Seventeen

Statistics, Probability Distributions and Sampling

CHAPTER OUTLINE

- Introduction
- Some Important Terms and Concepts of Statistics
- Correlation
- Regression
- Probability
- Random Variables
- Discrete Probability Distribution
- Discrete Distribution Function
- Measures of Central Tendency for Discrete Probability Distribution
- Continuous Probability Distribution
- Continuous Distribution Function
- Measures of Central Tendency for Continuous Probability Distribution
- Binomial Distribution
- Poisson Distribution
- Normal Distribution
- Test of Hypothesis
- Test of Significance for Large Samples
- Small Sample Tests
- Student's *t*-Distribution
- Chi-Square (χ^2) Test
- Snedecor's *F*-Distribution
- Fisher's *z*-Distribution

17.1 INTRODUCTION

Statistics is the science which deals with the collection, presentation, analysis, and interpretation of numerical data. Statistics should possess the following characteristics:

(i) Statistics are aggregates of facts.
(ii) Statistics are affected by a large number of causes.
(iii) Statistics are always numerically expressed.
(iv) Statistics should be enumerated or estimated.
(v) Statistics should be collected in a systematic manner.
(vi) Statistics should be collected for a pre-determined purpose.
(vii) Statistics should be placed in relation to each other.

The use of statistical methods help in presenting a complex mass of data in a simplified form so as to facilitate the process of comparison of characteristics in two or more situations. Statistics also provide important techniques for the study of relationship between two or more characteristics (or variables) in forecasting, testing of hypothesis, quality control, decision making, etc.

17.2 SOME IMPORTANT TERMS AND CONCEPTS OF STATISTICS

1. Arithmetic Mean The *arithmetic mean* of a set of observations is their sum divided by the number of observations. Let $x_1, x_2, ..., x_n$ be n observations. Then their average or arithmetic mean is given by

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

For example, the marks obtained by 10 students in Class XII in a physics examination are 25, 30, 21, 55, 40, 45, 17, 48, 35, 42. The arithmetic mean of the marks is given by

$$\bar{x} = \frac{\sum x}{n} = \frac{25+30+21+55+40+45+17+48+35+42}{10} = \frac{358}{10} = 35.8$$

If n observations consist of n distinct values denoted by $x_1, x_2, ..., x_n$ of the observed variable x occurring with frequencies $f_1, f_2, ..., f_n$ respectively then the arithmetic mean is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_ix_i}{N} = \frac{\sum fx}{N}$$

where $$N = \sum_{i=1}^{n} f_i = f_1 + f_2 + \cdots + f_n$$

(a) Arithmetic Mean of Grouped Data In case of grouped or continuous frequency distribution the arithmetic mean is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{\sum fx}{N}, \text{ where } N = \sum_{i=1}^{n} f_i$$

and x is taken as the midvalue of the corresponding class intervals.

(b) Arithmetic Mean from Assumed Mean If the values of x and (or) f are large, the calculation of mean becomes quite time-consuming and tedious. In such cases, the provisional mean 'a' is taken as that value of x (midvalue of the class interval) which corresponds to the highest frequency or which comes near the middle value of the frequency distribution. This number is called the *assumed mean*.

Let $d = x - a$

$$fd = f(x-a) = fx - af$$

$$\sum fd = \sum fx - a\sum f = \sum fx - aN$$

Dividing both the sides by n,

$$\frac{\sum fd}{N} = \frac{\sum fx}{N} - a = \bar{x} - a$$

$$\bar{x} = a + \frac{\sum fd}{N}$$

(c) Arithmetic Mean by the Step-Deviation Method When the class intervals in a grouped data are equal, calculation can be simplified by the step-deviation method. In such cases, deviation of variate x from the assumed mean a (i.e., $d = x - a$) are divided by the common factor h which is equal to the width of the class interval.

Let
$$d = \frac{x-a}{h}$$
$$\bar{x} = a + h\frac{\sum fd}{\sum f} = a + h\frac{\sum fd}{N}$$

where a is the assumed mean

$d = \frac{x-a}{h}$ is the deviation of any variate x from a

h is the width of the class interval

N is the number of observations

2. Median *Median* is the central value of the variable when the values are arranged in ascending or descending order of magnitude. It divides the distribution into two equal parts. When the observations are arranged in the order of their size, median is the value of that item which has equal number of observations on either side.

In case of ungrouped data, if the number of observations is odd then the median is the middle value after the values have been arranged in ascending or descending order of magnitude. If the number of observations is even, there are two middle terms and the median is obtained by taking the arithmetic mean of the middle terms.

For example, consider the following:

(i) The median of the values 20, 15, 25, 28, 18, 16, 30, i.e., 15, 16, 18, 20, 25, 28, 30 is 20 because $n = 7$, i.e., odd and the median is the middle value, i.e., 20.

(ii) The median of the values 8, 20, 50, 25, 15, 30, i.e., 8, 15, 20, 25, 30, 50 is the arithmetic mean of the middle terms, i.e., $\frac{20+25}{2} = 22.5$ because $n = 6$, i.e., even.

In case of discrete frequency distribution, the median is obtained by considering the cumulative frequencies. The steps for calculating the median are as follows:

(i) Arrange the values of the variables in ascending or descending order of magnitudes.

(ii) Find $\frac{N}{2}$, where $N = \sum f$.

(iii) Find the cumulative frequency just greater than $\frac{N}{2}$ and determine the corresponding value of the variable.

(iv) The corresponding value of x is the median.

Median for Continuous Frequency Distribution In case of continuous frequency distribution (less than frequency distribution), the class corresponding to the cumulative frequency just greater than $\frac{N}{2}$, is called the *median class*, and the value of the median is given by

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

where l is the lower limit of the median class
f is the frequency of the median class
h is the width of the median class
c is the cumulative frequency of the class preceding the median class
N is sum of frequencies, i.e., $N = \sum f$

In case of 'more than' or 'greater than' type of frequency distributions, the value of the median is given by

$$\text{Median} = u - \frac{h}{f}\left(\frac{N}{2} - c\right)$$

where u is the upper limit of the median class
f is the frequency of the median class
h is the width of the median class
c is the cumulative frequency of the class succeeding the median class

3. Mode *Mode* is the value which occurs most frequently in a set of observations and around which the other items of the set are heavily distributed. In other words, mode is the value of the variable which is most frequent or predominant in the series. In case of a discrete frequency distribution, mode is the value of x corresponding to the maximum frequency.

For example, consider the following:

(i) In the series 6, 5, 3, 4, 3, 7, 8, 5, 9, 5, 4, the value 5 occurs most frequently. Hence, the mode is 5.
(ii) Consider the following frequency distribution:

x	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	7	3

The value of x corresponding to the maximum frequency, viz., 25, is 4. Hence, the mode is 4.

