	1	2	3	4	5	6
Probability	0.05	0.10	0.20	0.30	0.20	0.15	0.05

Calculate the mean and standard deviation for this probability distribution.

P3.4 Table P3.4 gives the probability distribution where X represents the number of times the manufacturing process is recalibrated during a week whenever the quality of the component produced falls below certain specifications. Assume that X has the probability mass function as given in Table P3.4.

Table P3.4

x	0	1	2	3	4
p(x)	0.35	0.25	0.20	0.15	0.05

(a) determine the mean of X

(b) determine the variance and standard deviation for the random variable X.

P3.5 Table P3.5 shows the probability distribution of the number of breakdowns per week for a machine.

Table P3.5

Breakdowns per week	0	1	2	3
Probability	0.15	0.20	0.35	0.30

(a) represent the probability graphically

(b) find the probability that the number of breakdowns for this machine during a given week is

(i) exactly 2

(ii) 0 to 2

(iii) more than 1

(iv) at most 1

P3.6 For each of the following, indicate whether it is or is not a random variable.

(a) the number of automobiles sold at a dealership during a given month

(b) the height of a person

(c) the number of heart beats per minute

(d) the price of an automobile

(e) the time needed to complete a written examination

(f) the number of automobiles a household owns

(g) the number of tails obtained in four tosses of a balanced coin.

P3.7 The number of telephone calls received in a particular sales office between 10:00 am and 11:00 am has the probability function given in Table P3.7.

Table P3.7

Number of telephone calls received, x	0	1	2	3	4	5	6
Probability, P(x)	0.05	0.20	0.25	0.20	0.10	0.15	0.05

(a) is this a probability function?

(b) find the probability that there will be three or more calls

(c) find the probability that there will be an even number of calls.

P3.8 Table P3.8 gives the probability distribution of x, where x denotes the number of defective mechanical components produced in a manufacturing company.

Table P3.8

x	0	1	2	3	4	5
P(x)	0.02	0.20	0.30	0.30	0.10	0.08

Determine the mean and standard deviation of x.

P3.9 Table P3.9 gives the number of telephone calls received per hour in an office and the distribution.

Table P3.9

Number of telephone calls received, x	10	11	12	13	14	15
f (x)	0.08	0.15	0.30	0.20	0.20	0.07

Determine the mean and standard deviation of the number of telephone calls received per hour in that office.

P3.10 Table P3.10 gives the results of the experiment of rolling a die with the discrete random variable number of dots.

Table P3.10

Number of dots, x	1	2	3	4	5	6
Probability, f (x)	1/6	1/6	1/6	1/6	1/6	1/6

Determine the mean and variance of that random variable x.

P3.11 The time to failure of an electronic component is a continuous random variable known to have the density function $0.5e^{-0.5t}$ where t is in years. What is the probability that this component will fail within the first year of operation?

P3.12 Find the mean μ for the probability density function, $\rho(x)$, of the life of a projector bulb, random variable x, is given as:

and
$$\rho(x) = \begin{cases} 0 & ; \quad \text{for } x < 0 \\ 1/900\, e^{-x/900} & ; \quad \text{for } x \geq 0 \end{cases}$$

P3.13 The density function for a continuous random variable x is given as:

$f(x) = 0.25(x - 2)$ for $2 \leq x \leq 5$

Sketch the density and distribution functions.

P3.14 The daily consumption of an electric power of a certain machine (in units of power) is a random variable whose probability density is given by:

$$f(x) = \begin{cases} 1/9\, x e^{-x/3} ; & \text{for } x > 0 \\ 0 ; & \text{for } x \leq 0 \end{cases}$$

Determine the probabilities that on a given day

(a) the consumption of this machine is no more than 6 units

(b) the power supply is inadequate in the daily capacity if the supply is 9 units.

P3.15 The total lifetime (in years) of a certain machine is a random variable whose distribution function is given by:

$$f(x) = \begin{cases} 0 ; & \text{for } x \leq 5 \\ 1 - 25/2^2 ; & \text{for } x > 5 \end{cases}$$

Find the probabilities that such a machine system will have life

(a) beyond 10 years

(b) less than 8 years

(c) anywhere from 12 to 15 years.

P3.16 The probability density function of X is given by

$$f(x) = \begin{cases} 1.25(1 - x^4) ; & 0 < x < 1 \\ 0 ; & \text{otherwise} \end{cases}$$

where the random variable X denotes the clearance (in mm). The clearance is the difference between the radius of the hole drilled in a flat sheet-metal plate and a shaft inserted through the hole. Components with clearances larger than 0.8 mm are to be scrapped. What proportion of components are scrapped?

P3.17 Refer to Problem P3.16. Determine the mean clearance and the variance of the clearance.

P3.18 Determine the mean and variance for the function $f(x)$

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

where X is the uniform random variable on the interval $[a, b]$.

P3.19 A certain radioactive mass emits alpha particles from time to time. The time between emissions (in seconds) is random with the probability density function given by

$$f(x) = \begin{cases} 0.1e^{-0.1x} & ;\ x > 0 \\ 0 & ;\ x \le 0 \end{cases}$$

Determine the median time between emissions.

P3.20 Determine the mean and variance for the following continuous random variable X with probability density function $f(x)$ given by

$$f(x) = \begin{cases} \dfrac{1}{\lambda} e^{-x/\lambda} & ;\ x \ge 0 \text{ and } \lambda > 0 \\ 0 & ;\ \text{otherwise} \end{cases}$$

Section 3.2 Permutations and Combinations

P3.21 *(a)* Determine the number of permutations of the 26 letters of the alphabet.

(b) Determine the number of ways of 6 distinct products that can be lined up on a display shelf.

P3.22 A committee is to be formed consisting of 2 men and 2 women. There are 6 men and 7 women qualified and available to fill this position. How many different committees can be formed out of the finalists?

P3.23 Develop a table of the number of possibilities when 2, 3, 4 and 5 letters of the alphabet are used. Generate the possibilities. English alphabet letters are used with and without repetition.

P3.24 How many 5 digit numbers can be formed with the 10 digits 0, 1, 2, 3, …, 9 if

(a) repetitions are allowed

(b) repetitions are not allowed

(c) the last digit must be zero and repetitions are not allowed.

P3.25 How many ways can an executive committee of 5 can be chosen from a pool of 15 members?

P3.26 A company wants to purchase 4 electronic systems. After all the system models were reviewed, 8 foreign made and 10 U.S. made systems were considered to satisfy all the security requirements for the company.

(a) if the systems are chosen at random, find the probability that 2 of the systems selected are foreign made.

(b) what is the probability that all the four systems selected are U.S. made?

(c) what is the probability that all of the 4 systems selected are foreign made?

(d) what is the probability that at least 2 of the systems are U.S. made?

Permutations

P3.27 A pianist knows five pieces but will have enough stage time to play only four of them. Pieces played in a different order constitute a different program. How many different programs can be arranged?

P3.28 There are six persons on a sinking boat. There are four life jackets on board. How many combinations of survivors are there?

P3.29 A violinist knows four pieces but will have enough time to play only three of them. Pieces played in a different order constitute a different program. Determine the number of different programs that can be arranged.

P3.30 A bin contains 5 white balls, 2 red balls, and 3 green balls. Find the probability of getting either a white ball or a red ball in one drawn from the bin.

P3.31 Five defective parts were unintentionally mixed with 45 non-defective (good) ones. After a thorough mixing, 5 parts are picked simultaneously from the collection of 50.

(a) determine the probability that all 5 parts selected are non-defective ones

(b) what is the probability that no more than 1 part selected is defective?

(c) what is the probability that all 5 parts selected are defective?

P3.32 From 12 items, in how many ways can a selection of 6 be made?

(a) when a specified item is always included

(b) when a specified item is always excluded.

P3.33 Out of 6 manufacturing engineers and 9 design engineers, a committee consisting of 3 manufacturing engineers and 5 design engineers is to be formed. In how many ways can this be done if:

(a) any manufacturing engineer and any design engineer can be included

(b) one particular design engineer must be on the committee

(c) two particular manufacturing engineers cannot be on the committee.

P3.34 Determine the number of ways of splitting 10 persons into 2 groups containing 4 and 6 persons respectively.

P3.35 A committee of 2 men and 3 women needs to be formed out of 7 men and 7 women. Determine the number of ways this can be done if

(a) any man and any woman can be included

(b) one particular woman must be on the committee

(c) two particular men cannot be on the committee.

P3.36 Three cards are drawn in succession from a deck of 52 cards. Determine the number of ways this can be done

(a) with replacement

(b) without replacement.

P3.37 Determine the number of committees of 3 that can be formed from 8 persons.

P3.38 Determine the number of combinations of 4 taken 3 at a time.

P3.39 Find the number of ways of selecting two objects out of four objects.

P3.40 If 3 persons are to be selected randomly from 5 persons for a committee, determine the different possible combinations.

Section 3.3 Discrete Distributions

P3.41 A lot contains 140 electronic components and 20 are selected without replacement for quality testing.

(a) if 20 components are defective, what is the probability that at least one defective component is in the sample?

(b) if 5 components are defective, what is the probability that at least one defective component appears in the sample?

P3.42 A lot of mechanical parts contains 24 in a package. It was determined to test 4 parts in each package and if all 4 parts pass the tests, then they are all accepted. If in a package of parts in which 3 are non-conforming, what is the probability of rejection of these parts?

P3.43 A die is rolled three times. Find the probability of getting one 4 in the three rolls. Also find the probability of getting two 4's, three 4's, and no 4's in three rolls.

P3.44 In a production process the defective rate is 15 per cent. Assuming a random sample of 10 items is drawn from this process, find the probability that two of them are defective.

P3.45 Batches of 50 shock absorbers from a production process are tested for conformance to quality requirements. The mean number of non-conforming absorbers in a batch is 5. Assume that the number of non-conforming shock absorbers in a batch, denoted as x, is a binomial random variable.

(a) find n and p

(b) find $p(x \leq 2)$

(c) find $p(x \geq 49)$.

P3.46 A production process manufactures certain mechanical components for a machine system. On average, 1.5% of the components will not perform up to specifications. When a shipment of 100 components is received at the plant, they are tested, and if more than 2 are defective, the shipment is returned to the manufacturer. What is the probability of returning a shipment?

P3.47 A quality test engineer claims that 1 in 10 of certain manufactured parts is due to material defects. Using binomial distribution and rounding to four decimals, find the probability that at least of 3 out of 5 of the tested parts are due to material defects.

P3.48 At the ABC House Delivery Service, providing high-quality service to its customers is the top priority of the company. The company guarantees a refund of all charges if a package it is delivering does not arrive at it's destination by the specified time. It is known from past data that despite all efforts, 3% of the packages mailed through this company do not arrive at their destinations within the specified time. A corporation mailed 10 packages through ABC House Delivery Service on Monday.

(a) find the probability that exactly one of these 10 packages will not arrive at it's destination within the specified time.

(b) find the probability that at most one of these 10 packages will not arrive at it's destination within the specified time.

P3.49 Five per cent of a large batch of high-strength steel components purchased for a mechanical system are defective.

(a) if seven components are randomly selected, find the probability that exactly three will be defective

(b) find the probability that two or more components will be defective.

P3.50 The number of customers arriving a bank in the next period is a Poisson distribution having a mean of eight. Find the probability that exactly six customers will arrive in the next period.

P3.51 If the probability of a concrete beam failing in compression is 0.05, use the Poisson approximation to obtain the probability that from a sample of 50 beams

(a) at least three will fail in compression

(b) no beam will fail in compression.

P3.52 The number of defects on an electronic component which is used in a computerized system has been found to follow the Poisson distribution with $\lambda = 3$. Find the probability that a randomly selected electronic component will have two or less defects.

P3.53 The number of telephone calls made to a certain company's operator is a Poisson random variable with a mean of 5 calls per hour.

(a) what is the probability that 5 calls are received in one hour?

(b) what is the probability that 10 calls are received in 1.5 hour?

(c) what is the probability that less than 2 calls are received in 1 – 1/2 hours?

P3.54 A photocopying machine in an office breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this machine will have

(a) exactly two breakdowns

(b) at most one breakdown.

P3.55 The proportion of mechanical manufactured parts that are non-conforming is 0.04. Obtain the Poisson approximation to the binomial distribution for the probability of three or fewer non-conforming parts in a sample of 100.

P3.56 An examination consists of five questions, and to pass the examination a student has to answer at least four questions correctly. Each question has three possible answers, of which only one is correct. If a student guesses on each question, what is the probability that the student will pass the test?

P3.57 The probability that a person undergoes a heart operation will recover is 0.6. Determine the probability that of the six patients who undergo similar heart operation:

(a) none will recover

(b) all will recover

(c) half will recover

(d) at least half will recover.

P3.58 In a certain manufacturing process, it was found from quality control inspection that 20% of the machine components produced by the process are defective. Determine the probability that out of 4 machine components selected at random *(a)* 1, *(b)* 0, *(c)* at most 2 machine components will be defective.

P3.59 Given that the probability of an individual patient suffers a bad reaction from injection of a particular serum is 0.001. Determine the probability that out of 2000 individual patients.

(a) exactly 3 individuals will suffer a bad reaction

(b) more than 2 individuals will suffer a bad reaction. Use Poisson distribution.

P3.60 Use binomial distribution and repeat Problem 3.59.

Section 3.4 Continuous Probability Distributions

P3.61 The inside diameter of a finished shaft of uniform diameter is normally distributed with a mean of 4.50 cm and a standard deviation of 0.01 cm. What is the probability of obtaining a diameter exceeding 4.52 cm?

P3.62 The resistance of a foil strain gauge is normally distributed with a mean of 100 ohms and a standard deviation of 0.8 ohm. The specification limits are 100 ±1.0 ohms. What percentage of gauges will be defective?

P3.63 The measurement of the diameter of a special steel pipe is normally distributed with a mean of 5.01 cm and standard deviation of 0.03 cm. The specification limits are 5.00 ±0.05 cm. What percentage of pipes is not acceptable?

P3.64 The yield strength of metal component manufactured is $N(192.36)$. A purchaser of the metal components requires strength of at least 180 psi. The probability that a metal component will meet or exceed the specifications is given by $P(x \geq 180)$. Determine the probability that a metal component meets or exceeds the specifications.

P3.65 Find the probability that the yield strength of the metal components in Problem P3.64 is between 180 and 198 psi.

P3.66 The diameter of a machine shaft produced in a manufacturing company is normally distributed with a mean diameter of 0.001 inches and a standard deviation of 0.0002 inches.

(a) what is the probability that the diameter of the shaft exceeds 0.0013 inches?

(b) what is the probability that a diameter of the shaft is between 0.0007 and 0.0013 inches?

(c) what standard deviation of diameters is needed so that the probability in part (*b*) is 0.995?

P3.67 The mass, m, of a particular machine part is normally distributed with a mean of 66 kg and a standard deviation of 5 kg.

(a) what percentage of the parts will have a mass less than 72 kg?

(b) what percentage of parts will have a mass in excess of 72 kg?

(c) what per cent of the parts will have a mass between 61 and 72 kg?

P3.68 It was determined experimentally that the load X required to break a plate is normally distributed with mean 2.5 and standard deviation 0.24. Determine the probability that

(a) the plate breaks at a load of 2.61 or less

(b) the plate breaks at a load of more than 2.39

(c) the breaking load is at least 2.86

(d) the breaking load is in the interval (2.61, 2.86).

P3.69 Use the Table in Appendix-*E* to determine the following probabilities for the standard normal random variable z.

(a) $P(z < 3.0)$

(b) $P(z > -2.15)$

(c) $P(-3 < z < 3)$

(d) $P(0 < z < 1)$

(e) $P(z > 3)$

P3.70 Assume z has a standard normal distribution. Use Appendix-*E* table to determine the value for z that solves each of the following:

(a) $P(Z < z) = 0.5$

(b) $P(Z > z) = 0.9$

(c) $P(-z < Z < z) = 0.99$

(d) $P(-z < Z < z) = 0.68$

(e) $P(-1.24 < Z < z) = 0.8$

P3.71 The diameter of a component in a machine system is normally distributed with mean 0.2508 cm and standard deviation 0.0005 cm. The specifications on the component are 0.25 ±0.0015 cm. Determine the proportion of components confirms to specifications.

P3.72 The mass, μ, of a particular electronic component is normally distributed with a mean of 66 g and a standard deviation of 5 g. Determine

(a) the per cent of components that will have a mass less than 72 g

(b) the per cent of components that will have a mass in excess of 72 kg

(c) the per cent of components that will have a mass between 61 and 72 g.

P3.73 Customers buying copper rods supplied by a certain manufacturer require that the rods be between 9.9 cm and 10.5 cm, inclusive. The manufacturing process is such that the actual rod lengths are well approximated by a normal distribution with mean 10.1 cm and standard deviation 0.20 cm. Determine the percentage of the manufacturer's production is acceptable to the customer.

P3.74 Refer to Problem P3.73 and determine what rod length is exceeded by 95% of the manufacturer's product.

Section 3.5 Approximating Probability Distributions

P3.75 In a production lot of 100 components, five of them are found to be non-conforming. Approximate the probability that a random sample of 10 components contains no more than 1 non-conforming component (hypergeometric distribution) by the binomial distribution.

P3.76 The proportion of components manufactured that are non-conforming is 0.04. The probability of three of fewer non-conforming components in a random sample of 100 is given by the binomial distribution. Make an approximation using Poisson distribution.

P3.77 Out of a batch of 2800 electronic components received 25% of them were defective or non-conforming. In a sample of 50 randomly selected, the probability that between 12 and 14 are given by the binomial distribution. Use normal distribution to this binomial distribution.