In an asymmetrical frequency distribution, mean, median, and mode are not equal.

For an asymmetrical frequency distribution, the difference between the mean and the mode is approximately three times the difference between the mean and the median.

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$
$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

This is known as the *empirical formula for calculation of the mode.*

Mode for a Continuous Frequency Distribution In case of a continuous frequency distribution, the class in which the mode lies is called the *modal class* and the value of the mode is given by

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

where l is the lower limit of the modal class
h is the width of the modal class
f_m is the frequency of the modal class

f_1 is the frequency of the class preceding the modal class
f_2 is the frequency of the class succeeding the modal class

This method of finding mode is called the *method of interpolation*. This formula is applicable only to a unimodal frequency distribution.

4. Geometric Mean The *geometric mean* of a set of n observations is the n^{th} root of their product. If there are n observations, $x_1, x_2, .., x_n$ such that $x_i > 0$ for each i, their geometric mean GM is given by

$$GM = (x_1 \cdot x_2 \cdot \cdots \cdot x_n)^{\frac{1}{n}}$$

The n^{th} root is calculated with the help of logarithms. Taking logarithms of both the sides,

$$\log GM = \log(x_1 \cdot x_2 \cdots x_n)^{\frac{1}{n}} = \frac{1}{n}\log(x_1 \cdot x_2 \cdots x_n) = \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n) = \frac{\sum \log x}{n}$$

$$GM = \text{antilog}\left(\frac{\sum \log x}{n}\right)$$

In case of a frequency distribution consisting of n observations $x_1, x_2, .., x_n$ with respective frequencies $f_1, f_2, .., f_n$, the geometric mean is given by

$$GM = (x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n})^{\frac{1}{N}}, \text{ where } N = \sum f$$

Taking logarithms of both the sides,

$$\log GM = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n) = \frac{\sum f \log x}{N}$$

$$GM = \text{antilog}\left(\frac{\sum f \log x}{N}\right)$$

In case of a continuous or grouped frequency distribution, x is taken to be the value corresponding to the midpoints of the class intervals.

5. Harmonic Mean The *harmonic mean* of a number of observations, none of which is zero, is the reciprocal of the arithmetic mean of the reciprocals of the given values.

The harmonic mean of n observations $x_1, x_2, ..., x_n$ is given by

$$HM = \frac{1}{\frac{1}{n}\sum\left(\frac{1}{x}\right)} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

For example, the harmonic mean of 2, 4, and 5 is

$$HM = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{5}} = 3.16$$

In case of a frequency distribution consisting of n observations $x_1, x_2, ..., x_n$ with respective frequencies $f_1, f_2, ..., f_n$, the harmonic mean is given by

$$\text{HM} = \frac{f_1 + f_2 + \cdots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}} = \frac{\sum f}{\sum\left(\frac{f}{x}\right)}$$

6. Standard Deviation *Standard deviation* is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by the Greek letter σ. Let X be a random variable which takes on values, viz., $x_1, x_2, ..., x_n$. The standard deviation of these n observations is given by

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

where $\bar{x} = \dfrac{\sum x}{n}$ is the arithmetic mean of these observations.

This equation can be modified further.

$$\sigma = \sqrt{\frac{\sum (x^2 - 2x\bar{x} + \bar{x}^2)}{n}} = \sqrt{\frac{\sum x^2}{n} - 2\bar{x}\,\frac{\sum x}{n} + \bar{x}^2\,\frac{\sum 1}{n}}$$

$$= \sqrt{\frac{\sum x^2}{n} - 2\frac{\sum x}{n}\frac{\sum x}{n} + \left(\frac{\sum x}{n}\right)^2 \cdot \frac{n}{n}} \qquad \left[\because \sum 1 = n\right]$$

$$= \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \sqrt{\text{Mean of squares} - \text{Square of mean}}$$

In case of a frequency distribution consisting of n observations $x_1, x_2, ..., x_n$ with respective frequencies $f_1, f_2, ..., f_n$, the standard deviation is given by

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}}$$

This equation can also be modified.

$$\sigma = \sqrt{\frac{\sum f\,(x^2 - 2x\bar{x} + \bar{x}^2)}{N}} = \sqrt{\frac{\sum fx^2}{N} - \frac{2\bar{x}\sum fx}{N} + \bar{x}^2\,\frac{\sum f}{N}}$$

$$= \sqrt{\frac{\sum fx^2}{N} - 2\frac{\sum fx}{N}\frac{\sum fx}{N} + \left(\frac{\sum fx}{N}\right)^2} \qquad \left[\because \sum f = N \text{ and } \bar{x} = \frac{\sum fx}{N}\right]$$

$$= \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

(a) Variance The *variance* is the square of the standard deviation and is denoted by σ^2. The method for calculating variance is same as that given for the standard deviation.

(b) Standard Deviation from the Assumed Mean If the values of x and f are large, the calculation of fx, fx^2 becomes tedious. In such a case, the assumed mean a is taken to simplify the calculation.

Let a be the assumed mean.

$$d = x - a$$
$$x = a + d$$
$$\sum fx = \sum f(a+d) = Na + \sum fd$$

Dividing both the sides by N,

$$\frac{\sum fx}{N} = a + \frac{\sum fd}{N}$$
$$\bar{x} = a + \bar{d}$$
$$x - \bar{x} = d - \bar{d}$$
$$\sigma_x = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}} = \sqrt{\frac{\sum f(d-\bar{d})^2}{N}} = \sigma_d$$

Hence, the standard deviation is independent of change of origin.

$$\sigma_x = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

(c) Standard Deviation by Step-Deviation Method Let a be the assumed mean and h be the width of the class interval.

$$d = \frac{x-a}{h}$$
$$x = a + hd$$
$$\sum fx = \sum f(a+hd) = Na + h\sum fd$$

Dividing both the sides by N,

$$\frac{\sum fx}{N} = a + h\frac{\sum fd}{N}$$
$$\bar{x} = a + h\bar{d}$$
$$x - \bar{x} = h(d - \bar{d})$$
$$\sigma_x = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}} = \sqrt{\frac{\sum f\,h^2\,(d-\bar{d})^2}{N}} = h\sqrt{\frac{\sum f\,(d-\bar{d})^2}{N}} = h\,\sigma_d$$

Hence, the standard deviation is independent of change of origin but not of scale.

$$\sigma_x = h\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

(d) Coefficient of Variation The *standard deviation* is an absolute measure of dispersion. The coefficient of variation is a relative measure of dispersion and is denoted by CV.

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

where σ is the standard deviation and $\bar{x}$ is the mean of the given series. The coefficient of variation has great practical significance and is the best measure of comparing the variability of two series. The series or groups for which the coefficient of variation is greater is said to be more variable or less consistent. On the other hand, the series for which the variation is lesser is said to be less variable or more consistent.

7. Moment

Moment is the arithmetic mean of the various powers of the deviations of items from their assumed mean or actual mean. If the deviations of the items are taken from the arithmetic mean of the distribution, it is known as *central moment*. If the mean of the first power of deviations are taken, the first moment about the mean is obtained and is denoted by μ_1. The mean of the second power of the deviations gives the second moment about the mean and is denoted by μ_2. Similarly, the mean of the cubes of deviations gives third moment about the mean and is denoted by μ_3. The mean of the fourth power of the deviations from the mean gives the fourth moment about the mean and is denoted by μ_4. Thus, the mean of the r^{th} power of deviations gives the r^{th} moment about mean or r^{th} central moment and is denoted by μ_r.

(a) Central Moment or Moments about Actual Mean Let $x_1, x_2, \ldots, x_n$ be n observations with arithmetic mean $\bar{x}$. The various moments about actual mean are given by the following:

First moment about the mean $\mu_1 = \dfrac{\sum(x-\bar{x})}{n}$

Second moment about the mean $\mu_2 = \dfrac{\sum(x-\bar{x})^2}{n}$

Third moment about the mean $\mu_3 = \dfrac{\sum(x-\bar{x})^3}{n}$

Fourth moment about the mean $\mu_4 = \dfrac{\sum(x-\bar{x})^4}{n}$

In general,

r^{th} moment about the mean $\mu_r = \dfrac{\sum(x-\bar{x})^r}{n}$

In case of a frequency distribution consisting of n observations $x_1, x_2, \ldots, x_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ having arithmetic mean $\bar{x}$,

$$N = \sum f, \ \bar{x} = \frac{\sum fx}{N}$$

The various moments about the actual mean are given by the following:

First moment about the mean $\mu_1 = \dfrac{\sum f(x-\bar{x})}{N}$

Second moment about the mean $\mu_2 = \dfrac{\sum f(x-\bar{x})^2}{N}$

Third moment about the mean $\mu_3 = \dfrac{\sum f(x-\bar{x})^3}{N}$

Fourth moment about the mean $\mu_4 = \dfrac{\sum f(x-\bar{x})^4}{N}$

In general,

r^{th} moment about the mean $\mu_r = \dfrac{\sum f(x-\bar{x})^r}{N}$

Properties of Central Moment

(i) The first moment about the mean is always zero, i.e., $\mu_1 = 0$.

(ii) The second moment about the mean measures variance, i.e.,

$$\mu_2 = \sigma^2 \quad \text{or} \quad \text{SD} = \sigma = \pm\sqrt{\mu_2}$$

(iii) The third moment about the mean measures skewness.
If $\mu_3 > 0$, the distribution is positively skewed.
If $\mu_3 < 0$, the distribution is negatively skewed.
If $\mu_3 = 0$, the distribution is symmetrical.

$$\text{Skewness } \beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

(iv) The fourth moment about the mean measures kurtosis. It gives information about the peak of a frequency distribution, i.e., whether it is more peaked or more flat topped than a normal curve.

$$\text{Kurtosis } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

(v) In a symmetric distribution, all odd moments are zero, i.e., $\mu_1 = \mu_3 = \mu_5 = \cdots = \mu_{2r+1} = 0$.

(b) Raw Moment or Moments About Arbitrary Origin When the actual mean of a distribution is a fraction, it is tedious to calculate central moments. In such cases, moments about an arbitrary origin 'a' is calculated and then these moments are converted into the moments about actual mean. The moments about the arbitrary origin are known as raw moments and are denoted by μ_r'. Thus, μ_1' denotes the first moment about an arbitrary origin, μ_2' denotes the second moment about an arbitrary origin and so on.

The various raw moments are given by the following:

First moment about the arbitrary origin $\mu_1' = \dfrac{\sum (x-a)}{n}$

Second moment about the arbitrary origin $\mu_2' = \dfrac{\sum (x-a)^2}{n}$

Third moment about the arbitrary origin $\mu_3' = \dfrac{\sum (x-a)^3}{n}$

Fourth moment about the arbitrary origin $\mu_4' = \dfrac{\sum (x-a)^4}{n}$

In general,

r^{th} moment about the arbitrary origin $\mu_r' = \dfrac{\sum (x-a)^r}{n}$

In case of frequency distribution consisting of n observations $x_1, x_2, \ldots, x_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ having the arbitrary origin a,

$$N = \sum f,\ d = x - a$$

The various moments about the arbitrary origin are given by the following:

First moment about the arbitrary origin $\mu_1' = \dfrac{\sum fd}{N}$

Second moment about the arbitrary origin $\mu_2' = \dfrac{\sum fd^2}{N}$

Third moment about the arbitrary origin $\mu_3' = \dfrac{\sum fd^3}{N}$

Fourth moment about the arbitrary origin $\mu_4' = \dfrac{\sum fd^4}{N}$

In general,

r^{th} moment about the arbitrary origin $\mu_r' = \dfrac{\sum fd^r}{N}$

In case of frequency distribution with 'a' as arbitrary origin and h as width of the class interval,

$$N = \sum f,\ d = \frac{x-a}{h}$$

The various moments about the arbitrary origin are given by the following:

First moment about the arbitrary origin $\mu_1' = h\dfrac{\sum fd}{N}$

Second moment about the arbitrary origin $\mu_2' = h^2\dfrac{\sum fd^2}{N}$

Third moment about the arbitrary origin $\mu_3' = h^3\dfrac{\sum fd^3}{N}$

Fourth moment about the arbitrary origin $\mu_4' = h^4 \dfrac{\sum fd^4}{N}$

In general,

r^{th} moment about the arbitrary origin $\mu_r' = h^r \dfrac{\sum fd^r}{N}$

(c) Relation between Central and Raw Moments The moments about the actual mean, i.e., central moments and moments about the arbitrary origin, i.e., raw moments are related with each other by the following equations:

First central moment $\mu_1 = \mu_1' - \mu_1' = 0$

Second central moment $\mu_2 = \mu_2' - (\mu_1')^2$

Third central moment $\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2(\mu_1')^3$

Fourth central moment $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

Similarly, the raw moments can be expressed in terms of central moments.