P3.78 A lot of electronic components are known to contain $p = 0.12$ non-conforming. In a sample of 100, the probability that the fraction non-conforming is $0.10 \leq p \leq 0.20$ is given by the normal distribution. Make an approximation.

P3.79 The number of telephone calls received in a specified time by an operator in a manufacturing company is Poisson distribution with mean $\lambda = 14$. Determine the probability that between 10 and 18 calls are received by the operator in that specified time using normal approximation.

P3.80 Determine the probability that in a sample of 10 machine components chosen at random, exactly two will be defective by using

(a) the binomial distribution

(b) the Poisson approximate to the binomial distribution.

Given that 10% of the machine components produced in that manufacturing process are defective.

P3.81 In some manufacturing process that a production of 200 components contains 5 components that do not meet the quality specifications. Determine the probability that a random sample of 10 components will contain no non-conforming components. Use binomial approximation to the hypergeometric.

P3.82 Assume that an examination has 10 questions of the type true or false. If the student taking such an examination guesses at all 10 questions, determine the probability that the student answers either seven or eight questions correctly.

P3.83 *(a)* If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 35 and 45 inclusive.

(b) If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 35 and 45 exclusive.

P3.84 In a certain university, 25% of the students are over 21 years of age. In a sample of 400 students, what is the probability that more than 110 of them are over 21 years of age?

P3.85 A soft drink company conducted a taste survey marketing a new soft drink. The results of this survey showed that 70% of the people who tried the drink liked it. Encouraged by this result, the company decided to market the new soft drink. Assume that 70% of all people like this drink. On a certain day, 100 customers bought this soft drink.

(a) find the probability that exactly 65 out of 100 customers will like this drink

(b) find the probability that exactly 60 or less of the 100 customers will like this drink

(c) find the probability that exactly 75 to 80 of the 100 customers will like this drink

P3.86 The length of a bolt manufactured in a certain manufacturing process has a mean of 50 mm and standard deviation of 0.45 mm. Determine the largest possible value for the probability that the length of the bolt is outside the interval 49.1 — 50.9 mm.

P3.87 In a mechanical engineering design class, the mean for the final examination scores is 75 and the standard deviation is 5. Using Chebyshev's theorem, find the percentage of students who scored between 60 and 90.

P3.88 Given that X is a random variable with the $N(\mu, \sigma^2)$ distribution. Denoting the standardised X by Z, make the comparisons between the actual and the Chebyschev's bounds for $Z = 1$, 2, and 3($k = 1$, 2, and 3).

P3.89 In mechanics class, the mean for the midterm scores is 65 and the standard deviation is 8. Using Chebyshev's theorem, find the percentage of students who scored between 49 and 81.

P3.90 The 2007 gross sales of all firms in a large city have a mean of 3.3 million and a standard deviation of 0.6 million. Using Chebyshev's theorem, find at least what percentage of firms in this city had 2007 gross sales of

(a) \$2.1 to \$4.5 million

(b) \$1.8 to \$4.8 million

(c) \$1.5 to \$5.1 million

P3.91 *(a)* Let X be a random variable with mean $\mu = 50$ and standard deviation $\sigma = 6$. Use Chebyshev's inequality to find a value for b for which $P(50 - b \le X \le 50 + b) \ge 0.95$.

(b) Let X be a random variable with mean $\mu = 70$ and unknown standard deviation σ. Use Chebyshev's inequality to find a value for σ for which $P(65 \le X \le 75) \ge 0.90$.

P3.92 The age distribution of a sample of 6000 persons is bell-shaped with a mean of 50 years and a standard deviation of 10 years. Determine the approximate percentage of people who are 30 to 70 years old.

P3.93 A manufacturing company manufactures steel rods of 100 cm length and a standard deviation of 10 cm. The distribution of the rod is normal. Find the probability that a random sample of $n = 25$ rods will have an average length less than 95 cm.

P3.94 The following table gives the probability mass function of X, where X denotes the number of defects in a 1 m length of aluminium wire. One hundred wires are sampled from this population. Find the probability that the average number of defects per wire in this sample is less than 0.5.

x	0	1	2	3
P(X = x)	0.48	0.39	0.12	0.01

P3.95 Suppose that a random sample of size $n = 12$ is taken from the uniform distribution on the interval [0, 1]. Determine the value of $P\left(\left|\bar{X}_n - \frac{1}{2}\right| \le 0.1\right)$.

P3.96 Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2 \; ; & 4 \le x \le 6 \\ 0 \quad ; & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 50$. Use central limit theorem.

P3.97 At a large university, the mean age of the students is 21.3 years, and the standard deviation is 4 years. A random sample of 64 students is drawn. What is the probability that the average age of these students is greater than 22 years?

P3.98 The GPAs of all 5540 students enrolled at a university have an approximate normal distribution with a mean of 3.02 and a standard deviation of 0.29. Let $\bar{x}$ be the mean GPA of a random sample of 48 students selected from this university. Determine the mean and standard deviation of $\bar{x}$ and comment on the shape of its sampling distribution.

REVIEW QUESTIONS

1. Explain the meaning of probability distribution of a discrete random variable. What are the various ways to present the probability distribution of a discrete random variable?
2. Briefly explain the two-characteristics (conditions) of the probability distribution of a discrete random variable.
3. Define the terms mean and standard deviation of a discrete random variable.
4. Briefly explain the following:
   (a) binomial experiment
   (b) trial
   (c) binomial random variable
5. Describe the parameters of the binomial probability distribution.
6. Define the following:
   (a) *n*-factorial
   (b) combinations
   (c) permutations

7. What are the conditions that must be satisfied to apply Poisson probability distribution?
8. What is the parameter of the Poisson distribution?
9. Define the following terms:
   (a) Bernoulli trial
   (b) Binomial parameters
   (c) Binomial probability distribution
   (d) continuous random variable
   (e) discrete random variable
   (f) random variable
   (g) z-value or z-score
10. Describe the difference between the probability distribution of a discrete random variable and that of a continuous random variable.
11. Explain the main characteristics of a normal distribution.
12. Describe the standard normal distribution curve.
13. Describe the parameters of the normal distribution.
14. Describe the conditions for the normal distribution to be used as an approximation to the binomial distribution.

STATE TRUE OR FALSE

1. A random variable is a quantitative variable whose value depends on chance. (True/False)
2. A discrete random variable is a random variable whose possible values cannot be listed. (True/False)
3. Probability distribution is a listing of the possible values and corresponding probabilities of a discrete random variable, or a formula for the probabilities. (True/False)
4. Probability histogram is a graph of the probability distribution that displays the possible values of a discrete random variable on the horizontal axis and the probabilities of those values on the vertical axis. (True/False)
5. For any discrete random variable, $\Sigma P(X = 1) = 1$. (True/False)
6. The term expected value and expectation are commonly used in place of the term mean. (True/False)
7. $k! = k(k-1) \cdots 2 \cdot 1$. (True/False)
8. $0! = 1$. (True/False)
9. In Bernoulli trials, the trials are not independent. (True/False)
10. In Bernoulli trials, the probability of a success, called the success probability, remains the same from trial to trial. (True/False)
11. In Bernoulli trials, the experiment (each trial) has more than two possibilities. (True/False)
12. The binomial distribution is the probability distribution for the number of failures in a sequence of Bernoulli trials. (True/False)
13. In Bernoulli trials, the number of outcomes that contain exactly x successes equals the binomial coefficient $\begin{Bmatrix} n \\ x \end{Bmatrix}$. (True/False)

14. The mean of a binomial random variable with parameters n and p is $\mu = np$. (True/False)
15. The standard deviation of binomial random variable with parameters n and p is $\sigma = \sqrt{np(1-p)}$. (True/False)
16. The mean of a Poisson random variable with parameter λ is $\mu = \lambda$. (True/False)
17. The standard deviation of a Poisson random variable with parameter λ is $\sigma = \sqrt{\lambda}$. (True/False)
18. A random variable is a quantitative variable whose value depends on chance. (True/False)
19. A discrete random variable is a random variable whose possible values cannot be listed. (True/False)
20. The sum of the probabilities of the possible values of a discrete random variable equals 0. (True/False)
21. The number of possible permutations of m objects among themselves is $m!$. (True/False)
22. The number of possible samples of size n from a population of size N is NC_n. (True/False)
23. For any two events, the probability that one or the other of the events occurs equal the sum of the two individual probabilities. (True/False)
24. For any event, the probability that it occurs equals 1 minus the probability that it does not occur. (True/False)
25. Data obtained by observing values of one variable of a population are called univariate data. (True/False)
26. Data obtained by observing values of two variables of a population are called bivariate data. (True/False)
27. A frequency distribution for bivariate data is called contingency table, or two-way table. (True/False)
28. The joint probability equals the product of the marginal probabilities. (True/False)
29. For a normally distributed variable, the percentage of all possible observations within any specified range equals the corresponding area under its associated normal curve, expressed as a percentage. (True/False)
30. A normally distributed variable having mean 0 and standard deviation 1 is said to have the standard normal distribution. (True/False)
31. A normal distribution is completely determined by the mean and standard deviation. (True/False)
32. The shape of a normal distribution is completely determined by its standard deviation.(True/False)
33. The total area under the standard normal curve is –1. (True/False)
34. The standard normal curve extends indefinitely in both directions, approaching, but never touches, the horizontal axis as it does so. (True/False)
35. The standard normal curve is symmetric about 0. (True/False)
36. Almost all the area under the standard normal curve lies between –1 and +1. (True/False)
37. For a normally distributed variable, one can determine the percentage of all possible observations that lie within any specified range by first converting z-scores and then obtaining the corresponding area under the standard normal curve. (True/False)
38. The rule of thumb for using the normal approximation to the binomial is that np and $n(1-p)$ are 5 are greater. (True/False)

39. A variable is said to be normally distributed if its distribution has the shape of a normal curve. (True/False)
40. If a variable of a population is normally distributed and is the only variable under consideration, common practice is to say that the population is a normally distributed population. (True/False)
41. The parameters for a normal curve are the corresponding mean and standard deviation of the variable. (True/False)
42. Two variables that have the same mean and standard deviation have the same distribution. (True/False)
43. Two normally distributed variables that have the same mean and standard deviations have the same distribution. (True/False)
44. Two normal distributions that have the same mean are centred at the same place, regardless of the relationship between their standard deviations. (True/False)
45. Two normal distributions that have the same standard deviations have the same shape, regardless of the relationship between their means. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **True** 5. **True** 6. **True** 7. **True** 8. **True** 9. **False** 10. **True**
11. **False** 12. **False** 13. **True** 14. **True** 15. **True** 16. **True** 17. **True** 18. **True** 19. **True** 20. **False**
21. **True** 22. **True** 23. **False** 24. **True** 25. **True** 26. **True** 27. **True** 28. **True** 29. **True** 30. **True**
31. **True** 32. **True** 33. **False** 34. **True** 35. **True** 36. **False** 37. **True** 38. **True** 39. **True** 40. **True**
41. **True** 42. **False** 43. **True** 44. **True** 45. **True**

CHAPTER 4

Sampling Distributions

This chapter introduces the reader the basic sampling distributions. Inferential statistics is used here to draw conclusions about a population, based on a probabilistic model of random samples of the population. Since different random samples will most likely give different estimates, some knowledge of the variability of all possible estimates derived from random samples is needed to arrive at reasonable conclusions. *Population* is any finite set of objects being investigated. A sample of objects drawn from a population is a *random sample*.

If the populations contains N elements and a sample of n of them is to be selected, then if each of the $\frac{N!}{(N-n)!n!}$ possible samples has an equal probability of being chosen, the procedure employed is known as *random sampling*. Due to the difficulty in obtaining random samples in practice tables of random numbers are used.

Sampling theory is the study of relationships existing between a population and samples drawn from the population. Sampling theory is useful in finding whether observed differences between two samples are actually due to chance variation or whether they are really significant. A study of inferences made regarding a population by the use of samples drawn from it, along with indications of the accuracy of such inferences using probability theory is known as *statistical inferences*. A *statistic* is a function of the observations in a random sample, which is not dependent on unknown parameters. A *parameter* is in general an unknown constant. For example, the parameters of a normal distribution are μ and σ^2, whereas $\bar{X}$ and s^2 are statistics. The behaviour of sample statistics is needed in order to draw conclusions about a sample. The probabilistic distribution of a random variable defined on a space of random samples is called a *sampling distribution*. The behaviour of the sample statistics is described by a *sampling distribution*. Several sampling distributions are discussed in this chapter and their application to inferential statistics in Chapter 5 (Estimation) and Chapter 6 (Hypothesis Testing).

Suppose that $y_1, y_2, \ldots, y_n$ represents a sample. Then the *sample mean*

$$\bar{y} = \frac{\sum_{t=1}^{n} y_i}{n} \tag{4.1}$$

and the *sample variance*

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{4.2}$$

are both statistics. These quantities are measures of the central tendency and dispersion of the sample, respectively. Sometimes $s = \sqrt{s^2}$, called the *sample standard deviation*, is used as a measure of dispersion.

4.1 PROPERTIES OF SAMPLE MEAN AND VARIANCE

The sample mean $\bar{y}$ is a point estimator of the population mean μ, and the sample variance s^2 is a *point estimator* of the population variance σ^2. An *estimator* of an unknown parameter is a statistic that corresponds to that parameter. Note that a point estimator is a random variable. A numerical value of an estimator computed from sample data, is called an *estimate*. There are several properties required of good *point estimators*. Two of the most important are the following:

1. The point estimator must be *unbiased*. The long-run average of expected value of the point estimator should be the parameter that is being estimated. An unbiasedness is a desirable property and this alone does not always make an estimator a good one.
2. An unbiased estimator must have *minimum variance*. The minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter. Here, we show that $\bar{y}$ and s^2 are unbiased estimators of μ and σ^2, respectively. Using the properties of expectation,

$$E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \tag{4.3}$$

since the expected value of each observation y_i is μ. Thus, $\bar{y}$ is an unbiased estimator of μ.

Considering the sample variances s^2. We get

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = \frac{1}{n-1}E(SS) \tag{4.4}$$

where $SS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *correct sum of squares* of the observation y_i. Now

$$E(SS) = E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]$$

$$\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) = (n-1)\sigma^2 \tag{4.5}$$

Hence, $$E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2 \tag{4.6}$$

Therefore, s^2 is an unbiased estimator of σ^2. The quantity $n - 1$ in Equation (4.5) is called the *number of degrees of freedom* of the sum of squares SS. If y is a random variable with variance σ^2 and $SS = \sum (y_i - \bar{y})^2$ has degrees of freedom, then

$$E\left(\frac{SS}{\nu}\right) = \sigma^2 \tag{4.7}$$

The number of degrees of freedom of a sum of squares is equal to the number of independent elements in that sum of squares. For example, $SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ in Equation (4.5) consists of the sum of squares of the n elements $y_1 - \bar{y}$, $y_2 - \bar{y}$, …, $y_n - \bar{y}$. These elements are not all independent since $SS = \sum_{i=1}^{n} (y_i - \bar{y}) = 0$; in fact, only $n - 1$ of them are independent, implying that SS has $n - 1$ degrees of freedom.

The *standard error* of a statistic is the standard deviation of its sampling distribution. If the standard error involves unknown parameters whose values can be estimated, substitution of these estimates into the standard error results in an *estimated standard error*. Suppose, we are sampling from a normal distribution with mean μ and variance σ^2. Now the distribution of $\bar{X}$ is normal with mean μ and variance σ^2/n, so the *standard error* of $\bar{X}$ is

$$\frac{\sigma}{\sqrt{n}} \tag{4.7a}$$

The *estimated standard error* of $\bar{X}$ is

$$\frac{s}{\sqrt{n}} \tag{4.7b}$$

4.2 POPULATION AND SAMPLING DISTRIBUTIONS

Here, we introduce the basic concept of population distribution and sampling distribution.

4.2.1 Population Distribution

The *population distribution* is the probability distribution of the population data. It is the probability distribution derived from the information on all elements of a population.

4.2.2 Sampling Distribution

A sample provides data for only a portion of an entire population and therefore we cannot expect the sample to yield perfectly accurate information about the population. The value of a population parameter is always constant. Any population data set has only one value of the population mean, μ. We would expect different samples of the same size drawn from the same population to give different values of the sample mean, $\bar{x}$.

The sample mean, $\bar{x}$, is therefore a *random variable* and it possesses a probability distribution. For each sample we can find a statistic, such as the mean, standard deviation, etc., which will vary from sample to sample. As a consequence, we obtain a distribution of the statistic which is called its *sampling distribution*. Sample statistics such as the mean, median, mode and standard deviation all possess sampling distributions. In general, the probability distribution of a sample statistic is called its sampling distribution. The probability distribution of $\bar{x}$ is called its sampling distribution. It lists the values that $\bar{x}$ can assume and the probability of each value of $\bar{x}$.

4.3 SAMPLING AND NONSAMPLING ERRORS

Since a sample provides data for only a portion of an entire population, we cannot expect the sample to give perfectly accurate information about the population. Hence, we can expect a certain information about the population. Hence, we can expect a certain amount of error called *sampling error* - will result simply because we are sampling. *Sampling error* is the difference between the value of a sample statistic and the value of the corresponding population parameter. In the case of the mean,

$$\text{Sampling error} = \bar{x} - \mu \tag{4.8}$$

assuming that the sample is random and no nonsampling error has been made.

A sampling error occurs due to chance. The errors that occur in the collection, recording and tabulation of data are called *nonsampling errors*. Such errors occur due to human mistakes and not chance. The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean $\bar{x}$.