First raw moment $\mu_1' = \bar{x} - a$

Second raw moment $\mu_2' = \mu_2 + (\mu_1')^2$

Third raw moment $\mu_3' = \mu_3 + 3\mu_2 \mu_1' + (\mu_1')^3$

Fourth raw moment $\mu_4' = \mu_4 + 4\mu_3 \mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$

(d) Moments about Zero The moments about zero are denoted by $\nu_1, \nu_2, \nu_3, \nu_4$, etc. The various moments about zero are given by the following:

First moment about zero $\nu_1 = \dfrac{\sum fx}{N}$

Second moment about zero $\nu_2 = \dfrac{\sum fx^2}{N}$

Third moment about zero $\nu_3 = \dfrac{\sum fx^3}{N}$

Fourth moment about zero $\nu_4 = \dfrac{\sum fx^4}{N}$

In general, r^{th} moment about zero $\nu_r = \dfrac{\sum fx^r}{N}$

(e) Relation between Moments about Zero and Central Moments The moments about zero and central moments are related by the following equations:

First moment about zero $\nu_1 = a + \mu_1' = \bar{x}$

Second moment about zero $\nu_2 = \mu_2 + (\nu_1)^2$

Third moment about zero $\nu_3 = \mu_3 + 3\nu_1\nu_2 - 2(\nu_1)^3$

Fourth moment about zero $\nu_4 = \mu_4 + 4\nu_1\nu_3 - 6(\nu_1)^2\nu_2 + 3(\nu_1)^4$

8. Skewness *Skewness* is a measure that refers to the extent of symmetry or asymmetry in a distribution. A distribution is said to be *symmetrical* when its mean, median, and mode are equal, and the frequencies are symmetrically distributed about the mean. A symmetrical distribution when plotted on a graph will give a perfectly bell-shaped curve which is known as a *normal curve* (Fig. 17.1).

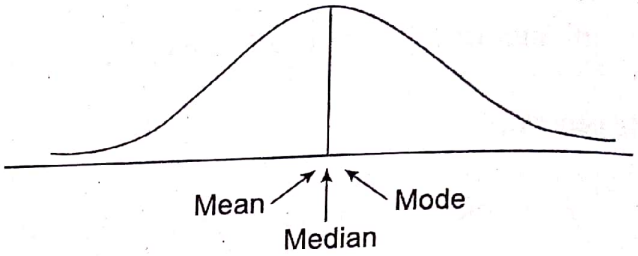
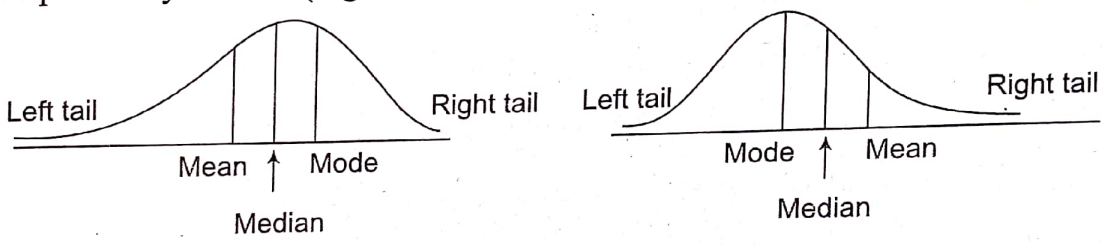


Fig. 17.1 Skewness Curve

A distribution is said to be asymmetrical or skewed when the mean, median, and mode are not equal, i.e., the mean, median, and mode do not coincide. If the curve has a longer tail towards the left, it is said to be a negatively skewed distribution (Fig. 17.2*a*). If the curve has a longer tail towards the right, it is said to be positively skewed (Fig. 17.2*b*).



(a) Negatively skewed distribution (b) Positively skewed distribution

Fig. 17.2 Skewed Distribution Curve

Skewness gives an idea of the nature and degree of concentration of observations about the mean.

(a) Measures of Skewness A measure of skewness gives the extent and direction of skewness of a distribution. These measures can be absolute or relative. The absolute measures are also known as measures of skewness.

$$\text{Absolute skewness} = \text{Mean} - \text{Mode}$$

If the value of the mean is greater than the mode, the skewness will be positive and if the value of the mean is less than the mode, the skewness will be negative.

The relative measures of skewness is called the *coefficient of skewness*.

(b) Karl Pearson's Coefficient of Skewness Karl Pearson's coefficient of skewness denoted by S_k, is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

When the mode is ill-defined and the distribution is moderately skewed, the averages have the following relationship:

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

$$S_k = \frac{\text{Mean} - (3\,\text{Median} - 2\,\text{Mean})}{\text{Standard Deviation}} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

The coefficient of skewness usually lies between -1 and 1.
For a positively skewed distribution, $S_k > 0$.
For a negatively skewed distribution, $S_k < 0$.
For a symmetrical distribution, $S_k = 0$.

HISTORICAL DATA

Karl Pearson (1857–1936) was an influential English mathematician and biostatician. He has been credited with establishing the discipline of mathematical statistics and contributed significantly to the field of biometrics, meteorology, theories of social Darwinism and eugenics.

In 1911, he founded the world's first university statistics department at University College London.

His major contributions include correlation coefficient, methods of moments, Pearson's system of continuous curves, Chi-distance, p-value, the statistical hypothesis testing theory and the statistical decision theory, Pearson's chi-squared test and principal component analysis.

17.3 CORRELATION

Correlation and regression are the most commonly used techniques for investigating the relationship between two quantitative variables. *Correlation* refers to the relationship of two or more variables. It measures the closeness of the relationship between the variables. *Regression* establishes a functional relationship between the variables. The coefficient of correlation is a relative measure whereas the regression coefficient is an absolute figure.

Two variables are said to be correlated if a change in one variable affects a change in the other variable. Such a data connecting two variables is called *bivariate data*. Thus, correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other. Some examples of such a relationship are as follows:

1. Relationship between heights and weights.
2. Relationship between price and demand of commodity.
3. Relationship between rainfall and yield of crops.
4. Relationship between age of husband and age of wife.

17.3.1 Karl Pearson's Coefficient of Correlation

The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r.

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where cov (X, Y) is covariance of variables X and Y,

σ_X is the standard deviation of variable X,

and σ_Y is the standard deviation of variable Y.

This expression is known as **Karl Pearson's** coefficient of correlation or Karl Pearson's product-moment coefficient of correlation.

$$\text{cov}(X,Y) = \frac{1}{n}\sum (x-\bar{x})(y-\bar{y})$$

$$\sigma_X = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$\sigma_Y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$$

The above expression can be further modified.