Example E4.1

The mean ages of all students in a large university follow a distribution that is skewed to the right is 24 years and a standard deviation of 4 years. Find the probability that the mean age for a random sample of 36 students would be

(a) between 23 and 25 years

(b) less than 23 years.

SOLUTION:

Given the population mean = 24 years, and $n = 36$.

The standard deviation of the sample mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{36}} = 0.6667 \text{ years}$$

(a) For $\bar{x} = 23$: $z = \dfrac{23 - 24}{0.6667} = -1.50$

For $\bar{x} = 25$: $z = \dfrac{25 - 24}{0.6667} = 1.50$

$$P(23 < \bar{x} < 25) = P(-1.50 < z < 1.50) = (0.933193 - 0.066807) = 0.86639$$

(from the table in Appendix-*E*)

(b) For $\bar{x} = 23$: $z = \dfrac{23 - 24}{0.6667} = -1.50$

$P(\bar{x} < 25) = P(z < -1.50) = 0.066807$ (from the table in Appendix-*E*).

4.4 MEAN AND STANDARD DEVIATION OF *x*

For a variable *x* and a given sample size, the distribution of the variable $\bar{x}$ is called *the sampling distribution of the sample mean*.

The mean and standard deviation of the sampling distribution of $\bar{x}$ are called the *mean and standard deviation of* $\bar{x}$ and are denoted by $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ respectively.

The standard deviation of $\bar{x}$ is also called the *standard error of* $\bar{x}$. There is a simple relationship between the mean of the variable $\bar{x}$ and the mean of the variable under consideration. They are equal, or $\mu_{\bar{x}} = \mu$. Thus, for any particular sample size, the mean of all possible samples means equals the populations mean. This equality holds true regardless of the size of the sample.

Hence, the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population.

$$\mu_{\bar{x}} = \mu \tag{4.9}$$

The sample mean $\bar{x}$, is called an *estimate* of the population mean, μ. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that simple statistic is said to be an *unbiased estimator*. For the sample mean $\bar{x}$, $\mu_{\bar{x}} = \mu$. Therefore, $\bar{x}$ is an unbiased estimator of μ.

Standard Deviation of the Sample Mean

For samples of size *n*, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size.

Hence, $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.10}$$

The above Eq. (4.10) assumes the sampling is done with replacement from a finite population or when it is done from an infinite population. When sampling is done without replacement from a finite population, the appropriate formula is

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}} \tag{4.11}$$

where *n* denotes the sample size and *N* the population size.

The sample size is considered to be small compared to the population size if the sample is equal to or less than 5% of the population size.

That is, if $n/N \leq 0.05$.

When the sample size is small relative to the population size, there is little difference between sampling with and without replacement. In such cases, the two formulae Eqs. (4.10) and (4.11) for $\sigma_{\bar{x}}$ yield almost

the same values. However, in most practical applications, the sample size is small relative to the population size.

The possible sample means cluster more closely around the population mean as the sample size increases, and therefore the larger the sample size, the smaller the sampling error tends to be in estimating a population mean by a sample mean.

The spread of the sampling distribution of $\bar{x}$ is smaller than the spread of the corresponding population distribution. In other words, $\sigma_{\bar{x}} < \sigma$. The standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases.

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is said to be a *consistent estimator*. From Eq.(4.10), $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$, it is clear that as n increases, the value of $\sqrt{n}$ also increases and as a result, the value of $\dfrac{\sigma}{\sqrt{n}}$ decreases. Hence, the sample mean $\bar{x}$ is a consistent estimator of the population mean, μ.

Example E4.2

A population random variable X has mean 120 and standard deviation 15. Find the mean and standard deviation of the sample mean $\bar{X}$ for random samples of size 5 drawn with replacement.

SOLUTION:

For the population, $\mu = 120$, $\sigma = 15$.

The mean $\mu_{\bar{X}}$ and the standard deviation $\sigma_{\bar{X}}$ of $\bar{X}$ are given by

$$\mu_{\bar{X}} = \mu = 120$$

and $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{5}} = 6.7082$$

Example E4.3

A population random variable X has mean 120 and standard deviation 15. Find the mean and standard deviation of $\bar{X}$ if the population size is 300 and the samples of size 5 are drawn without replacement.

SOLUTION:

Here $N = 300$ and $n = 5$

Mean $\mu_{\bar{X}} = \mu = 120, \sigma = 15.$

and the standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{15}{\sqrt{5}}\sqrt{\frac{300-5}{300-1}} = 6.6632$$

Example E4.4

Let $S = \{1, 3, 7, 9\}$. Find the probability distribution of the sample mean $\bar{x}$ for random samples of size 2 drawn with replacement.

SOLUTION:

Since S has 4 elements, there are $4^2 = 16$ random samples of size 2 drawn with replacement. These pairs and their average values are given in Table P4.4.

Table P4.4

Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$
(1, 1)	1	(1, 3)	2	(1, 7)	4	(1, 9)	5
(3, 1)	2	(3, 3)	3	(3, 7)	5	(3, 9)	6
(7, 1)	4	(7, 3)	5	(7, 7)	7	(7, 9)	8
(9, 1)	5	(9, 3)	6	(9, 7)	8	(9, 9)	9

The probability distribution of $\bar{x}$ is given in Table P4.4(a)

Table P4.4(a)

$\bar{x}$	1	2	3	4	5	6	7	8	9
$p(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Example E4.5

Let $S = \{1, 3, 7, 9\}$

(a) list all samples of size 3, without replacement

(b) how many samples, without replacement, are there of size 4, size n?

SOLUTION:

(a) A sample of size 3, without replacement is a subset of S containing 3 elements. There are

$$\binom{4}{3} = 4$$

Subsets: $\{1, 3, 7\}$, $\{1, 3, 9\}$, $\{1, 7, 9\}$, $\{3, 7, 9\}$.

(b) For $n = 1, 2, 3, 4$, there are $\binom{4}{n}$ samples of size n; for $n > 4$, there are no samples of size n.

Example E4.6

Let $S = \{1, 3, 7, 9\}$. Find the probability distribution of the sample mean $\bar{x}$ for random samples of size 2 drawn without replacement.

SOLUTION:

Since S has 4 elements, there are $\binom{4}{2} = 6$ random samples of size 2 drawn without replacement. These, their average value, and the probability distribution of $\bar{x}$ are given in Tables P4.6 and P4.6(a).

Table P4.6

Sample	$\bar{x}$
(1, 3)	2
(1, 7)	4
(1, 9)	5
(3, 7)	5
(3, 9)	6
(7, 9)	8

Table P4.6(a)

$\bar{x}$	$p(\bar{x})$
2	1/5
4	1/6
5	2/6
6	1/6
8	1/6

Example E4.7

How many teams of 5 students can be randomly selected from a class of 10 female students and 15 male students? *(a)* how many teams will have all male students? *(b)* how many teams will have all female students? *(c)* how many teams will have 3 female students and 2 male students?

SOLUTION:

The number of 5-student teams is the number of ways at 5 students can be selected from a class of 25 students, or the number of samples of size 5 that can be selected, without replacement, from a population of size 25, which is

$$\binom{25}{5} = 53,130 \text{ (from the table in Appendix-}B\text{)}$$

(a) The number of teams that will have all male students is

$$\binom{15}{5} = 3003 \text{ (from the table in Appendix-}B\text{)}$$

(b) The number of teams that have all female students is

$$\binom{10}{5} = 252 \text{ (from the table in Appendix-}B\text{)}$$

(c) The number of teams that have 3 female students and 2 male students is

$$\binom{10}{3}\binom{15}{2} = (120)(150) = 12,600 \text{ (see Appendix-}B\text{).}$$

Example E4.8

Determine the most likely breakdown of male students and female students in a team of 5 randomly selected from 15 male and 10 female students.

SOLUTION:

The ratio of 15 male students to 10 female students is 3 to 2. Hence, a team of 3 male and 2 female students would occur at random. From Problem P4.7, we have

5 male students in a team = 3003

5 female students in a team = 252

3 female students and 2 male students in a team = 12600

In a similar manner, we obtain the following counts: (use the table in Appendix-*B*)

$$\text{1 male and 4 female students in a team} = \binom{15}{1}\binom{10}{4} = (15)(210) = 3150$$

$$\text{3 male and 2 female students in a team} = \binom{15}{3}\binom{10}{2} = (455)(45) = 20{,}475$$

$$\text{4 male and 1 female students in a team} = \binom{15}{4}\binom{10}{1} = (1365)(10) = 13{,}650$$

4.5 SHAPE OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

In Section 4.4, we described the sampling distribution of the sample mean, that is, the distribution of the variable $\bar{x}$. It was shown there that the mean and standard deviation of $\bar{x}$ can be expressed in terms of the sample size and the population mean and standard deviation:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In this section, we describe the shape of the sampling distribution of $\bar{x}$ as related to the following two cases:

1. The population from which samples are drawn has a normal distribution.
2. The population from which samples are drawn does not have a normal distribution.

4.5.1 Sampling from a Normally Distributed Population

If the population variable x of a population is normally distributed with mean μ and standard deviation σ, then the sampling distribution of the sample mean, $\bar{x}$, will also be normally distributed with the mean μ and standard deviation $\sigma/\sqrt{n}$.

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

and $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \frac{n}{N} \le 0.05 \tag{4.12}$$

From the sampling distribution of $\bar{x}$ for different sample sizes, n, it can be observed that the spread of the sampling distribution of $\bar{x}$ decreases as the sample size increases.

Example E4.9

The lengths of all machine parts made by a company have a distribution that is skewed to the right with a mean of 68 mm and a standard deviation of 4 mm. Find the probability that the mean length of a random sample of 100 parts produced by this company would be

(a) less than 67.8 mm

(b) between 67.5 mm and 68.7 mm

(c) within 0.6 mm of the population mean

(d) lower than the population mean by 0.5 mm or more.

SOLUTION:

Given $\mu = 68$ mm, $\sigma = 4$ mm and $n = 100$.

The standard deviation of the sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{100}} = 0.4 \text{ mm}$$

(a) For $\bar{x} = 67.8$ mm: $z = \dfrac{67.8 - 68}{0.4} = -0.50$

$P(\bar{x} < 67.8) = P(z < -0.50) = 0.308538$ (from the table in Appendix-*E*).

(b) For $\bar{x} = 67.5$: $z = \dfrac{67.5 - 68}{0.4} = -1.25$

For $\bar{x} = 68.7$: $z = \dfrac{68.7 - 68}{0.4} = 1.75$

$P(67.5 \le \bar{x} \le 67.8) = P(-1.25 \le z \le 1.75) = (-0.105650 + 0.959941) = 0.85429$

(from the table in Appendix-*E*).

(c) $P(\bar{x}$ within 0.6 mm of $\mu) = P(67.4 \le \bar{x} \le 68.6)$

For $\bar{x} = 67.4$: $z = \dfrac{67.4 - 68}{0.4} = -1.50$

For $\bar{x} = 68.6$: $z = \dfrac{68.6 - 68}{0.4} = 1.50$

$P(67.4 \le \bar{x} \le 68.6) = P(-1.50 \le z \le 1.50) = 0.933193 - 0.066807 = 0.86639$

(from the table in Appendix-*E*).

(d) $P(\bar{x}$ lower than μ by 0.5 mm or more$) = p(\bar{x} \le 67.5)$

$P(\bar{x} \le 67.5) = P(z \le -1.25) = 0.10565$ (from the table in Appendix-*E*).

For $\bar{x} = 67.5$: $z = \dfrac{67.5 - 68}{0.4} = -1.25$

4.5.2 Sampling from a Population that is not Normally Distributed

If the sampling is done from a population that is not normally distributed, then the shape of the sampling distribution of $\bar{x}$ is inferred from central limit theorem (see Chapter 3, Section 3.8).

4.6 APPLICATIONS OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

It was shown in Chapter 3 (Section 3.8) on central limit theorem that for large samples, the sampling distribution of $\bar{x}$ is approximately normal, regardless of the distribution of the variable under considerations. The approximation becomes better with increasing sample size. In general, the farther the variable under consideration is from being normally distributed, the larger the sample size must be for a normal distribution to provide an adequate approximation to the distribution of $\bar{x}$. A sample size of 30 or more ($n \geq 30$) is large enough.

The sampling distribution of the sample mean can be summarised as follows:

If the variable x of a population has mean μ and standard deviation σ, then for samples of size n:

1. The mean of $\bar{x}$ equals the population mean, or $\mu_{\bar{x}} = \mu$.
2. The standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, or $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.
3. If x is normally distributed, so is $\bar{x}$, regardless of sample size.
4. If the sample size is large, $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.

Example E4.10

Refer to Problem P4.2. Suppose the random variable X in Problem P4.2 is approximately normally distributed, determine $P(115 \leq \bar{X} \leq 125)$ for samples of size 5 drawn with replacement.

SOLUTION:

From the solution of P4.2, the mean and standard deviation of $\bar{X}$ are $\mu_{\bar{x}} = 120$ and $\sigma_{\bar{x}} = 6.7082$. $\bar{X}$ is approximately normally distributed. Hence,

$$P(115 \leq \bar{X} \leq 125) = P\left(\frac{115-120}{6.7082} \leq \frac{\bar{X}-120}{6.7082} \leq \frac{125-120}{6.7082}\right)$$

$$= P(-0.745 \leq z \leq 0.745) \qquad \text{(See the table in Appendix-}E\text{).}$$

Where z is the standard normal random variable.

$$= (-0.229650 + 0.770350) = 0.5407$$

Example E4.11

Refer to Problem P4.3. Suppose the random variable X in Problem P4.3 is approximately normally distributed, find $P(115 \leq \bar{X} \leq 125)$ for samples of size 5 drawn with replacement.

SOLUTION:

Refer to the solution of Problem P4.3. The mean and standard deviation of $\bar{X}$ are $\mu_{\bar{x}} = 120$ and $\sigma_{\bar{x}} = 6.66.2$.

$\bar{X}$ is approximately normally distributed. Hence,

$$P(115 \le \bar{X} \le 125) = P\left(\frac{115-120}{6.6632} \le \frac{\bar{X}-120}{6.6632} \le \frac{125-120}{6.6632}\right) = P(-0.75039 \le z \le 0.75039)$$

where z is the standard normal random variable.

Using the standard normal table in Appendix-*E*, we have

$$P(-0.75039 \le z \le 0.75039) = (-0.226627 + 0.773373) = 0.54675$$

4.7 POPULATION AND SAMPLE PROPORTIONS

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. In general, when we pick a sample,

$$\text{Sample proportion of an attribute} = \frac{\text{number of items in the sample having the attribute}}{\text{sample size}} \tag{4.13}$$

Hence, if X is the number of items having a certain attribute in a sample of size n, then *the sample proportion having the attribute* is the random variable X/n. The probability distribution of this statistic is called the *sampling distribution of the proportion*.

The *population* and *sample proportions*, denoted by p and $\hat{p}$, respectively, are calculated as

$$p = \frac{X}{N} \quad \text{and} \quad \hat{p} = \frac{x}{n} \tag{4.14}$$

where N = total number of elements in the population

n = total number of elements in the sample

X = number of elements in the population that possess a specific characteristic

x = number of elements in the sample that possess a specific characteristic

The sampled population can be described by the following probability distribution:

x	Probability
0	1 – p
1	p

The computations for finding the mean and the variance of this population are shown in Table 4.1.

Table 4.1

Value x	Probability p(x)	xp(x)	$x^2p(x)$
0	1 – p	0(1 – p)	$0^2(1-p)$
1	p	1p	1^2p
Σ	1	P	p

From column 3 in Table 4.1, $\Sigma xp(x) = p$.

Hence, the population mean μ is p or $\mu = p$ (4.15)

Similarly the sum in column 4 is $\Sigma x^2 p(x) = p$.

Hence, the variance of the population, σ^2 is given by

$\sigma^2 = \Sigma x^2 p(x) - \mu^2 = p - p^2 = p(1-p)$

Therefore, population standard deviation = $\sqrt{p(1-p)}$ (4.16)

4.8 SAMPLING DISTRIBUTION OF $\hat{p}$

Similar to the sample mean $\bar{x}$, the sample proportion, $\hat{p}$, is also a random variable. It possesses a probability distribution called its *sampling distribution*. It gives the various values that $\hat{p}$ can assume and their probabilities.

4.9 MEAN AND STANDARD DEVIATION OF $\hat{p}$

The *mean of* $\hat{p}$, which is the same as the mean of the sampling distribution $\hat{p}$, is always equal to the population proportion, p.

Hence, $\mu_{\hat{p}} = p$ (4.17)

The sample proportion, $\hat{p}$, is called an *estimator* of the population proportion, p. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that sample statistic is said to be an *unbiased estimator*. Since for the sample population, $\mu_{\hat{p}} = p$, $\hat{p}$ is an unbiased estimator of p. The *standard deviation of* $\hat{p}$, denoted by $\sigma_{\hat{p}}$, is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \tag{4.18}$$

where p is the population proportion, $q = 1 - p$, and n is the sample size.

For large values of n ($n \geq 30$) the sampling distribution is very closely normally distributed. Note that the population is binomially distributed. The Eq. (4.18) is also valid for a finite population in which sampling is with replacement.

However, if $n/N \geq 0.05$, then $\sigma_{\hat{p}}$ is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{N-n}{N-1}} \tag{4.19}$$

where the factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor.

In general, the sample size, n is small compared to the population size, N and Eq.(4.18) is used.

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is called the *consistent estimator*. It is clear from Eq.(4.18) that as n increases, the value of $\sqrt{pq/n}$ decreases and the sample proportion, $\hat{p}$, is said to be the *consistent estimator* of the population proportion, p.

The shape of the sampling distribution of $\hat{p}$ is inferred from the central limit theorem described in Section 3.8 of Chapter 3. According to the central limit theorem, the *sampling distribution of* $\hat{p}$ is approximately normal for a sufficiently large sample size. In the case of a proportions, the sample size is considered to be sufficiently large if np and $n(1-p)$ are both greater than 5. That is, if $np > 5$ and $n(1-p) > 5$. This condition is the same as required for the application of the normal approximation to the binomial probability distribution described in Chapter 3 (Section 3.5).

Example E4.12

A company manufactures engine cylinders. The machine that is used to make these cylinders is known to produce 6% defective cylinders. If a sample of 100 cylinders are selected every week and inspected them for being good or defective. If 8% or more of the cylinders in the sample are defective, the process is stopped and the machine is readjusted. Determine the probability that based on a sample of 100 cylinders the process will be stopped to readjust the machine.

SOLUTION:

Here, $p = 0.06$, $q = 1 - p = 1 - 0.06 = 0.94$ and $n = 100$.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.06(0.94)}{100}} = 0.023749$$

For $\hat{p} = 0.08$, we have

$$z = \frac{(0.08 - 0.06)}{0.023749} = 0.84$$

Therefore, $P(\hat{p} \geq 0.08) = P(z \geq 0.84) = P(1 - z \leq 0.84)$. Refer to the table in Appendix-*E*.

$P(\hat{p} \geq 0.08) = (1 - 0.799546) = 0.20045$

Example E4.13

Let m be the mean annual salary of faculty members in a college for 1995. Assume that the standard deviation of the salaries of these faculty members is \$50,000. What is the probability that the 1995 mean salary of a random sample of 100 faculty members was within \$10,000 of the population mean, m? Assume that $\frac{n}{N} \leq 0.05$.

SOLUTION:

Given $\sigma = \$50{,}000$ and $n = 100$

Standard deviation of the sample mean

$$\mu_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50000}{\sqrt{100}} = \$5000$$

The required probability is

$P[\mu - 10{,}000 \leq \bar{x} \leq \mu + 10{,}000]$

For $\bar{x} = \mu - 10{,}000$: $z = [(\mu - 10{,}000) - \mu]/5{,}000 = -2$

For $\bar{x} = \mu + 10{,}000$: $z = [(\mu + 10{,}000) - \mu]/5{,}000 = 2$

Hence $P(\mu - 10{,}000 \le \bar{x} \le \mu + 10{,}000) = P(-2 \le z \le 2)$

$= (-0.022750 + 0.977250)$ from the table in Appendix-*E*

$= 0.9545$

Example E4.14

A particular city is planning to build a nuclear power plant to generate the electric power. An independent survey fund that 53% of the voters in that city favour the building of that plant. Assume that this result holds true for the population of all the voters in this city.

(a) what is the probability that more than 50% of the voters in a random sample of 200 voters selected from this city will favour the building of this plant?

(b) a city official would like to take a random sample of voters in which over 50% would favour the plant building. How large a sample should be selected so that the city official is 95% sure of this outcome? Assume $\frac{n}{N} \le 0.05$.

SOLUTION:

Given $p = 0.53$, $n = 200$ and $n/N = \le 0.05$

(a) Standard deviation of the sample mean

$$\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.53)(0.47)}{200}} = 0.03529$$

The shape of the sampling distribution is approximately normal. In order to have over 50% in favour in a sample of 200 requires 101 or more in favour of the plant building. Hence, we require

$$P\left(\hat{p} \ge \frac{101}{200}\right) = P(\hat{p} \ge 0.505)$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.505 - 0.53}{0.03529} = -0.71$$

$$P(\hat{p} \ge 0.505) = P(z \ge -0.71) = 1 - P(z \le -0.71) = 1 - 0.238852 = 0.76115$$

(from the table in Appendix-*E*).

(b) $P(z > -1.65) = 1 - P(z < -1.65) = 1 - 0.049471 = 0.95053$ (from the table in Appendix-*E*).

Since $z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$

$$\sigma_{\hat{p}} = \frac{\hat{p} - p}{z} = \frac{0.5 - 0.53}{-1.65} = 0.01818$$

Since $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$$n = \frac{p(1-p)}{\left(\sigma_{\hat{p}}\right)^2} = \frac{0.53(0.47)}{(0.01818)^2} = 753.53 \approx 754$$

Hence, the sample should include at least 754 voters.

Example E4.15

A sample of 49 is picked at random from a population of manufactured steel circular rods. If the standard deviation of the distribution of their diameters is known to be 3 cm, find the standard error of the mean if

(a) the population consists of 1000 steel rods

(b) the population is extremely large

SOLUTION:

(a) We are giving that $n = 49$, $N = 1000$, and $\sigma = 3$. Hence, the standard error of the mean is

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{49}}\sqrt{\frac{1000-49}{1000-1}} = \frac{3}{7}(0.97568) = 0.418$$

(b) If the population size is extremely large (practically infinite), the standard error is given by

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{49}} = \frac{3}{7} = 0.4286.$$

4.10 THE CHI-SQUARE DISTRIBUTION

The *chi-square distribution* is sometimes written as χ^2 *distribution* and read as *chi-square distribution*. The values of a chi-square distribution are denoted by the symbol χ^2, just as the values of the standard normal distribution and the *t*-distribution are defined by *z* and *t*, respectively. A variable has a *chi-square distribution* if its distribution has the shape of a specific type of right-skewed curve, called a *chi-square (χ^2) curve*. There are infinitely many chi-square distributions. The chi-square curve is identified by its number of degrees of freedom.

Let $Z_1, Z_2, \ldots, Z_k$ be normally and independently distributed random variables, with mean $\mu = 0$ and variance $\sigma^2 = 1$. Then the random variable

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

has the probability density function

$$f_{\chi^2}(u) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} u^{(k/2)-1} e^{-u/2} \qquad u > 0$$

$$= 0 \qquad \text{otherwise} \tag{4.20}$$

and is said to follow the chi-square distribution with *k* degrees of freedom abbreviated χ_k^2.

The mean and variance of the χ_k^2 distribution are

$$\mu = k \tag{4.21}$$

and

$$\sigma^2 = 2k \tag{4.22}$$

Several chi-square distributions are shown in Fig. 4.1.

Some basic properties of χ^2-curves are:

1. The total area under a χ^2-curve equals 1.
2. A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. A χ^2-curve is right-skewed.
4. As the number of degrees of freedom becomes larger, χ^2-curves look increasingly like normal curves.

Percentages (and probabilities) for a variable having a chi-square distribution are equal to areas under its associated χ^2-curve.

The chi-square random variable is non-negative, and the probability distribution is skewed to the right. As k increases, the distribution becomes more symmetric. As $k \to \infty$, the limiting form of the chi-square distribution is the normal distribution. The percentage points of the χ^2_k distribution are given in Appendix-*F*. Define $\chi^2_{\alpha,k}$ as the percentage point or value of the chi-square random variable with k degrees of freedom such that the probability that χ^2_k exceed this value is α. That

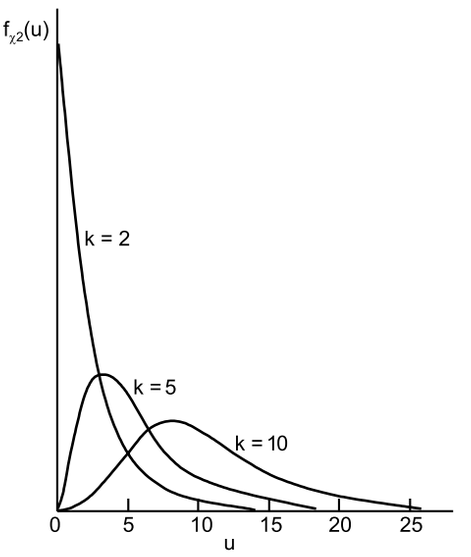


Fig. 4.1: Several χ^2 distributions

$$P\left\{\chi^2_k \geq \chi^2_{\alpha,k}\right\} = \int_{\chi^2_{\alpha,k}}^{\infty} f_{\chi^2}(\mu)\, du = \alpha \tag{4.23}$$

The probability in Eq. (4.23) is shown in Fig. 4.2.

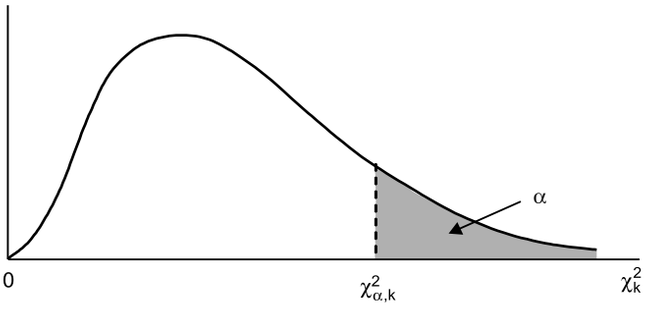


Fig. 4.2: Percentage point $\chi^2_{\alpha,k}$ of the chi-square distribution

The two outside columns of table in Appendix-*F*, labeled *df*, display the number of degrees of freedom. The symbol χ^2_α denotes the χ^2-value having area α to its right under a χ^2-value. Hence, the column headed $\chi^2_{0.995}$, for example, contains χ^2-values having area 0.995 to their right.

Example E4.16

Determine the χ^2-value having area 0.05 to the left for a χ^2-curve with $df = 9$ as shown in Fig. E4.16(*a*).

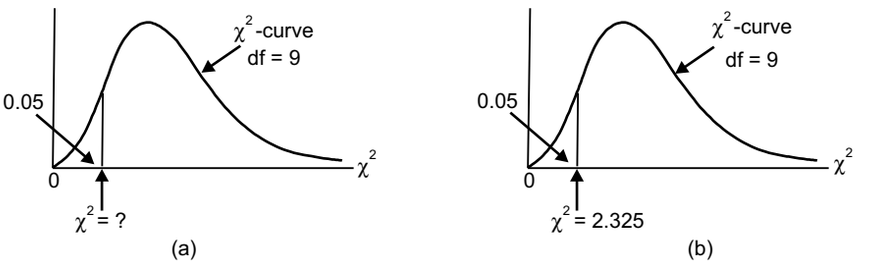


Fig. E4.16

SOLUTION:

The total area under a χ^2-curve equals $1 - 0.05 = 0.95$. Hence, the required χ^2-value is $\chi^2_{0.995}$. From the table in Appendix-*F* with $df = 9$, $\chi^2_{0.995} = 3.325$. Therefore, for a χ^2-curve with $df = 9$, the χ^2-values having area 0.05 to its left is 3.325 as shown in Fig. E4.16(*b*).

Example E4.17

For a χ^2-curve with $df = 25$, determine the two χ^2-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas as shown in Fig. E4.17(*a*).

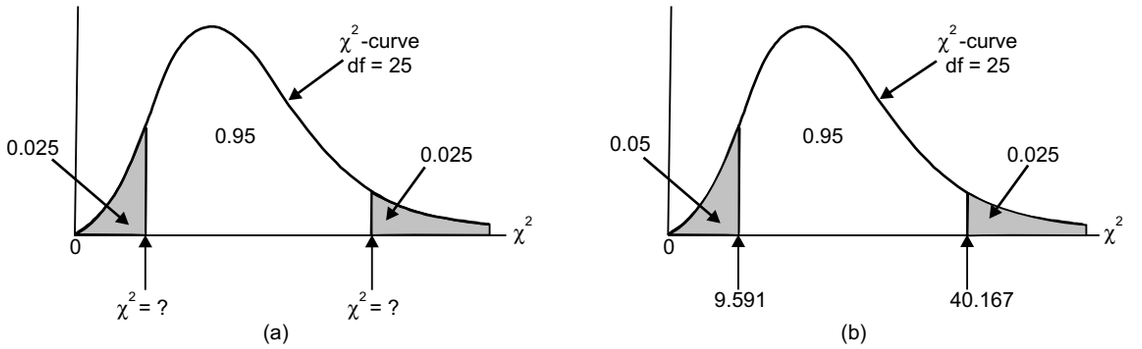


Fig. E4.17

SOLUTION:

First, we find the χ^2-value on the right in Fig. E4.17(*a*). Because the shaded area on the right is 0.025, the χ^2-value on the right is $\chi^2_{0.025}$. From the table in Appendix-*F* with $df = 25$, $\chi^2_{0.025} = 40.647$.

Next, we find the χ^2-value on the left in Fig. E4.17(*a*). Because the area to the left of that χ^2-value is 0.025, the area to its right is $1 - 0.025 = 0.975$. Hence, the χ^2-value on the left is $\chi^2_{0.975}$, which, by table in Appendix-*F* equals 13.120 for $df = 25$.

Consequently, for a χ^2-curve with $df = 25$, the two χ^2-values that the area under the curve into a middle 0.95 area and two outside 0.025 areas are 13.120 and 40.167, as shown in Fig. E4.17(*b*).

4.11 THE t-DISTRIBUTION

The standardised version of $\bar{x}$ has the standard normal distribution. William Gosset in 1908 determined the distribution of the studentised version of $\bar{x}$, a distribution now called *student's t-distribution* or, simply, the *t-distribution*.

There is a different *t*-distribution for each sample size. A particular *t*-distribution is identified by its number of *degrees of freedom* (*df*). For the studentised version of $\bar{x}$, the number of degrees of freedom is 1 less than the sample size, which is indicated symbolically by $df = n - 1$. The mean of the *t*-distribution is 0 just like the standard normal distribution. But unlike the standard normal distribution, whose standard deviation is 1, the standard deviation of a *t*-distribution is $\sqrt{\dfrac{df}{(df-2)}}$ which is always greater than 1. Hence, the standard deviation of a *t*-distribution is larger than the standard deviation of the standard normal distribution.

The number of degree of freedom is the only parameter of the *t*-distribution. A variable with a *t*-distribution has an associated curve, called a *t-curve*. There is a different *t*-distribution for each number of degrees of freedom.

The *t*-distribution is a specific type bell-shaped distribution with a height and a wider spread than the standard normal distribution. As the sample size becomes larger, the *t*-distribution approaches the standard normal distribution.

Let $Z \sim N(0, 1)$ and V be a chi-square random variable with k degrees of freedom. If Z and v are independent, then the random variable

$$\Gamma = \frac{Z}{\sqrt{V/k}} \tag{4.24}$$

has the probability density function

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{\pi k}\,\Gamma(k/2)} \times \frac{1}{[(t^2/k)+1]^{(k+1)/2}} \qquad -\infty < t < \infty \tag{4.25}$$

and is said to follow the *t*-distribution with k degrees of freedom, abbreviated t_k.

The mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k-2)$ for $k > 2$, respectively. Several *t*-distributions are shown in Fig. 4.3.

The basic properties of *t*-curves are:

1. The total area under a *t*-curve equals 1.
2. A *t*-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.
3. A *t*-curve is symmetric about 0.
4. As the number of degrees of freedom becomes larger, *t*-curves look increasingly like the standard normal curve.

Note that if $k = \infty$, the t-distribution becomes the standard normal distribution. A table of percentage points of the t-distribution is given in Appendix-G. If $x_1, x_2, \ldots, x_n$ is a random sample from $N(\mu, \sigma^2)$ distribution, then the quantity

$$t = \frac{x - \mu}{s/\sqrt{n}}$$

is distributed as t with $n - 1$ degrees of freedom.

We will let $t_{\alpha,k}$ be the value of the random variable Γ with k degrees of freedom above which we find an area (or probability) α. Thus $t_{\alpha,k}$ is an upper-tail 100α percentage point of the t-distribution with k degrees of freedom. This percentage point is shown in Fig. 4.4. In the Appendix-G, the α values are the column headings, and the degrees of freedom are listed in the left column. To illustrate the use of the table, note that the t-value with 15 degrees of freedom having an area of 0.05 to the right is $t_{0.05,10} = 1.753$. That is,

$$P(\Gamma_{10} > t_{0.05,15}) = P(\Gamma_{10} > 1.753) = 0.05$$

Since t-distribution is symmetric about zero, we have $t_{1-\alpha} = -t_{\alpha}$; that is, the t-value having an area of $1 - \alpha$ to the right (and therefore an area of α to the left) is equal to the negative of the t-value that has area α in the right tail of the distribution. Therefore, $t_{0.95,15} = -t_{0.05,15} = -1.753$.

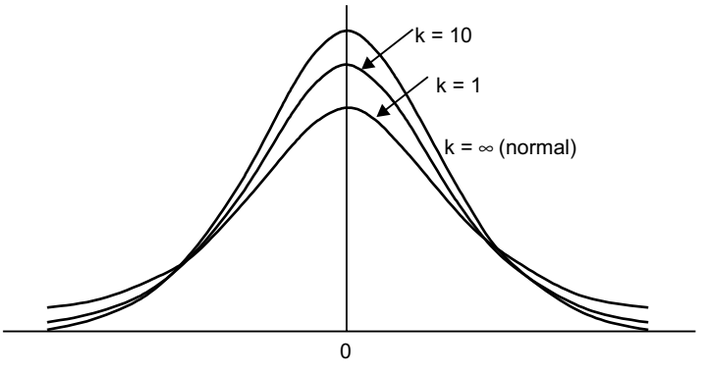


Fig. 4.3: Several *t*-distributions

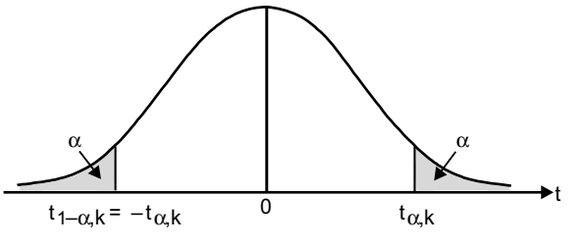


Fig. 4.4: Percentage points of the *t*-distribution

Example E4.18

Find a t-curve with 17 degrees of freedom, determine $t_{0.05}$. In other words, find the t-value having area 0.05 to its right, as shown in Fig. E4.18(a).