Expanding the terms,

$$r = \frac{\sum (xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y})}{\sqrt{\sum (x^2 - 2x\bar{x} + \bar{x}^2)}\sqrt{\sum (y^2 - 2y\bar{y} + \bar{y}^2)}} = \frac{\sum xy - \bar{y}\sum x - \bar{x}\sum y + \bar{x}\bar{y}\sum 1}{\sqrt{\sum x^2 - 2\bar{x}\sum x + \bar{x}^2 \sum 1}\sqrt{\sum y^2 - 2\bar{y}\sum y + \bar{y}^2 \sum 1}}$$

$$= \frac{\sum xy - \frac{\sum y}{n}\sum x - \frac{\sum x}{n}\sum y + \frac{\sum x}{n}\frac{\sum y}{n}\cdot n}{\sqrt{\sum x^2 - 2\frac{\sum x}{n}\sum x + \left(\frac{\sum x}{n}\right)^2 n}\sqrt{\sum y^2 - 2\frac{\sum y}{n}\sum y + \left(\frac{\sum y}{n}\right)^2 n}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

17.3.2 Properties of Coefficient of Correlation

1. The Coefficient of Correlation Lies between –1 and 1, i.e., $-1 \le r \le 1$.

Proof Let $\bar{x}$ and $\bar{y}$ be the mean of x and y series and σ_x and σ_y be their respective standard deviations.

Let $\sum \left(\frac{x-\bar{x}}{\sigma_x} \pm \frac{y-\bar{y}}{\sigma_y}\right)^2 \ge 0$ [$\because$ sum of squares of real quantities cannot be negative]

$$\frac{\sum (x-\bar{x})^2}{\sigma_x^2} + \frac{\sum (y-\bar{y})^2}{\sigma_y^2} \pm \frac{2\sum (x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y} \ge 0$$

$$n + n \pm 2nr \ge 0$$

$$2n \pm 2nr \ge 0$$

$$2n(1 \pm r) \ge 0$$

$$1 \pm r \ge 0$$

i.e., $1 + r \ge 0$ or $1 - r \ge 0$

$r \ge -1$ or $r \le 1$

Hence, the coefficient of correlation lies between –1 and 1, i.e., $-1 \le r \le 1$.

2. Correlation Coefficient is Independent of Change of Origin and Change of Scale.

Proof Let $d_x = \frac{x-a}{h}$, $d_y = \frac{y-b}{k}$

$$x = a + hd_x, \quad y = b + kd_y$$

where a, b, h (>0) and k(>0) are constants.

$$x = a + hd_x \Rightarrow \bar{x} = a + h\bar{d}_x \Rightarrow x - \bar{x} = h(d_x - \bar{d}_x)$$

$$y = b + kd_y \Rightarrow \bar{y} = b + h\bar{d}_y \Rightarrow y - \bar{y} = k(d_y - \bar{d}_y)$$

$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{\sum h(d_x-\bar{d}_x)\,k(d_y-\bar{d}_y)}{\sqrt{\sum h^2(d_x-\bar{d}_x)^2}\sqrt{\sum k^2(d_y-\bar{d}_y)^2}}$$

$$= \frac{\sum (d_x-\bar{d}_x)(d_y-\bar{d}_y)}{\sqrt{\sum (d_x-\bar{d}_x)^2}\sqrt{(d_y-\bar{d}_y)^2}} = r_{d_x d_y}$$

Hence, the correlation coefficient is independent of change of origin and change of scale.

Note Since correlation coefficient is independent of change of origin and change of scale,

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}}$$

3. Two Independent Variables are Uncorrelated.

Proof If random variables X and Y are independent,

$$\sum (x-\bar{x})(y-\bar{y}) = 0 \quad \text{or} \quad \text{cov}(X, Y) = 0$$

$$\therefore \quad r = 0$$

Thus, if X and Y are independent variables, they are uncorrelated.

Note The converse of the above property is not true, i.e., two uncorrelated variables may not be independent.

EXAMPLE 17.1 *Calculate the correlation coefficient between x and y using the following data:*

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Solution: $n = 6$

x	y	x^2	y^2	xy
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\sum x = 36$	$\sum y = 60$	$\sum x^2 = 266$	$\sum y^2 = 706$	$\sum xy = 293$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}} = \frac{293 - \frac{(36)(60)}{6}}{\sqrt{266 - \frac{(36)^2}{6}}\sqrt{706 - \frac{(60)^2}{6}}} = -0.9203$$

EXAMPLE 17.2 *Calculate the correlation coefficient between for the following values of demand and the corresponding price of a commodity:*

Demand in Quintals	65	66	67	67	68	69	70	72
Price in Rupees Per Kg	67	68	65	68	72	72	69	71

Solution: Let the demand in quintal be denoted by x and the price in rupees per kg be denoted by y.

$$n = 8$$

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
65	67	–3	–2	9	4	6
66	68	–2	–1	4	1	2
67	65	–1	–4	1	16	4
67	68	–1	–1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
$\sum x = 544$	$\sum y = 552$	$\sum(x-\bar{x}) = 0$	$\sum(y-\bar{y}) = 0$	$\sum(x-\bar{x})^2 = 36$	$\sum(y-\bar{y})^2 = 44$	$\sum(x-\bar{x})(y-\bar{y}) = 24$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{24}{\sqrt{36}\sqrt{44}} = 0.603$$

EXAMPLE 17.3 *Calculate the coefficient of correlation between the ages of cars and annual maintenance costs.*

Age of Cars (Year)	2	4	6	7	8	10	12
Annual Maintenance Cost (₹)	1600	1500	1800	1900	1700	2100	2000

Solution: Let the ages of cars in years be denoted by x and annual maintenance costs in rupees be denoted by y.

Let $a = 7$ and $b = 1800$ be the assumed means of x and y series respectively.

Let $h = 1$, $k = 100$

$$d_x = \frac{x-a}{h} = \frac{x-7}{1} = x-7$$

$$d_y = \frac{y-b}{k} = \frac{y-1800}{100}$$

$$n = 7$$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
2	1600	–5	–2	25	4	10
4	1500	–3	–3	9	9	9
6	1800	–1	0	1	0	0
7	1900	0	1	0	1	0
8	1700	1	–1	1	1	–1
10	2100	3	3	9	9	9
12	2000	5	2	25	4	10
		$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 = 70$	$\sum d_y^2 = 28$	$\sum d_x d_y = 37$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}\sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}} = \frac{37-0}{\sqrt{70-0}\sqrt{28-0}} = 0.836$$

EXAMPLE 17.4

A computer operator while calculating the coefficient between two variates x and y for 25 pairs of observations obtained the following constants:

$$n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508$$

It was later discovered at the time of checking that he had copied down two pairs as (6, 14) *and* (8, 6) *while the correct pairs were* (8,12) *and* (6, 8). *Obtain the correct value of the correlation coefficient.*

Solution: $n = 25$

$$\text{Corrected } \sum x = \text{Incorrect } \sum x - (\text{Sum of incorrect } x) + (\text{Sum of correct } x)$$
$$= 125 - (6+8) + (8+6)$$
$$= 125$$