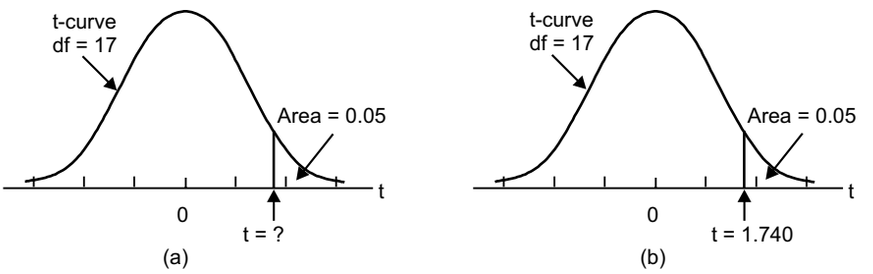


Fig. E4.18

SOLUTION:

The number degrees of freedom is 17, so we first go down the outside columns, (from the table in Appendix-*E*) labeled *df*, to "17". Then, going across that row to the column labeled $t_{0.05}$, we reach 1.740. This number is the *t*-value having area 0.05 to its right, as shown in Fig. E4.18(*b*). In other words, for a *t*-curve with $df = 17$, $t_{0.05} = 1.740$.

4.12 THE F-DISTRIBUTION

A variable is said to have an *F-distribution* if its distribution has the shape of a special type of right-skewed curve, called an *F-curve*. There are infinitely many *F*-distributions. The shape of a particular *F-distribution* curve depends on the number of degrees of freedom. There are two numbers of degrees of freedom for the *F*-distribution curve. The first number of degrees of freedom for an *F*-curve is called the *degrees of freedom for the numerator* and the second the *degrees of freedom for the denominator.*

The random variable *F* is defined to be the ratio of two independent chi-square random variables, each divided by its number of degrees freedom. That is

$$F = \frac{W/u}{Y/v} \tag{4.26}$$

where *W* and *Y* are independent chi-square random variables with *u* and *v* degrees of freedom, respectively.

Let *W* and *Y* be independent chi-square random variables with *u* and *v* degrees of freedom, respectively. Then the ratio

$$F = \frac{W/u}{Y/v} \tag{4.27}$$

has the probability density function

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \qquad 0 < x < \infty \tag{4.28}$$

and is said to follow the *F*-distribution with *u* degrees of freedom in the numerator and *v* degrees of freedom in the denominator. It is usually abbreviated as $F_{u,v}$.

The mean and variance of *F*-distribution are $\mu = v/(v-2)$ for $v > 2$, and

$$\sigma^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)} \qquad v > 4$$

The F random variable is non-negative, and the distribution is skewed to the right. The F-distribution looks very similar to the chi-square distribution as shown in Fig. 4.5; however the two parameters u and v provide extra flexibility regarding shape.

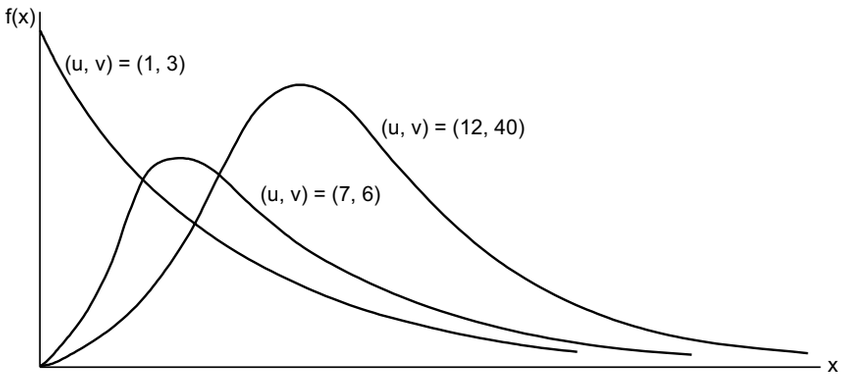


Fig. 4.5: Three *F*-distribution curves

The basic properties of F-curves are:

1. The total area under an F-curve equals 1.
2. An F-curve starts on 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. An F-curve is right-skewed.

For an F-curve with $df = (u, v)$, the F-value having area α to its left equals the reciprocal of the F-value having area α to its right for an F-curve with $df = (v, u)$.

The percentage points of F distribution are given in Appendix-*H*. Let $f_{\alpha,u,v}$ be the percentage point of F-distribution, with the numerator degrees of freedom u and the denominator degrees of freedom v such that the probability that the random variable F exceeds this value is

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)\,dx = \alpha \tag{4.29}$$

This is illustrated in Fig. 4.6.

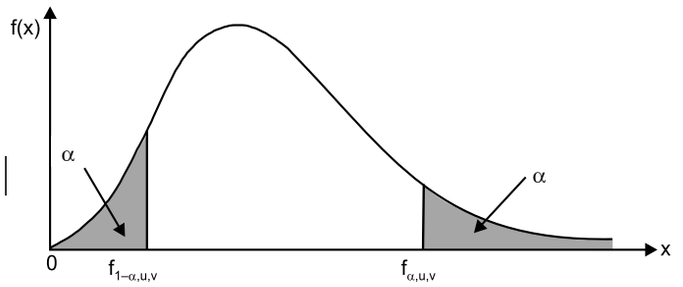


Fig. 4.6: Upper and lower percentage points of the F-distribution

For example, if $u = 6$ and $v = 10$, we find from the table in Appendix-*H*, that

$$P(F > f_{0.05,5,10}) = P(F_{5,10} > 3.22) = 0.05$$

That is, the upper 5 percentage point of $F_{5,10}$ is $f_{0.05,5,10} = 3.22$. The table in Appendix-*H* contains only upper-tail percentage points (for selected values of $f_{\alpha,u,v}$ for $\alpha \leq 0.25$) of the *F*-distribution. The lower-tail percentage point $f_{1-\alpha,u,v}$ can be found as follows:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,u,v}} \tag{4.30}$$

For example, to find the lower-tail percentage point of $f_{0.95,6,10}$, note that

$$f_{0.95,6,10} = \frac{1}{f_{0.05,10,6}} = \frac{1}{4.06} = 0.246$$

Example E4.19

For an *F*-curve with *df*(5, 13), find $F_{0.05}$. That is, find an *F*-value having area 0.05 to its right as shown in Fig. E4.19(*a*).

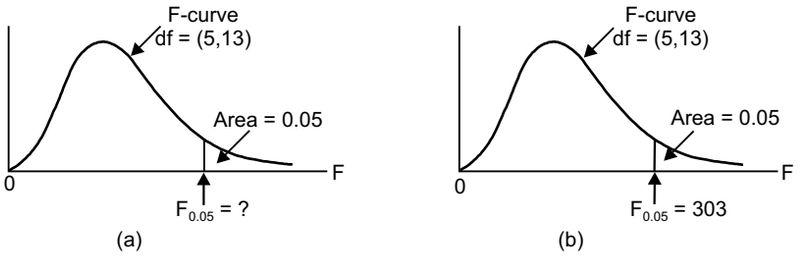


Fig. E4.19

SOLUTION:

To find the *F*-value, we use the table in Appendix-*H*. In this case, $\alpha = 0.05$ (area in the right tail under the *F*-distribution curve). The degrees of freedom for the numerator is 5, and the degrees of freedom for the denominator is 13.

We first go down the degree of freedom to "13". Next, we concentrate on the row for α labeled 0.05. Then, going across that row to the column labeled "5" (from the table in Appendix-*E*), we reach 3.03. This number is the *F*-value having area 0.05 to its right, as shown in Fig. E4.19(*b*). In other words, for an *F*-curve with *df*(5, 15), $F_{0.05} = 3.03$.

Example E4.20

For an *F*-curve with *df* = (50, 8) having an area 0.05 to its left.

SOLUTION:

The required *F*-value is the reciprocal of the *F*-value having area 0.05 to its right for an *f*-curve with *df* = (8, 50). From the table in Appendix-*H*, this latter *F*-value equals 2.13. Consequently, the required *F*-value is $\frac{1}{2.13}$ or 0.4695, as shown in Fig. E4.20.

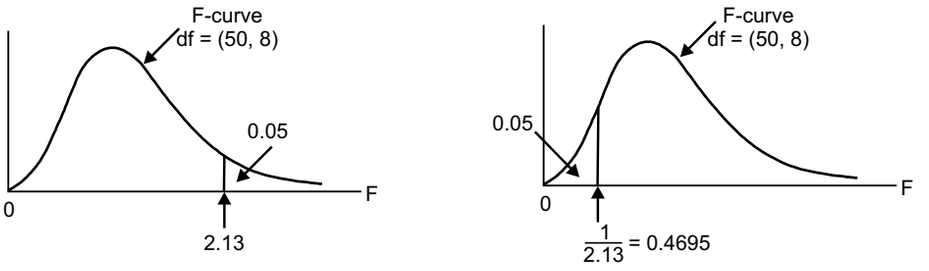


Fig. E4.20

Example E4.21

For an F-curve with $df = (10, 8)$, determine the two F-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas as shown in Fig. E4.21(a).

SOLUTION:

First, we find the F-value on the right in Fig. E4.21(a). Because, the shaded area on the right is 0.025, the F-value on the right is $F_{0.025}$. From the table in Appendix-H, with $df = (10, 8)$, $F_{0.025} = 4.30$.

Next, we find the F-value on the left in Fig. E4.21(a). The F-value is the reciprocal of the F-value having area 0.025 to its right for an F-curve with $df(10, 8)$. From the table in Appendix-H, we find that this latter F-value equals $\dfrac{1}{4.30} = 0.2326$.

Consequently, for an F-curve with $df(10, 8)$, the two F-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas are 0.2326 and 4.30, as shown in Fig. E4.21(b).

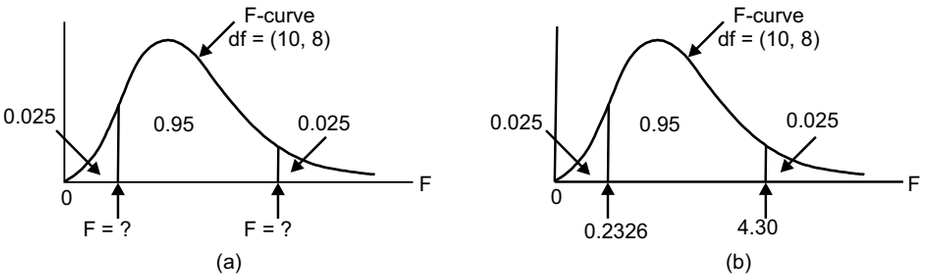


Fig. E4.21

4.13 SUMMARY

In Chapter 3, we discussed the probability distributions of discrete and continuous random variables. In this chapter, we extended the concept of probability distribution to that of a *sample statistic*. A sample statistic is a numerical summary measure calculated for sample data. The numerical summary measures calculated for population data are called *population parameters*. A population parameter is always a constant, whereas a sample statistic is always a random variable. Also, since every random variable must possess a probability distribution, each sample statistic possesses a probability distribution. The probability distribution of a sample statistic is more generally called its *sampling distribution*. In this chapter, we presented the sampling distributions of the sample mean and the sample proportion. The concepts presented in this chapter forms the foundation for the inferential statistics discussed in Chapters 5 and 6.

PROBLEMS

P4.1 A population consists of the values 1, 2 and 5.

(a) list all the possible samples (with replacement) of size $n = 2$ along with the sample means and their individual probabilities.

(b) find the mean of the sampling distribution

(c) does the sample mean target the values of the population mean?

P4.2 A manufacturing company produces machine parts that have a mean tensile strength of 100 MPa and a standard deviation of 10 MPa. The distribution of tensile strength is normal. Find the probability that a tensile strength less than 95 MPa.

P4.3 Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 50$.

P4.4 Assume that the weights of 4000 male students at a university are normally distributed with a mean of 60 kg and standard deviation 3.0 kg. If 100 samples consisting 25 students each are obtained, what would be the expected mean and standard deviation of the resulting sampling distribution of means if sampling were done (*a*) with replacement, (*b*) without replacement?

P4.5 A survey showed that the mean expenditure incurred by a student in 1995 was \$9000 and the standard deviation of the expenditure was \$800. Find the approximate probability that the mean expenditure of 64 students picked at random was

(a) more than \$8820

(b) between \$8800 and \$9120

P4.6 Refer to Problem P4.4. In how many samples of Problem P4. 4 would you expect to find the mean

(a) between 58.8 and 60.3 kg

(b) less than 58.4 kg

P4.7 The length of life (in hours) of a certain type of machine component is a random variable with a mean of 600 hours and a standard deviation of 70 hours. What is the approximate probability that a random sample of 196 machine parts will have a mean life between 588 and 605 hours?

P4.8 Five hundred electronic components have a mean weight of 6.02 g and a standard deviation of 0.3 g. Find the probability that a random sample of 100 electronic components chosen from this group will have a combined weight of

(a) between 596 and 600 g

(b) more than 610 g

P4.9 Assume that the weights of machine parts of a heavy machine are normally distributed with mean 55 kg and standard deviation 2 kg. A number of samples of 100 parts each are taken at random with replacement from the population. Determine

(a) the mean and standard deviation of the sampling distribution of the mean

(b) the probability that the sample mean will differ from the population mean by less than 0.4 kg.

P4.10 It was found in a particular survey that adults spend an average of 10 hours a day at work and commuting. Let the daily work and commute times for all adults have a mean of 10 hours and a standard deviation of 2 hours. Find the probability that the mean of the daily work and commute times for a random sample of 100 adults will be

(a) greater than 10.45 hours

(b) between 9.75 and 10.50 hours

(c) within 0.25 hours of the population mean

(d) lower than the population mean by 0.50 hours or more.

P4.11 A random sample of size 121 is taken from a normal population with $\sigma = 27.5$. What is the probability that the mean of the sample will differ from the mean of the population by 3 or more either way?

P4.12 The diameters of ball bearings manufactured by a company are normally distributed with mean 12 mm and standard deviation 0.1 mm. A sample of 25 ball bearings was taken each day during a given month. Find

(a) the measured standard deviation of the distribution of the sample means

(b) the probability that the sample mean of ball bearings will

(i) exceed 12.01 mm

(ii) be less than 11.98 mm

(iii) lie between 11.98 and 12.1 mm

P4.13 A random sample of size of 81 is taken from a normal population with $\mu = 61.4$ and $\sigma = 7.65$. What is the probability that the mean of the sample will

(a) exceed 62.9

(b) fall between 60.5 and 62.3

(c) be less than 60.6

P4.14 A random sample of 300 air compressors had a standard deviation of 50 months. Determine the probability that the estimate on the population mean lifetime of these compressors will be within 6 months of the true mean from the sample estimate.

P4.15 According a particular survey, the standard deviation of the lengths of time that men with one job are employed during the first 10 years of their career is 100 weeks. Length of time employed during the first 10 years of career is left-skewed variable. For this variable, find the following:

(a) the sampling distribution of the sample mean for simple random samples of 100 men with one job

(b) the probability that the sampling error made in estimating the mean length of time employed by all men with one job by that of a random sample of 100 such men will be at most 25 weeks.

P4.16 In a large university, it was found that the score in statistics class with large number students over a number of years has a distribution with mean 80 and standard deviation 16. A random sample of 64 students of every class for 20 classes is taken. For the sampling distribution of mean, find

(a) the mean and standard deviation

(b) the probability that

(i) $\bar{X} > 84$

(ii) $\bar{X} > 86$

(iii) $\bar{X} \leq 80$

where $\bar{X}$ is the sample mean found for the sample size of 64.

P4.17 According to a particular survey, the mean annual salary of private classroom teachers is \$45,000 and standard deviation of \$9000.

(a) find the sampling distribution of the sample mean for sample size of 81

(b) repeat part *(a)* for samples of 289

(c) is the assumption that classroom teachers salaries are normally distributed to answers in parts *(a)* and *(b)* necessary?

(d) what is the probability that the sampling error made in estimating the population mean salary of all classroom teachers by the mean salary of a sample of 81 classroom teachers will be at most \$1000

(e) repeat part *(d)* for samples of size 289.

P4.18 The weight of electronic components packed in certain containers is a random variable with a mean weight of 16 g and a standard deviation of 0.6 g. If the containers are shipped in boxes of 36, find, approximately, the probability that a randomly picked box will weigh over 585 g.

P4.19 According a particular survey it was found that the standard deviation of the lengths of hospital stay on the intervention ward is 9 days.

(a) for the variable "length of hospital stay", determine the sampling distribution of the sample mean for samples of 100 patients on the intervention ward

(b) the distribution of the length of hospital stay is right-skewed. Does this validate or invalidate the results found in part *(a)*?

(c) find the probability that the sampling error made in estimating the population mean length of stay on the intervention ward by the mean length of stay of a sample of 100 patients will be at most 3 days.

P4.20 The proportion of employed men between the ages 21 and 40 years in a city is 2/3. Suppose random samples of size 25 are drawn with replacement from all men in that city between the ages 21 and 40. What are the mean and standard deviation of the proportion $\hat{p}$ for all such samples?