Similarly,

$$\text{Corrected } \sum y = 100 - (14+6) + (12+8) = 100$$

$$\text{Corrected } \sum x^2 = 650 - (6^2+8^2) + (8^2+6^2) = 650$$

$$\text{Corrected } \sum y^2 = 460 - (14^2+6^2) + (12^2+8^2) = 436$$

$$\text{Corrected } \sum xy = 508 - (84+48) + (96+48) = 520$$

Correct value of correlation coefficient

$$r=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{\sum x^2-\frac{(\sum x)^2}{n}}\sqrt{\sum y^2-\frac{(\sum y)^2}{n}}}=\frac{520-\frac{(125)(100)}{25}}{\sqrt{650-\frac{(125)^2}{25}}\sqrt{436-\frac{(100)^2}{25}}}=0.67$$

EXERCISE 17.1

1. Calculate the coefficient of correlation between x and y.

x	2	4	5	6	8	11
y	18	12	10	8	7	5

[Ans.: –0.92]

2. Find the coefficient of correlation between x and y for the following data:

x	10	12	18	24	23	27
y	13	18	12	25	30	10

[Ans.: 0.223]

3. From the following information relating to the stock exchange quotations for two shares A and B, ascertain by using Pearson's coefficient of correlation how shares A and B are correlated in their prices?

Price Share (A) ₹	160	164	172	182	166	170	178
Price Share (B) ₹	292	280	260	234	266	254	230

[Ans.: –0.96]

4. Find the correlation coefficient between the income and expenditure of a wage earner.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Income	46	54	56	56	58	60	62
Expenditure	36	40	44	54	42	58	54

[Ans.: 0.769]

5. From the following data, examine whether the input of oil and output of electricity can be said to be correlated.

Input of Oil	6.9	8.2	7.8	4.8	9.6	8.0	7.7
Output of Electricity	1.9	3.5	6.5	1.3	5.5	3.5	2.2

[Ans.: 0.696]

6. For the following data, show that cov $(x, x^2)=0$.

x	–3	–2	–1	0	1	2	3
x^2	9	4	1	0	1	4	9

7. Find the coefficient of correlation between x and y for the following data:

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

[Ans.: 0.9032]

8. The following data gave the growth of employment in lacs in the organized sector in India between 1988 and 1995:

Year	1988	1989	1990	1991	1992	1993	1994	1995
Public Sector	98	101	104	107	113	120	125	128
Private Sector	65	65	67	68	68	69	68	68

Find the correlation coefficient between the employment in public and private sectors.

[Ans.: 0.77]

9. Calculate Karl Pearson's coefficient of correlation from the following data, using

20 as the working mean for price and 70 as working mean for demand.

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

[Ans.: –0.954**]**

10. A sample of 25 pairs of values x and y lead to the following results:

$$\sum x = 127, \sum y = 100, \sum x^2 = 760,$$
$$\sum y^2 = 449, \sum xy = 500$$

Later on, it was found that two pairs of values were taken as (8, 14) and (8, 6) instead of the correct values (8, 12) and (6, 8). Find the corrected coefficient between x and y.

[Ans.: –0.31**]**

17.3.3 Rank Correlation

Let a group of n individuals be arranged in order of merit with respect to some characteristics. The same group would be given a different order (rank) for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the *rank correlation* in the characteristics A and B for that group of individuals.

Spearman's Rank Correlation Coefficient Let x, y be the ranks of the i^{th} individual in two characteristics A and B respectively, where $i = 1, 2, ..., n$. Assuming that no two individuals have the same rank either for x or y, each of the variables x and y take the values 1, 2, ..., n.

$$\bar{x} = \bar{y} = \frac{1+2+3+\cdots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\sum (x-\bar{x})^2 = \sum (x^2 - 2x\bar{x} + \bar{x}^2) = \sum x^2 - 2\bar{x}\sum x + \bar{x}^2 \sum 1$$

$$= \sum x^2 - 2n\bar{x}^2 + n\bar{x}^2 \qquad \left[\because \sum x = n\bar{x} \text{ and } \sum 1 = n\right]$$

$$= \sum x^2 - n\bar{x}^2 = (1^2 + 2^2 + \cdots + n^2) - n\left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{1}{12}(n^3 - n)$$

Similarly, $\sum (y-\bar{y})^2 = \frac{1}{12}(n^3 - n)$

If d denotes the difference between the ranks of the i^{th} individual in the two variables,

$$d = x - y = (x-\bar{x}) - (y-\bar{y}) \qquad [\because \bar{x} = \bar{y}]$$

Squaring and summing over i from 1 to n,

$$\sum d^2 = \sum [(x-\bar{x}) - (y-\bar{y})]^2 = \sum (x-\bar{x})^2 + \sum (y-\bar{y})^2 - 2\sum (x-\bar{x})(y-\bar{y})$$

$$\sum (x-\bar{x})(y-\bar{y}) = \frac{1}{2}\left[\sum (x-\bar{x})^2 + \sum (y-\bar{y})^2 - \sum d^2\right] = \frac{1}{12}(n^3 - n) - \frac{1}{2}\sum d^2$$

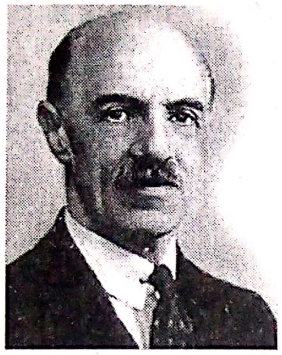
Hence, the coefficient of correlation between these variables is

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{\frac{1}{12}(n^3 - n) - \frac{1}{2}\sum d^2}{\frac{1}{12}(n^3 - n)} = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

This is called Spearman's rank correlation coefficient and is denoted by r.

Note $\sum d = \sum (x - y) = \sum x - \sum y = n(\bar{x} - \bar{y}) = 0$

HISTORICAL DATA



Charles Spearman (1863–1945) was an English psychologist known for his work in statistics, as a pioneer of factor analysis, and for Spearman's rank correlation coefficient. He also did seminal work on models for human intelligence, including the theory that disparate cognitive test scores reflect a single General intelligence factor and coining the term g factor.

In statistics, Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s.

EXAMPLE 17.5

Ten competitors in a musical test were ranked by the three judges A, B, and C in the following order:

Rank by A	1	6	5	10	3	2	4	9	7	8
Rank by B	3	5	8	4	7	10	2	1	6	9
Rank by C	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method, find which pair of judges has the nearest approach to common liking in music.