P4.21 In a particular survey in a large metropolitan city nearly 20% of all students were at least somewhat afraid of being attacked in or nearby their schools. Suppose 20% of all students in grade 10 are afraid of such attacks. If $\hat{p}$ is the proportion of students in a random sample of 49 of grade 10 students who fear such attacks, find the mean and standard deviation of $\hat{p}$.

P4.22 Refer to Problem P4.20. Suppose the city has 250 men between ages 21 and 40 years, and the sampling is without replacement. What are the mean and standard deviation of $\hat{p}$?

P4.23 Suppose 25% of all workers in a state belong to a labour union. What is the probability that in a random sample of 100 workers in that state, at least 18% will belong to a labour union?

REVIEW QUESTIONS

1. Explain the meaning of a population distribution and a sampling distribution.
2. What is sampling error? Does such an error occur only in a sample survey or can it occur both in a sample survey and a census?
3. Explain the meaning of nonsampling error. Do such errors occur only in a sample survey or can they occur both in a sample survey and a census?
4. Describe the condition or conditions that must hold true for the sampling distribution of the sample mean to be normal when the sample size is less than 30.
5. Describe the central limit theorem (Refer to Chapter 3, Section 3.8).
6. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.5 standard deviations of the population mean? (Chapter 3, Section 3.9).
7. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 1.5 standard deviations of the population mean? (Chapter 3, Section 3.9).
8. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.0 standard deviations of the population mean? (Chapter 3, Section 3.9).
9. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 3.0 standard deviations of the population mean? (Chapter 3, Section 3.9).
10. Define the following terms:
    (a) Consistent estimator
    (b) Estimator
    (c) Mean of $\hat{p}$
    (d) Mean of $\overline{x}$
    (e) Population proportion, p
    (f) Unbiased estimator
11. Explain the following terms:
    (a) Sample proportion
    (b) Sampling distribution of $\hat{p}$
    (c) Sampling distribution of $\overline{x}$
    (d) Standard deviation of $\hat{p}$
    (e) Standard deviation of $\overline{x}$
12. Describe very briefly the following distributions:
    (a) t-distribution
    (b) F-distribution
    (c) chi-square distribution

STATE TRUE OR FALSE

1. Sampling error is the error resulting from using a sample to estimate a population characteristic. (True/False)
2. For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called the sampling distribution of the population mean. (True/False)
3. The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean $\bar{x}$. (True/False)
4. For a sample of size n, the mean of the variable $\bar{x}$ equals the mean of the variable under consideration. (True/False)
5. For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration multiplied by the square root of the sample size. (True/False)
6. A normal distribution is determined by the mean and standard deviation. (True/False)
7. The mean of all possible sample means (i.e., of the variable $\bar{x}$) always equals the population mean. (True/False)
8. The larger the sample size, the larger is the standard deviation $\bar{x}$. (True/False)
9. The smaller the standard deviation of $\bar{x}$, the more closely the possible values $\bar{x}$ of (the possible sample means) cluster around the mean of $\bar{x}$. (True/False)
10. For a relatively large sample size, the variable $\bar{x}$ is approximately normally distributed.(True/False)
11. The total area under χ^2-curve equals 1. (True/False)
12. A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so. (True/False)
13. A χ^2-curve is left skewed. (True/False)
14. As the number of degrees of freedom becomes larger, χ^2-curves look decreasingly like normal curves. (True/False)
15. The larger the standard deviation of $\bar{x}$, the more closely the possible values of $\bar{x}$ (the possible sample means) cluster around the mean of $\bar{x}$. (True/False)
16. For samples of size n, the variable $\chi^2 = \left(\frac{n-1}{\sigma^2}\right)s^2$ has the chi-square distribution with $n - 1$ degrees of freedom. (True/False)
17. A variable is said to have a chi-square distribution if its distribution has the shape of a special type of right-skewed curve, called a chi-square curve. (True/False)
18. Different chi-square distributions are identified by their degrees of freedom. (True/False)
19. The total area under an F-curve equals 1. (True/False)
20. An F-curve starts at 0, on the horizontal axis and extends indefinitely to the left, approaching, but never touching, the horizontal axis as it does so. (True/False)
21. An F-curve with $df = (v_1, v_2)$, the F-value having area α to its left equals the reciprocal of the F-value having area α to its right for an F-curve with $df = (v_2, v_1)$. (True/False)
22. An F-distribution and its corresponding F-curve are identified by stating its two numbers of degrees of freedom. (True/False)
23. An F-distribution has two numbers of degrees of freedom. (True/False)

24. The first degree of freedom in F-distribution is called the degrees of freedom for the numerator. (True/False)
25. The second degree of freedom in F-distribution is called the degrees of freedom for the denominator. (True/False)
26. A χ^2-curve looks increasingly like a normal curve as the number of degrees of freedom becomes larger. (True/False)
27. An F-curve is right-skewed. (True/False)
28. For an F-curve with $df = (15, 5)$, the F-value having area 0.05 to its left equals the reciprocal of the F-value having area 0.05 to its right for an F-curve with $df = (5, 15)$. (True/False)
29. The observed value of a variable having an F-distribution must be greater than or equal to 1. (True/False)
30. The total area under a t-curve equals 1. (True/False)
31. A t-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so. (True/False)
32. A t-curve is not symmetric about 0. (True/False)
33. As a number of degrees of freedom becomes larger, t-curves look increasingly like the F-distribution curve. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **True** 5. **False** 6. **True** 7. **True** 8. **False** 9. **True** 10. **True**
11. **True** 12. **True** 13. **False** 14. **False** 15. **False** 16. **True** 17. **True** 18. **True** 19. **True** 20. **False**
21. **True** 22. **True** 23. **True** 24. **True** 25. **True** 26. **True** 27. **True** 28. **True** 29. **False** 30. **True**
31. **True** 32. **False** 33. **False**

CHAPTER 5

Estimation

Inferential statistics is the part of statistics that helps us to make decisions about some characteristics of a population based on sample information. *Estimation* is a procedure by which numerical value or values are assigned to a population parameter based on the information collected from a sample.

In inferential statistics, μ is called the *true population mean* and p is called the *true population proportion*. The value(s) assigned to a population parameter based on the value of a sample statistic is called an *estimate*. The sample statistic used to estimate a population parameter is called an *estimator*.

If the mean of the sampling distribution of a statistic equals to the corresponding population parameter, the statistic is called an *unbiased estimator* of the parameter, otherwise it is called a *biased estimator*. The corresponding values of such statistics are called *unbiased* or *biased estimates* respectively. If the sampling distributions of two statistics have the same mean (or expectation), the statistic with the smaller variance is called an *efficient estimator* of the mean while the other statistic is called an *inefficient estimator*. The corresponding values of the statistic are called *efficient* or *inefficient estimates* respectively.

Statistical inference is divided into two major categories: (1) Parameter estimation and (2) Hypothesis testing (see Chapter 6). Parameter estimation is further classified into two major types: (1) Point estimation and (2) Interval estimation.

5.1 POINT ESTIMATION

Estimation is a procedure by which numerical value or values are assigned to a population parameter based on the information collected from a sample. In inferential statistics, μ is called the *true population mean* and p is called the *true population proportion*. The value(s) assigned to a population parameter based on the value of a simple statistic is called an *estimate* of the population parameter.

If the population is small, we can ordinarily determine μ exactly by first taking a census and then computing μ from the population data. But if the population is large, as it often is in practice, taking a census is generally impractical, extremely expensive, or impossible. Nonetheless, we can usually obtain sufficiently accurate information about μ by taking a sample from the population.

An estimate of a population parameter given by a single number is called *point estimate* of parameter. An estimate of a population parameter is given by two numbers between which the parameter may be

considered to lie is called an *interval estimator* of the parameter. The sample statistic used to estimate a population parameter is called an *estimator*. Hence, the sample mean, $\bar{X}$ is an estimator of the population mean, μ and the sample population, $\hat{p}$ is an estimator of the population proportion, p.

The desired estimators depend on the distribution used. These estimators are not always given by $\bar{X}$ and S^2. The common distributions and their estimators are given below.

Distribution	Parameter	Suggested Estimator(s)
Poisson	λ	$\hat{\lambda} = \bar{X}$
Binominal	p	$\hat{p} = \bar{X}$
Normal	μ, σ^2	$\hat{\sigma}^2 = S^2$ (denominator is $n - 1$)

After a sample is taken and the data are analysed, the result is an *estimate*.

The estimation procedure involves the following steps:

1. Select a sample
2. Collect the required information from the members of the sample
3. Calculate the value of the sample statistic
4. Assign value(s) to the corresponding population parameter

If we select a sample and compute the value of the sample statistic for this sample, the value gives the *point estimate* of the corresponding population parameter.

5.2 INTERVAL ESTIMATION

An interval estimate for a population parameter is called a *confidence interval*. In *interval estimation*, instead of assigning a single value to a population parameter, an interval is constructed around the point estimate and then a probabilistic statement that this interval contains the corresponding population parameter is made. This probabilistic statement is given by the *confidence level*. An interval that is constructed based on the confidence level is called a *confidence interval*. The confidence level associated with a confidence interval states how much confidence we have that the interval contains the true population parameter. The confidence level is denoted by $(1 - \alpha)100\%$. When expressed as probability, it is called the *confidence coefficient* and is denoted by $(1 - \alpha)$. Note that α is called the *significance level*.

A *tolerance interval* is another important type of interval estimate. For a normal distribution, we know that 95% of the distribution is in the interval $\mu - 1.96\,\sigma, \mu + 1.96\,\sigma$.

Confidence and tolerance intervals bound unknown elements of a distribution. A *prediction interval* provides bounds on one (or more) future observations from the population.

Summarising, the purpose of the 3 types of interval estimates are:

A confidence interval bounds population or distribution parameter (such as the mean weight of a person).

A tolerance interval bounds a selected proportion of a distribution.

A prediction interval bounds future observations from the population or distribution.

A point estimate of a parameter consists of a single value with no indication of the accuracy of the estimate. A confidence interval consists of an interval of numbers obtained from a point estimate of the

parameter together with a percentage that specifies how confident we are that the parameter lies in the interval.

A confidence interval estimate of a parameter consists of an interval of number obtained from the point estimate of the parameter together with a 'confidence level' that specifies how confident we are that the interval contains the parameter. This is superior to a point estimate because it provides some information about the accuracy of the estimate whereas a point estimate does not.

Interval estimates indicate the precision or accuracy of an estimate and are therefore preferable to point estimates. A statement of the error of precision of an estimate is often called its *reliability*. An interval estimate consists of two sample statistics, L and U. The probability that the interval formed by L and U contains the true value of a parameter is $1 - \alpha$. For instance, to construct an interval estimate on the parameter θ, we find two statistics, L and U, such that

$$P(L \leq \theta \leq U) = 1 - \alpha \tag{5.1}$$

The resulting interval $[L \leq \theta \leq U]$ is called a $100(1 - \alpha)\%$ confidence interval on unknown parameter θ. If a large number of independent samples were taken from the population being considered, approximately $100(1 - \alpha)\%$ of the intervals formed would be expected to contain the true value of θ. Confidence intervals that contain both an L, or lower side, and a U, or upper side, are called *two-sided confidence intervals*. Confidence intervals that contain only a lower or an upper side are called *one-sided confidence intervals*. A one-sided $100(1 - \alpha)\%$ confidence interval on θ is given by

$$L \leq \theta$$

with the probability property

$$P(L \leq \theta) = 1 - \alpha \tag{5.2}$$

Similarly, the one-sided upper $100(1 - \alpha)\%$ confidence interval on θ is given by $\theta \leq U$ with the probability property.

$$P(\theta \leq U) = 1 - \alpha \tag{5.3}$$

The length of the observed confidence interval is an important measure of the quality of the information obtained from the sample. The half-interval length $\theta - L$ or $U - \theta$ is called the *accuracy* of estimator. The longer the confidence interval, the more confident we are that interval actually contains the true value of θ and the less information we have about the true value of θ.

Example E5.1

The following readings give the weights of 6 persons (in kg) picked at random from college students in a particular college: 50, 52, 55, 60, 65 and 70. Find estimates of the following:

(a) the true (population) mean weight of all the students

(b) the true variance of weights of all the students

(c) the true standard deviation of weights of all the students.

SOLUTION:

$$n = 6$$

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{352}{6} = 58.6667 \text{ kg}$$

$$\Sigma(X_i - \bar{X})^2 = (50 - 58.6667)^2 + (52 - 58.6667)^2 + \cdots = 303.3333$$

$$\text{Variance} = \frac{\Sigma(X_i - X)^2}{n-1} = \frac{303.3333}{6-1} = 60.6667 \text{ kg}^2$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{60.6667} = 7.7889 \text{ kg}$$

(a) The estimate of the true mean weight is $\bar{X}$, that is, 58.6667 kg

(b) The estimate (in kg^2) of the true variance is

$$\frac{\Sigma(X_i - \bar{X})^2}{n-1} = 60.6667$$

(c) The estimate (in kg) of the true standard deviation is $\sqrt{60.6667} = 7.7889$ kg

Example E5.2

A simple random sample of 44 engineers in New York city yielded the following data on their monthly income from employment (in thousands of dollars):

8	14	8	11	7	4	5	7	5
12	6	7	7	4	8	12	7	9
12	6	9	6	5	7	3	10	10
12	8	8	10	5	6	3	8	11
4	5	7	6	11	7	6	8	

(a) use the data to obtain a point estimate for the mean monthly income of all engineers working in New York city

(b) is your point estimate in part *(a)* likely to equal μ exactly?

SOLUTION:

(a) $\Sigma X_i = 334$, and $n = 44$

Hence $\bar{X} = 334/44 = 7.6$

(b) It is not likely that $\bar{X}$ is exactly equal to μ (population mean). Some sampling error is to be expected.

5.3 CONFIDENCE INTERVAL ON MEAN, VARIANCE KNOWN

This section presents methods for using sample data to find a point estimate and confidence interval estimate of a population mean. The key requirement in this section is that in addition to having sample data, we also know σ, the standard deviation of the population.

For obtaining a confidence interval for μ, in this section, we make the following assumptions:

1. The sample consists of n independent observations, that is, any one observation does not influence any other.
2. The underlying population is normally distributed with mean μ, which is, of course, unknown ($n > 30$).
3. The population standard deviation σ is known.

Let X be a random variable with unknown mean μ and known variance σ^2, and suppose that a random sample of size n, $X_1, X_2, \ldots, X_n$ is taken. A $100(1-\alpha)\%$ confidence interval on μ can be obtained by considering the sampling distribution of the sample mean $\bar{X}$. The mean of $\bar{X}$ is μ and the variance σ^2/n. Therefore, the distribution of the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is taken to be a standard normal distribution.

The distribution of $Z = \left(\bar{X} - \mu\right)\Big/\left(\sigma/\sqrt{n}\right)$ is shown in Fig. 5.1. From this figure we note that

$$P\{-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\} = 1 - \alpha$$

or
$$P\left\{-Z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right\} = 1 - \alpha$$

This can be arranged as

$$P\left\{\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right\} = 1 - \alpha \tag{5.4}$$

From Eqs. (5.1) and (5.4), we note that the $100(1-\alpha)\%$ two-sided confidence interval on μ is

$$\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n} \tag{5.5}$$

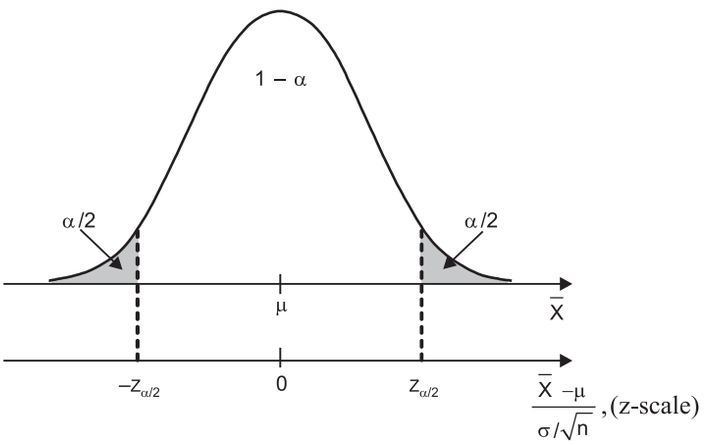


Fig. 5.1: The distribution of z

From Eq. (5.4), we observe that $\bar{X}$ will differ from μ by at most $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ with probability $(1-\alpha)$. The quantity $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ is, therefore, called the *maximum error of estimate* of μ at the $(1-\alpha)100$ per cent level. The maximum error of estimate of μ is commonly called the *margin of error* or the *sampling error*.

Hence the margin of error of estimate of μ at the $(1 - \alpha)100$ per cent level is the maximum error in estimating μ and is given by

$$\text{margin of error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The interval $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is a random interval and the probability that it covers the mean μ is $(1 - \alpha)$. The end points

$$L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and

$$U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

of the interval are random variables. Their values depend on the value of $\bar{X}$ which, in turn, depends on the sample values.

Therefore, a confidence interval for μ is often given simply by specifying the limits as

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the sample mean based on n observations.

The above construction of the confidence interval for μ was based on the assumption that the population is normally distributed. This assumption was necessary since it permitted us to proceed by stating that $\bar{X}\left(\text{and } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)$ has a normal distribution. If n is large (at least 30), the assumption of a normal distribution is crucial, because the central limit theorem allows us to proceed by stating that $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ has approximately a normal distribution (we still get the same confidence limits given above, but now they are approximate limits). When we say that the confidence interval is *exact*, we mean that the true confidence level equals $1 - \alpha$; similarly, when we say that the confidence interval is *approximately correct*, we mean that the true confidence level only approximately equals $1 - \alpha$.