Solution: $n = 10$

Rank by A x	*Rank by B* y	*Rank by C* z	$d_1 = x - y$	$d_2 = y - z$	$d_3 = z - x$	d_1^2	d_2^2	d_3^2
1	3	6	–2	–3	5	4	9	25
6	5	4	1	1	–2	1	1	4
5	8	9	–3	–1	4	9	1	16
10	4	8	6	–4	–2	36	16	4
3	7	1	–4	6	–2	16	36	4
2	10	2	–8	8	0	64	64	0
4	2	3	2	–1	–1	4	1	1
9	1	10	8	–9	1	64	81	1
7	6	5	1	1	–2	1	1	4
8	9	7	–1	2	–1	1	4	1
			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 200$	$\sum d_2^2 = 214$	$\sum d_3^2 = 60$

$$r(x, y) = 1 - \frac{6\sum d_1^2}{n(n^2-1)} = 1 - \frac{6(200)}{10\left[(10)^2-1\right]} = -0.21$$

$$r(y, z) = 1 - \frac{6\sum d_2^2}{n(n^2-1)} = 1 - \frac{6(214)}{10\left[(10)^2-1\right]} = -0.296$$

$$r(z, x) = 1 - \frac{6\sum d_3^2}{n(n^2-1)} = 1 - \frac{6(60)}{10\left[(10)^2-1\right]} = 0.64$$

Since $r(z, x)$ is maximum, the pair of judges A and C have the nearest common approach.

EXAMPLE 17.6

The coefficient of rank correlation of the marks obtained by 10 students in physics and chemistry was found to be 0.5. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the rank coefficient of the rank correlation.

Solution:

$$n = 10$$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$0.5 = 1 - \frac{6\sum d^2}{10(100-1)}$$

$$\therefore \quad \sum d^2 = 82.5$$

$$\text{Correct } \sum d^2 = \text{Incorrect } \sum d^2 - (\text{Incorrect rank difference})^2 + (\text{Correct rank difference})^2$$

$$= 82.5 - (3)^2 + (7)^2 = 122.5$$

$$\text{Correct coefficient of rank correlation } r = 1 - \frac{6(122.5)}{10(100-1)} = 0.26$$

Tied Ranks If there is a tie between two or more individuals ranks, the rank is divided among items equally, for example, if two items have fourth rank, the 4th and 5th rank is divided between them equally and is given as $\frac{4+5}{2} = 4.5^{\text{th}}$ rank to each of them. If three items have the same 4th rank, each of them is given $\frac{4+5+6}{3} = 5^{\text{th}}$ rank. As a result of this, the following adjustment or correction is made in the rank correlation formula. If m is the number of items having equal ranks then the factor $\frac{1}{12}(m^3 - m)$ is added to $\sum d^2$. If there are more than one cases of this type, this factor is added corresponding to each case.

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{n(n^2-1)}$$

EXAMPLE 17.7 *Obtain the rank correlation coefficient from the following data:*

x	10	12	18	18	15	40
y	12	18	25	25	50	25

Solution: $n = 6$

x	y	Rank x	Rank y	$d = x - y$	d^2
10	12	1	1	0	0
12	18	2	2	0	0
18	25	4.5	4	0.5	0.25
18	25	4.5	4	0.5	0.25
15	50	3	6	–3	9
40	25	6	4	2	4
					$\sum d^2 = 13.5$

There are two items in the x series having equal values at the rank 4. Each is given the rank 4.5. Similarly, there are three items in the y series at the rank 3. Each of them is given the rank 4.

$m_1 = 2, m_2 = 3$

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)} = 1 - \frac{6\left[13.5 + \frac{1}{12}(8-2) + \frac{1}{12}(27-3)\right]}{6\left[(6)^2 - 1\right]} = 0.5429$$

EXERCISE 17.2

1. Compute Spearman's rank correlation coefficient from the following data:

x	18	20	34	52	12
y	39	23	35	18	46

[Ans.: –0.9]

2. Two judges gave the following ranks to a series of eight one-act plays in a drama competition. Examine the relationship between their judgements.

Judge A	8	7	6	3	2	1	5	4
Judge B	7	5	4	1	3	2	6	8

[Ans.: 0.62]

3. From the following data, calculate Spearman's rank correlation between x and y.

x	36	56	20	42	33	44	50	15	60
y	50	35	70	58	75	60	45	80	38

[Ans.: 0.92]

4. Ten competitors in a voice test are ranked by three judges in the following order:

Rank by First Judge	6	10	2	9	8	1	5	3	4	7
Rank by Second Judge	5	4	10	1	9	3	8	7	2	6
Rank by Third Judge	4	8	2	10	7	6	9	1	3	6

Use the method of rank correlation to gauge which pairs of judges has the nearest approach to common liking in voice.

[**Ans.:** The first and third judge]

5. The following table gives the scores obtained by 11 students in English and Tamil translation. Find the rank correlation coefficient.

Scores in English	40	46	54	60	70	80	82	85	85	90	95
Scores in Tamil	45	45	50	43	40	75	55	72	65	42	70

[**Ans.:** 0.36]

6. Calculate Spearman's coefficient of rank correlation for the following data:

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

[**Ans.:** –0.905]

7. Following are the scores of ten students in a class and their IQ:

Score	35	40	25	55	85	90	65	55	45	50
IQ	100	100	110	140	150	130	100	120	140	110

Calculate the rank correlation coefficient between the score IQ.

[**Ans.:** 0.47]

17.4 REGRESSION

Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated. *Regression analysis* is used to predict or estimate one variable in terms of the other variable. It is a highly valuable tool for prediction purpose in economics and business. It is useful in statistical estimation of demand curves, supply curves, production function, cost function, consumption function, etc.

17.4.1 Lines of Regression

If the variables, which are highly correlated, are plotted on a graph then the points lie in a narrow strip. If all the points cluster around a straight line, the line is called the *line of regression*. The line of regression is the line of best fit and is obtained by the principle of least squares.

1. Line of Regression of y on x It is the line which gives the best estimate for the values of y for any given values of x. The regression equation of y on x is given by

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

It is also written as:

$$y = a + bx$$

2. Line of Regression of x on y It is the line which gives the best estimate for the values of x for any given values of y. The regression equation for x on y is given by

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

It is also written as:

$$x = a + by$$

where $\bar{x}$ and $\bar{y}$ are means of x series and y series respectively, σ_x and σ_y are standard deviations of x series and y series respectively, r is the correlation coefficient between x and y.

17.4.2 Regression Coefficients

The slope b of the line of regression of y on x is also called the *coefficient of regression* of y on x. It represents the increment in the value of y corresponding to a unit change in the value of x.

$$b_{yx} = \text{Regression coefficient of } y \text{ on } x = r\frac{\sigma_y}{\sigma_x}$$

Similarly, the slope b of the line of regression of x on y is called the coefficient of regression of x on y. It represents the increment in the value of x corresponding to a unit change in the value of y.

$$b_{xy} = \text{Regression coefficient of } x \text{ on } y = r\frac{\sigma_x}{\sigma_y}$$

17.4.3 Expressions for Regression Coefficients

(i) $$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$$

$$\sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

and $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$$

(ii) $$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

$$\sigma_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\sigma_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}$$

(iii) $$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$\sigma_x = \sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}$$

$$\sigma_y = \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}$$

$$b_{yx} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{(\sum d_y)^2}{n}}$$

and $$b_{yx} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{(\sum d_y)^2}{n}}$$

17.4.4 Properties of Regression Coefficients

1. ***The coefficient of correlation is the geometric mean of the coefficients of regression, i.e.,*** $r = \sqrt{b_{yx} b_{xy}}$.