The procedure to find a confidence interval for a population mean when the standard deviation is known is shown in Table 5.1.

Example E5.3

Refer to Example E5.1. Construct a 90% confidence interval for μ, the true mean weight of the students.

SOLUTION:

From Example E5.1, we have

$$n = 6,\ \bar{X} = 58.6667 \text{ kg and } \sigma = 7.7889 \text{ kg}.$$

For a 90% confidence interval, $\alpha = 0.10$. Thus, $z_{\alpha/2} = z_{0.05} = 1.645$ (from the table in Appendix-*E*). The confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$58.6667 - 1.645\frac{7.7889}{\sqrt{6}} \le \mu \le 58.6667 + 1.645\frac{7.7889}{\sqrt{6}}$$

or

$$58.6667 - 5.2308 \le \mu \le 58.6667 + 5.2308$$

$$53.4359 \le \mu \le 63.8975$$

which can be simplified to $53.44 < \mu < 63.90$, giving the confidence interval (53.44, 63.90). Hence, with 90% confidence we estimate that the mean weight μ of the students in that college is between 53.44 and 63.90 kg.

The term *normal population* is used here as an abbreviation for "the variable under consideration is normally distributed". The *z*-interval procedure works reasonably well even when the variable is not normally distributed and the sample size is small, or moderate, provided the variable is not too far from being normally distributed. For large samples, say, of size 30 or more ($n \ge 30$), the *z*-procedure can be used essentially without restriction. For samples of moderate size ($15 \le n \le 30$), the *z*-interval procedure can be used unless the data contains outliers or the variable under consideration is far from being normally distributed. For small samples, say, of size less than 15 ($n \le 15$), the *z*-interval procedure should be used only when the variable under consideration is normally distributed or very close to being so. If outliers are present then their removal is justified in a data set in order to use the *z*-interval procedure.

Table 5.1: Procedure to find a confidence interval for a population mean when σ is known

Assumptions:
1. Simple random sample
2. Normal population or large sample
3. σ known

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*E* to find $z_{\alpha/2}$.

Step 2: The confidence interval for μ is from

$$\bar{X} - z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is found in step 1, n is the sample size and $\bar{X}$ is computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.4

Refer to Example E5.2. Assume that the recent monthly income of engineers from employment in New York city are normally distributed with a standard deviation of $8100.

(a) determine a 95.44% confidence interval for the mean cost, μ, of all recent engineers in New York city

(b) interpret your result in part *(a)*

(c) does the mean monthly income from employment of all engineers in New York city lie in the confidence interval obtained in part *(a)*?

Explain your answer.

SOLUTION:

(a) The confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $x_{\alpha/2} = 2.0$ (from the table in Appendix-*E*).

Therefore $7.6 - 2\frac{(2.4)}{\sqrt{44}}$ to $7.6 + 2\frac{(2.4)}{\sqrt{44}}$

or 6.9 to 8.3

(b) Since we know that 95.44% of all samples of 44 engineers employment monthly income have the property that the interval from $\bar{X} - 0.7$ to $\bar{X} + 0.7$ contains μ, we can be 95.44% confident that the interval from 6.9 to 8.3 contains μ.

(c) The confidence interval in part *(a)* would be exact if the population of engineers' monthly employment income were exactly a normal distribution. However, since engineers' monthly employment income is a discrete random variable and the normal distribution is continuous, monthly employment income cannot follow a normal distribution exactly.

Example E5.5

It has been established that in a certain packaging process, the packages are of an average weight of μ but that μ changes over time as the process adjustment changes through wear. It was also found that the variance of the package weight is a constant at 16 kg even though the mean changes overtime. A sample of 36 packages has revealed their mean weight $\bar{X}$ of 100 kg. Assuming that the weight of the individual packages are normally distributed around μ, *(a)* find the 90% confidence limits for μ, *(b)* find the 95% confidence limits for μ.

SOLUTION:

(a) Here $1 - \alpha = 0.90$ or $\alpha = 0.10$ and $\alpha/2 = 0.05$.

$z_{0.05} = 1.645$ (from the table in Appendix-*E*), $\alpha = \sqrt{16} = 4$ kg

By computation,

$$L = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 - \frac{1.645(4)}{\sqrt{36}} = 100 - 1.0967 = 98.9033$$

$$U = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 + \frac{1.645(4)}{\sqrt{36}} = 100 + 1.0967 = 101.0967$$

Hence, we are 90% confident that m lies between 98.9033 and 101.0967 kg.

(b) Here $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*)

$$L = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 - \frac{1.96(4)}{\sqrt{36}} = 100 - 1.3067 = 98.6933$$

$$U = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 + \frac{1.96(4)}{\sqrt{36}} = 100 + 1.3067 = 101.3067$$

Therefore, we are 95% confident that μ lies between 98.6933 and 101.3067 kg.

One-sided Confidence Intervals

One-sided confidence intervals for μ are obtained by setting either $L = -\infty$ or $U = \infty$ and replacing $z_{\alpha/2}$ by z_{α}. The $100(1 - \alpha)\%$ upper-confidence interval for μ is

$$\mu \le \bar{X} + z_{\alpha}\sigma/\sqrt{n} \tag{5.6}$$

and the $100(1 - \alpha)\%$ lower-confidence interval for μ is

$$\bar{X} - z_{\alpha}\sigma/\sqrt{n} \le \mu \tag{5.7}$$

It is most preferable to obtain and pinpoint the parameter with 100% certainty. Due to the variability inherent in the population, it is hard to accomplish this. One can try to achieve the turn goal of a high level of confidence and a narrow interval.

Length of the confidence interval for μ: The length of any interval from *a* to *b* is $(b - a)$. Hence, the length of the confidence interval is

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Therefore, the length of the confidence interval for μ

$$= 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2 \text{ (margin of error)}$$

The length of the confidence interval is also called its *width*.

Hence, the length of the confidence interval is twice the margin of error or the margin of error is one-half the length of the confidence interval.

We also note that the length of the confidence interval for μ does not depend on $\bar{X}$, but it depends on $z_{\alpha/2}$, σ, and *n*. Hence, in order to achieve a high accuracy in estimating μ (that is, a narrow confidence interval with high degree of confidence) then one way to accomplish this goal is to pick an appropriately large sample. The margin of error is the standard error of the mean multiplied by $z_{\alpha/2}$. The length of a confidence interval, and thus the precision with which $\bar{X}$ estimates μ, is determined by the margin of error. Increasing the confidence level while keeping the sample size the same will increase the value of $z_{\alpha/2}$ and hence the

length of the confidence interval (decrease the precision of the estimate). Increasing the sample size while keeping the same confidence level will decrease the margin of error and the length of the confidence interval (increase the precision).

Example E5.6

A random sample of 18 venture-capital investments in the fiber optics sector yields the following data, in crores of rupees:

2.04	5.48	5.60	5.96	6.27	10.51
4.13	5.58	5.74	5.95	6.67	8.63
4.21	4.98	6.66	7.71	8.64	9.21

(a) determine a 95% confidence interval for the mean amount, μ, of all venture, capital investments in the fiber optics sector. Assume that the population standard deviation is 2.04 crores of rupees

(b) interpret your finding in part *(a)*

(c) find a 99% confidence interval for μ

(d) why is the confidence interval found in part (*c*) longer than the one in part *(a)*?

(e) which confidence interval yields a more precise estimate of μ? Explain your answer.

SOLUTION:

(a) $n = 18$ and $\Sigma X = 113.97$ crores

$$\bar{X} = \frac{\Sigma X}{n} = \frac{113.97}{18} = 6.33 \text{ crores}$$

The 95% confidence interval for μ is

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{\omega/2} = 1.96$ (from the table in Appendix-*E*),

$$6.33 - 1.96\frac{(2.04)}{\sqrt{18}} \text{ to } 6.33 + 1.96\frac{(2.04)}{\sqrt{18}}$$

or 5.39 to 7.27 crores.

(b) We can be 95% confident that the interval from 5.39 crores to 7.27 crores contains the population mean venture capital investment in the fiber-optics sector.

(c) $n = 18$, $\bar{X} = 6.33$ and $\sigma = 2.04$

$\alpha = 0.01$ and $z_{\omega/2} = z_{0.005} = 2.575$ (from the table in Appendix-*E*),

Hence $$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$6.33 - 2.575\frac{(2.04)}{\sqrt{18}} \text{ to } 6.33 + 2.575\frac{(2.04)}{\sqrt{18}}$$

or 5.09 to 7.57 crores.

(d) The confidence interval in *(c)* is longer than the one part *(a)* because we have changed the confidence level from 95% in part *(a)* to 99% in part *(c)*. We notice that increasing the confidence level from 95% to 99% increases the $z_{\alpha/2}$ value from 1.96 to 2.575. The larger z-value, in turn, results in a longer interval. In order to accomplish a higher level of confidence that the interval contains the population mean, we need a longer interval.

(e) See Fig. E5.6.

The 95% confidence interval is shorter and therefore provides a more precise estimate of μ.

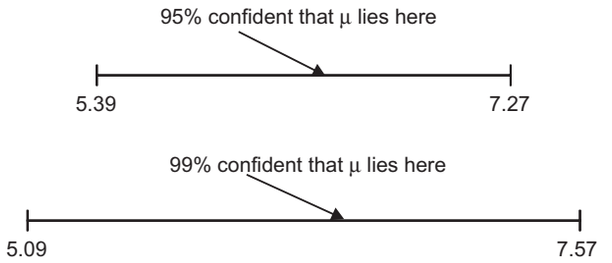


Fig. E5.6

5.4 CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE UNKNOWN

The method presented here is valid for any arbitrating sample size but is particularly important when the sample size is small. For setting a confidence interval for μ, we make the following assumptions:

1. The sample is a simple random sample.
2. The observations are picked from the population under study and are independent.
3. The population has a normal distribution.

Here we find a confidence interval on the mean of a distribution when the variance is unknown. Suppose a random sample of size n, $X_1, X_2, \ldots, X_n$ is available and $\bar{x}$ and S^2 are the sample mean and sample variance, respectively. In order to use a valid confidence interval when the sample size small, we make an assumption that the underlying population is normally distributed. This leads to confidence intervals based on the t-distribution. Therefore, let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. We know the sampling distribution of the statistic

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is the t-distribution with $n - 1$ degrees of freedom.

The distribution of $t = \left(\bar{X} - \mu\right) \Big/ \left(S/\sqrt{n}\right)$ is shown in Fig. 5.2. Letting $t_{\alpha/2, n-1}$ be the upper $\alpha/2$ percentage point of the t-distribution with $n - 1$ degree of freedom, we note from Fig. 5.2, that

$$P\left\{-t_{\alpha/2, n-1} \le t \le t_{\alpha/2, n-1}\right\} = 1 - \alpha$$

or
$$P\left\{-t_{\alpha/2,\,n-1} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2,\,n-1}\right\} = 1-\alpha \tag{5.8}$$

Rearranging Eq. (5.8)

$$P\left\{\bar{X} - t_{\alpha/2,\,n-1}S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\,n-1}S/\sqrt{n}\right\} = 1-\alpha \tag{5.9}$$

Comparing Eqs. (5.4) and (5.9), we see that a 100 $(1-\alpha)$% two-sided confidence interval on μ is

$$\bar{X} - t_{\alpha/2,\,n-1}S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\,n-1}S/\sqrt{n} \tag{5.10}$$

A 100 $(1-\alpha)$ % lower-confidence interval on μ is given by

$$\bar{X} - t_{\alpha/2,\,n-1}S/\sqrt{n} \le \mu \tag{5.11}$$

and a 100 $(1-\alpha)$ % upper-confidence interval on μ is given by

$$\mu \le \bar{X} + t_{\alpha/2,\,n-1}S/\sqrt{n} \tag{5.12}$$

The above method assumes that the samplings are from a normal population.

The procedure to find a confidence interval for a population mean when the standard deviation σ is unknown, is given in Table 5.2. Properties and guidelines for use of the *t*-interval procedure are the same as those for the *z*-interval procedure. The *t*-interval procedure is robust to moderate violations of the normality assumptions.

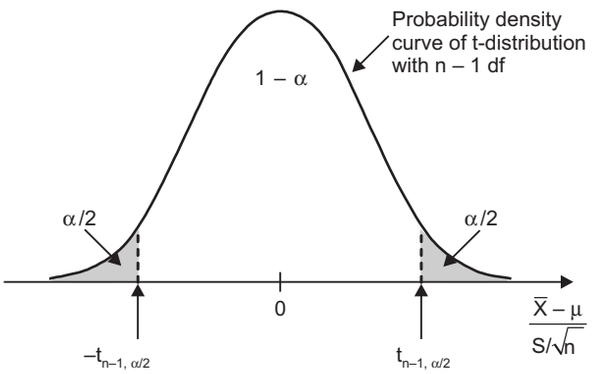


Fig. 5.2: t-values such that there is an area α/2 in the right tail and an area α/2 in the left tail of the distribution

Table 5.2: Procedure to find a confidence interval for a population mean when σ is unknown

Assumptions:
1. Simple random sample
2. Normal population or large sample
3. σ unknown

Step 1: For a confidence level of $1 - \alpha$, use the Table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n - 1$, where n is the sample size

Step 2: The confidence interval for μ is from

$$\bar{X} - t_{\alpha/2} \bullet \frac{S}{\sqrt{n}} \text{ to } \bar{X} + t_{\alpha/2} \bullet \frac{S}{\sqrt{n}}$$

where $t_{\alpha/2}$ is found in step 1 and $\bar{X}$ and S are computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.7

Find the following confidence intervals for μ_d assuming that the populations of paired differences are normally distributed.

(a) $n = 9$, $\bar{X} = 25$, $S = 13$, confidence level 99%.

(b) $n = 26$, $\bar{X} = 13$, $S = 5$, confidence level 95%.

(c) $n = 12$, $\bar{X} = 35$, $S = 12$, confidence level 90%.

SOLUTION:

(a) The *t*-value for $\nu = n - 1 = 9 - 1 = 8$ d.o.f. and 0.005 area in the right tail is 3.355 (from the table in Appendix-*G*). The 99% confidence interval with $t_{0.005,8} = 3.355$ is

$$\bar{X} - t_{\alpha/2,\nu} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\nu} S/\sqrt{n}$$

$$25 - 3.355(13)/\sqrt{9} \le \mu \le 25 + 3.355(13)\sqrt{9}$$

$$25 - 14.538 \le \mu \le 25 + 14.538$$

$$10.462 \le \mu \le 39.538$$

The complete statement of the confidence interval, with the associated probability is

$$P(10.462 \le \mu \le 39.538) = 0.99$$

(b) Using the data given, a 95% confidence interval with $t_{0.025,25} = 2.060$ from the table in Appendix-*G*, is given

$$13 - 2.060(5)/\sqrt{26} \le \mu \le 13 + 2.060(5)/\sqrt{26}$$

$$13 - 2.02 \le \mu \le 13 + 2.02$$

$$10.98 \le \mu \le 15.02$$

The complete statement of the confidence interval with the associated probability is

$P(10.98 \le \mu \le 15.02) = 0.95$

(c) Using the data given, a 90% confidence interval with $t_{0.05,11} = 1.796$ from the table in Appendix-*G*, is given by

$$35 - 1.796(12)/\sqrt{12} \le \mu \le 35 + 1.796(12)/\sqrt{12}$$

$$35 - 0.16 \le \mu \le 35 + 0.16$$

$$34.84 \le \mu \le 35.16$$

The complete statement of the confidence interval with the associated probability is

$P(34.84 \le \mu \le 35.16) = 0.90$

Example E5.8

Measurements on the percentage of enrichment of 12 fuel rods were reported as follows:

2.94	3	2.9	2.75	3	2.95
2.75	2.95	2.92	2.81	3.05	2.8

Find a 99% two-sided confidence interval on the mean percentage of enrichment of fuel rods. Can we state the mean percentage of enrichment is 2.95%? Why?

SOLUTION:

Here $n = 12$ and $\Sigma X = 34.82$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{34.82}{12} = 2.9017$$

Also $\Sigma(X_i - \bar{X})^2 = 0.112567$

Hence $$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{0.112567}{12-1}} = 0.10116$$

The 99% two-sided confidence interval on the mean per cent enrichment is found from

$$\bar{X} - t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right) \le \mu \le \bar{X} + t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right)$$

For $\alpha = 0.01$ and $n = 12$, $t_{\alpha/2,\, n-1} = t_{0.005,11} = 3.106$ (from the table in Appendix-*G*).

Therefore $$2.9017 - 3.106\left(\frac{0.10116}{\sqrt{12}}\right) \le \mu \le 2.9017 + 3.106\left(\frac{0.10116}{\sqrt{12}}\right)$$

$$2.9017 - 0.896625$$

or $$2.005 \le \mu \le 3.798325$$

We can state that the mean percentage of enrichment 2.95 is included in the 99% two-sided confidence interval.

5.5 CONFIDENCE INTERVAL ON THE VARIANCE OF A NORMAL DISTRIBUTION

Here, we accept S^2 as providing a good estimate for σ^2. For setting a confidence interval for σ^2, we make the following assumptions:

1. The observations are independent.
2. The parent population has a normal distribution.
3. The sample is a simple random sample.

Suppose that X is normally distributed with unknown mean μ and unknown variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n, let S^2 be the sample variance. We note that the sampling distribution of

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is chi-square with $n - 1$ degrees of freedom. This distribution is shown in Fig. 5.3.

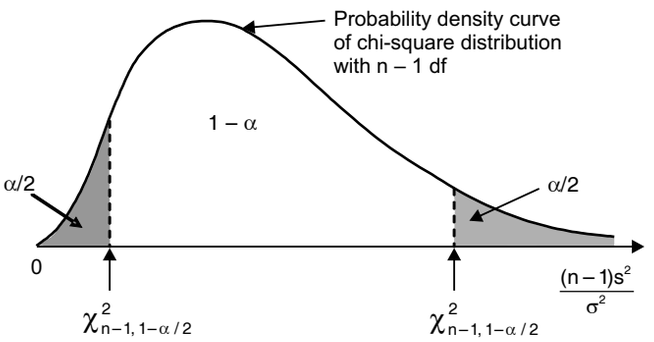


Fig. 5.3: Chi-square values such that areas 1 – α/2 and α/2 are to their right

We observe from Fig. 5.3 that,

$$P\left\{\chi^2_{1-\alpha/2,n-1} \le \chi^2 \le \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha$$

or

$$P\left\{\chi^2_{1-\alpha/2,n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha \tag{5.13}$$

Equation (5.13) can be rearranged to give

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha \tag{5.14}$$

Comparing Eqs. (5.14) and (5.4), we see that a 100 $(1 - \alpha)$ % two-sided confidence interval on σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{5.15}$$

when S^2 is the sample variance based on n observations from a normal population.

In order to find a 100 $(1-\alpha)$% lower-confidence interval on σ^2, set $U = \infty$ and replace $\chi^2_{\alpha/2,n-1}$ with $\chi^2_{\alpha,n-1}$, giving

$$\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2 \tag{5.16}$$

The 100 $(1-\alpha)$% upper-confidence interval is found by setting $L = 0$ and replacing $\chi^2_{1-\alpha/2,n-1}$ with $\chi^2_{1-\alpha,n-1}$, resulting in

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} \tag{5.17}$$

In all the above Eqs. (5.13 to 5.17), $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$ represent values from the chi-square distribution with $n-1$ degrees of freedom such that they leave, respectively, area of $1-\alpha/2$ and $\alpha/2$ to their right, as shown in Fig. 5.3. The procedure to find a confidence interval for a population standard deviation is given in Table 5.3.

Table 5.3: Procedure to find a confidence interval for a population standard deviation, σ

Assumptions:	
1.	Simple random sample
2.	Normal population
Step 1:	For a confidence level of $1-\alpha$, use the table in Appendix-*F* to find $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ with $df = n-1$
Step 2:	The confidence interval for σ is from $\sqrt{\frac{(n-1)}{\chi^2_{\alpha/2}}S}$ to $\sqrt{\frac{(n-1)}{\chi^2_{1-\alpha/2}}S}$ where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are found in step 1, n is the sample size and S is computed from the sample data obtained
Step 3:	Interpret the confidence interval.

Example E5.9

Data are collected on the driving time required to reach place *A* from place *B*. The driving time is assumed to follow a normal distribution. Twenty-one driving times are collected and the sample variance is calculated to be 2 hours. A 99% confidence interval on the true variance is desired.

SOLUTION:

The two-sided confidence interval is given by (Appendix-*F*)

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\nu}}$$

where $\nu = n-1$.

From the table in Appendix-*F* with $\nu = 21 - 1 = 20$ and $\alpha/2 = 0.005$, we find that $\chi^2_{0.005,20} = 39.997 = \chi^2_{0.995,20} = 7.434$.

Therefore,

$$\frac{(21-1)(2^2)}{39.997} \le \sigma^2 \le \frac{(21-1)(2^2)}{7.434} \text{ or } 2 \le \sigma^2 \le 10.761$$

The complete statement of the confidence interval with the associated probability is

$$P(2 \text{ hrs} \le \sigma^2 \le 10.761 \text{ hrs}) = 0.99$$

Example E5.10

The capability of a specific gauge can be studied by measuring the weight of paper. The data for repeated measurements of one sheet of paper are given in Table E5.10. Construct a 95% one-sided, upper confidence interval for the standard deviation of these measurements.

Table E5.10

3.481	3.449	3.484	3.476	3.473
3.478	3.473	3.465	3.475	3.471
3.472	3.471	3.478	3.473	3.475

SOLUTION:

A 95% upper-confidence interval is found from Eq. (5.17) as follows:

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{0.95,14}}$$

$n = 15$, $S = 0.001354$ and $\chi^2_{0.95,14} = 6.57$ (from the table in Appendix-*F*).

$$\Sigma X_i = 52.094\,,\ \Sigma(X_i - \bar{X})^2 = 0.000901$$

$$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = 0.001354 \text{ and } S^2 = 1.83 \times 10^{-6}$$

$$\sigma^2 = \frac{14(1.83 \times 10^{-6})}{6.57}$$

$$\sigma^2 \le 3.9073 \times 10^{-6}$$

The above can be converted into a confidence interval on the standard deviation σ by taking the square root of both sides, resulting in $\sigma = 0.00197667$.

Therefore, at the 95% level of confidence, the data indicates that the process standard deviation could be as large as 0.00197667.

5.6 CONFIDENCE INTERVAL ON A POPULATION PROPORTION

We will now give an interval estimate for population proportion under the following assumptions:

1. The sample consists of n independent observations.
2. The sample size is large.
3. The population proportion is not too close to 0 or 1.

If a random sample size n has been taken from a large (possibly infinite) population, and $X(\leq n)$ observations in this sample belong to a class of interest, then $\hat{p} = X/n$ is the point estimator of the proportion of the population belonging. n and p are the parameters of binomial distribution. The sampling distribution $\hat{p}$ is approximately normal mean p and variance $p(1-p)/n$, if p is not too close to either 0 or 1, and if n is relatively large. Therefore, the distribution of

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal distribution.

We also observe that

$$P\{-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\} = 1 - \alpha$$

or

$$P\left\{-Z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq Z_{\alpha/2}\right\} = 1-\alpha \tag{5.18}$$

Rearranging Eq. (5.18)

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right\} = 1-\alpha \tag{5.19}$$

The quantity of $\sqrt{p(1-p)/n}$ is called the *standard error of the point estimator* $\hat{p}$. Replacing p by $\hat{p}$ in the standard error, giving an estimated standard error.

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} = 1-\alpha \tag{5.20}$$

The above formula, Eq. (5.20), is applicable only if the observed sample proportion is not too close to 0 or 1 and the observed number of successes X and the observed number of failures $(n - X)$ both exceed 5. If these conditions are not met, the procedure is not recommended.

The above procedure depends on the adequacy of the normal approximation to the binomial. Conservatively speaking, this requires that np and $n(1 - p)$ be greater than or equal to 5. In situations where the approximation is inappropriate (especially when n is small) other methods must be used. Tables of binomial distribution could be used to obtain a confidence interval for p.

We may find approximate one-sided confidence bounds on p by a simple modification of Eq. (5.20).

The approximate $100(1 - \alpha)\%$ lower and upper confidence bounds are

$$\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \quad \text{and} \quad p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

respectively.

The procedure to find a confidence interval for a population proportion, p is given in Table 5.4.

Table 5.4: Procedure to find a confidence interval for a population standard proportion, p

Assumptions:
1. Simple random sample
2. The number of successes, x, and the number of failures, $n - x$ are both 5 or greater

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*E* to find $z_{\alpha/2}$

Step 2: The confidence interval for p is from

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ to } \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is found in step 1, n is the sample size and $\hat{p} = x/n$ is the sample proportion

Step 3: Interpret the confidence interval.

Example E5.11

In a random sample of 300 automobile crankshaft bearings, 12 have a surface finish that is rougher than the specifications allow.

(a) Calculate a 95% two-sided confidence interval on the fraction of defective crankshafts produced by this particular process.

(b) Calculate a 95% upper confidence bound on the fraction of defective crankshafts.

SOLUTION:

(a) 95% confidence interval on the fraction defectives produced with the process.

$\hat{p} = \dfrac{12}{300} = 0.04$, $n = 30$, $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*).

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.04 - 1.96\sqrt{\frac{0.04(1-0.04)}{300}} \le p \le 0.04 + 1.96\sqrt{\frac{0.04(1-0.04)}{300}}$$

$0.01 - 0.022175 \le p \le 0.04 + 0.02275$

$0.017825 \le p \le 0.062175$

(b) 95% upper confidence bound:

$z_{\alpha/2} = z_{0.05} = 1.65$ (from Appendix-*E*)

$$p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$p \le 0.04 + 1.65\sqrt{\frac{0.04(1-0.04)}{300}}$$

or $\le 0.04 + 0.018668$

Hence $p \le 0.058668$

Example E5.12

A sample of 300 of electronic components produced showed 36 non-conforming ones. Construct a 95% confidence interval for this. Assume binomial distribution.

SOLUTION:

A two-sided confidence interval on p is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Since n is large and the underlying distribution is binomial, $\hat{p} = 36/300 = 0.12$

A 95% confidence interval, with $Z_{0.025} = 1.96$ (from the table Appendix-*E*) is given by

$$0.12 - 1.96\sqrt{\frac{(0.12)(0.88)}{300}} \le p \le 0.12 + 1.96\sqrt{\frac{(0.12)(0.88)}{300}}$$

or
$$0.12 - 1.96\,(0.01876) \le p \le 0.12 + 1.96\,(0.01876)$$
$$0.12 - 0.0368 \le p \le 0.12 + 0.0368$$
$$0.0832 \le p \le 0.1568$$

The complete statement of the confidence interval with the associated probability is

$$P(0.0832 \le p \le 0.1568) = 0.95$$

Margin of Error:

The margin of error from Eq. (5.20) is given by

$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

It can be shown that the maximum value of $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is $z_{\alpha/2}\sqrt{\frac{1}{4n}}$ and occurs when $\hat{p} = \frac{1}{2}$. This called the *safe* or *conservative margin of error*, whatever the actual value of $\hat{p}$ in the population.

Hence, conservative margin of error = $z_{\alpha/2}\sqrt{\frac{1}{4n}}$.

Example E5.13

Determine the margin of error in Example E5.12.

SOLUTION:

The margin of error is given by

$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*), $\hat{p} = 36/300 = 0.12$

Hence, margin of error = $1.96\sqrt{\dfrac{0.12(1-0.12)}{300}} = 0.0368$

5.7 CONFIDENCE INTERVAL ON THE DIFFERENCE IN TWO MEANS, VARIANCE KNOWN

The procedure presented in this section for finding a confidence interval is valid if the following assumptions are justified:

1. The two populations are normally distributed. This assumption is not very important if both sample sizes are at least 30 (large sample size).
2. The standard deviations of the two populations are known.
3. Two random samples are picked, one from each population. They are independent, that is, any outcome in one sample does not influence any outcome in the other sample.
4. Within each sample the outcomes are independent.

Consider two independent random variables X_1 with unknown mean μ_1 and known variance σ_1^2 and X_2 with unknown mean μ_2 and variance σ_2^2. We wish to find a 100 $(1 - \alpha)$% confidence interval on the difference in means μ_1, μ_2. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample of n_1 observations from X_1; and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample of n_2 observations from X_2. If $\bar{X}_1$ and $\bar{X}_2$ are the sample means, the statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is standard normal if X_1 and X_2 are normal or approximately standard normal if the conditions of the central limit theorem apply, respectively. Also, σ_1 and σ_2 are the population standard deviations. From Fig. 5.2 it is clear that

$$P\{-z_{\alpha/2} \le z \le z_{\alpha/2}\} = 1 - \alpha$$

or

$$P\left\{-z_{\alpha/2} \le \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right\} = 1 - \alpha$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Rearranging, we have

$$P\left\{\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} = 1 - \alpha \tag{5.21}$$

Comparing Eqs. (5.4) and (5.21), we note that 100$(1 - \alpha)$% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.22}$$

One-sided confidence intervals on $\mu_1 - \mu_2$ may also be obtained. A 100 $(1 - \alpha)$% upper-confidence interval on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.23}$$

and a $100(1 - \alpha)$ % lower-confidence interval is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \tag{5.24}$$

The confidence level $(1 - \alpha)$ is exact when the populations are normal. For non-normal populations, the confidence level is approximately valid for large sample sizes.

The margin of error E is

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.24a}$$

Example E5.14

The mean hourly wage for male workers was \$15 and \$13 for female workers in a particular manufacturing sector in the year 2000. These two estimations were based on random samples of 1000 and 1200 workers taken, respectively from two independent populations. The standard deviations of the two populations are known to be \$2 and \$1.50 respectively. Construct a 95% confidence interval for the difference between the mean hourly wages of the two populations.

SOLUTION:

For the 95% confidence level $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*).

The 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$15 - 13 \pm 1.96\sqrt{\frac{2^2}{1000} + \frac{1.5^2}{1200}} = 2 \pm 1.96\,(0.2046) = 2 \pm 0.4011 = \$1.60 \text{ to } \$2.40$$

We have obtained (\$1.60, \$2.40) as an 95% confidence interval for $\mu_1 - \mu_2$.

Example E5.15

Two different formulations of oxygenated motor fuel are being tested to study their road octane numbers. The variances of road octane number for formulation *A* is $\sigma_1^2 = 1.5$ and for formulation *B* it is $\sigma_2^2 = 1.2$. Two random samples of sizes $n_1 = 15$ and $n_2 = 20$ are tested, and the mean road octane numbers observed are

$\bar{X}_1$ = 90 and $\bar{X}_2$ = 93. Assume normality and calculate a 95% confidence interval on the difference in means.

SOLUTION:

95% confidence interval: $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$ (from the table in Appendix-*E*).

$$\left(\bar{X}_1 - \bar{X}_2\right) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$90 - 93 - 1.96\sqrt{\frac{1.5}{15} + \frac{1.2}{20}} \le \mu_1 - \mu_2 \le 90 - 93 + 1.96\sqrt{\frac{1.5}{15} + \frac{1.2}{20}}$$

$$-3 - 0.784 \le \mu_1 - \mu_2 \le -3 + 0.784$$

or $$-3.784 \le \mu_1 - \mu_2 \le -2.216$$

With 95% confidence, we believe the mean road octane number for formulation *B* exceeds that of formulation *A* since 0 is not included in the confidence interval.

5.8 CONFIDENCE INTERVAL ON THE DIFFERENCE IN MEANS OF TWO NORMAL DISTRIBUTIONS, VARIANCES UNKNOWN

We make the following assumptions:

1. The two populations are normally distributed.
2. The population variances σ_1^2 and σ_2^2, though unknown are the same, say, each equal to σ^2.
3. Two random and independent samples are picked, one from each population.

Consider two independent normal random variables, say X_1 with mean μ_1 and variance σ_1^2, and X_2 with mean μ_2 and variance σ_2^2. Both the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 are unknown. Assume that both variances are equal; that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to find a 100 $(1 - \alpha)$% confidence interval on the difference in means $\mu_1 - \mu_2$.

Random samples of size n_1 and n_2 are taken on X_1 and X_2, respectively; let the sample means be denoted by $\bar{X}_1$ and $\bar{X}_2$, and the sample variances be denoted S_1^2 and S_2^2. Since both S_1^2 and S_2^2 are estimates of the common variance σ^2, we may obtain a combined (or "pooled") estimator of σ^2 as the pooled estimate of the common population standard deviation, S_p, is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{5.25}$$

To find the confidence interval for $\mu_1 - \mu_2$, note that the distribution of the statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Hence,

$$P\left\{-t_{\alpha/2,n_1+n_2-2} \le t \le t_{\alpha/2,n_1+n_2-2}\right\} = 1-\alpha$$

or
$$P\left\{-t_{\alpha/2,n_1+n_2-2} \le \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le t_{\alpha/2,n_1+n_2-2}\right\} = 1-\alpha$$

Here $t_{\alpha/2,n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Rearranging,

$$P\left\{\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right\} = 1-\alpha \tag{5.26}$$

Hence, a 100 $(1 - \alpha)$% two-sided confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{5.27}$$

A one-sided 100 $(1 - \alpha)$% lower-confidence interval on $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \tag{5.28}$$

and a one-sided 100 $(1 - \alpha)$% upper-confidence interval on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{5.29}$$

The margin of error E is given by

$$E = z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Since σ is not known, we replace it with its pooled estimate S_p and thereby we cannot use z-values from the normal table. Here, we use values from the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. This is especially important if the degrees of freedom are less than 30 because then the t-values will be markedly different from the z-values. Hence, the marginal of error is

$$E = t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where
$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{\Sigma(X_{1i}-\bar{X}_1)^2 + \Sigma(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, based, respectively on n_1 and n_2 observations, and S_p is the pooled estimate of σ. The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.5.

Table 5.5: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:
1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n_1 + n_2 - 2$

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.16

The overall distance travelled by a golf ball is tested by striking the ball with a specially made mechanical golfer. Ten randomly selected balls of two different brands *A* and *B* are tested and the overall distance measured as shown below:

Brand *A*	263	267	271	273	275	276	279	283	286	287
Brand *B*	244	258	261	263	265	268	270	271	273	280

Construct a 95% two-sided confidence interval on the mean difference in overall distance between the two brands of golf balls.

SOLUTION:

Here $n_1 = 10, n_2 = 10, \Sigma X_{1i} = 2760, \Sigma X_{2i} = 2653$

$$\bar{X}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{2760}{10} = 276, \quad \bar{X}_2 = \frac{\Sigma X_{2i}}{n_2} = \frac{2653}{10} = 265.3$$

$$\Sigma(X_{1i} - \bar{X}_1)^2 = 564 \text{ and } \Sigma(X_{2i} - \bar{X}_2)^2 = 868.1$$

$$S_1^2 = \frac{\Sigma(X_{1i} - \bar{X}_1)^2}{n_1 - 1} = 62.6667 \text{ and } S_1 = 7.9162$$