Proof $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$b_{yx}\, b_{xy} = r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y} = r^2$$

$$r = \sqrt{b_{yx}\, b_{xy}}$$

2. *If one of the regression coefficients is greater than one, the other must be less than one.*

Proof Let $b_{yx} > 1$

$$r^2 \le 1 \text{ and } r^2 = b_{yx}\, b_{xy}$$

$$b_{yx}\, b_{xy} \le 1$$

$$b_{yx} \le \frac{1}{b_{xy}}$$

Hence, if $b_{yx} > 1$ then $b_{xy} < 1$

3. *The arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation.*

Proof To prove that

$$\frac{1}{2}(b_{yx} + b_{xy}) \ge r$$

i.e., $$\frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right) \ge r$$

$$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \ge 2$$

$$\sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y \ge 0$$

i.e., $$(\sigma_y - \sigma_x)^2 \ge 0$$

which is always true, since the square of a real quantity is always ≥ 0.

4. *Regression coefficients are independent of the change of origin but not of scale.*

Proof Let $d_x = \dfrac{x-a}{h}$, $d_y = \dfrac{y-b}{k}$

$$x = a + hd_x, \quad y = b + kd_y$$

where a, b, h (> 0) and k(> 0) are constants.

$$r_{d_x d_y} = r_{xy}, \quad \sigma_{d_x}^2 = \frac{1}{h^2}\sigma_x^2, \quad \sigma_{d_y}^2 = \frac{1}{k^2}\sigma_y^2$$

$$b_{d_x d_y} = r_{d_x d_y}\frac{\sigma_{d_x}}{\sigma_{d_y}} = r_{xy}\frac{\sigma_x}{h}\frac{k}{\sigma_y} = \frac{k}{h}r_{xy}\frac{\sigma_x}{\sigma_y} = \frac{k}{h}b_{xy}$$

Similarly, $b_{d_y d_x} = \frac{h}{k} b_{yx}$

5. ***Both regression coefficients will have the same sign i.e., either both are positive or both are negative.***
6. ***The sign of correlation is same as that of the regression coefficients, i.e., $r > 0$ if $b_{xy} > 0$ and $b_{yx} > 0$; and $r < 0$ if $b_{xy} < 0$ and $b_{yx} < 0$.***

17.4.5 Properties of Lines of Regression (Linear Regression)

1. The two regression lines x on y and y on x always intersect at their means $(\bar{x}, \bar{y})$.
2. The regression lines become identical if $r = \pm 1$. It follows from the regression equations that $x = \bar{x}$ and $y = \bar{y}$. If $r = 0$, these lines are perpendicular to each other.

EXAMPLE 17.8

The regression lines of a sample are $x + 6y = 6$ and $3x + 2y = 10$. Find
(i) sample means $\bar{x}$ and $\bar{y}$, and
(ii) the coefficient of correlation between x and y.
(iii) Also, estimate y when $x = 12$.

Solution:

(i) The regression lines pass through the point $(\bar{x}, \bar{y})$.

$$\bar{x} + 6\bar{y} = 6 \quad \ldots(1)$$

$$3\bar{x} + 2\bar{y} = 10 \quad \ldots(2)$$

Solving Eqs (1) and (2),

$$\bar{x} = 3, \quad \bar{y} = \frac{1}{2}$$

(ii) Let the line $x + 6y = 6$ be the line of regression of y on x.

$$6y = -x + 6$$

$$y = -\frac{1}{6}x + 1$$

$$\therefore \quad b_{yx} = -\frac{1}{6}$$

Let the line $3x + 2y = 10$ be the line of regression of x on y.

$$3x = -2y + 10$$

$$x = -\frac{2}{3}y + \frac{10}{3}$$

$$\therefore \quad b_{xy} = -\frac{2}{3}$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{\left(-\frac{1}{6}\right)\left(-\frac{2}{3}\right)} = \frac{1}{3}$$

Since b_{yx} and b_{xy} are negative, r is negative.

$$r = -\frac{1}{3}$$

Estimated value of y when $x = 12$ is

$$y = -\frac{1}{6}(12) + 1 = -1$$

EXAMPLE 17.9

From the following results, obtain the two regression equations and estimate the yield when the rainfall is 29 cm and the rainfall, when the yield is 600 kg:

	Yield in kg	*Rainfall in cm*
Mean	508.4	26.7
SD	36.8	4.6

The coefficient of correlation between yield and rainfall is 0.52.

Solution: Let rainfall in cm be denoted by x and yield in kg be denoted by y.

$$\bar{x} = 26.7, \quad \bar{y} = 508.4, \quad \sigma_x = 4.6, \quad \sigma_y = 36.8, \quad r = 0.52$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.52\left(\frac{36.8}{4.6}\right) = 4.16$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.52\left(\frac{4.6}{36.8}\right) = 0.065$$

The equation of the line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 508.4 = 4.16(x - 26.7)$$

$$y = 4.16x + 397.328$$

The equation of the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 26.7 = 0.065(y - 508.4)$$

$$x = 0.065y - 6.346$$

Estimated yield when the rainfall is 29 cm is

$$y = 4.16\ (29) + 397.328 = 517.968 \text{ kg}$$

Estimated rainfall when the yield is 600 kg is

$$x = 0.065\ (600) - 6.346 = 32.654 \text{ cm}$$

EXAMPLE 17.10

The number of bacterial cells (y) per unit volume in a culture at different hours (x) is given below:

x	0	1	2	3	4	5	6	7	8	9
y	43	46	82	98	123	167	199	213	245	272

Fit lines of regression of y on x and x on y. Also, estimate the number of bacterial cells after 15 hours.

Solution: $n = 10$

x	y	x^2	xy	y^2
0	43	0	0	1849
1	46	1	46	2116
2	82	4	164	6724
3	98	9	294	9604
4	123	16	492	15129
5	167	25	835	27889
6	199	36	1194	39601
7	213	49	1491	45369
8	245	64	1960	60025
9	272	81	2448	73984
$\sum x = 45$	$\sum y = 1488$	$\sum x^2 = 285$	$\sum xy = 8924$	$\sum y^2 = 282290$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{8924 - \frac{(45)(1488)}{10}}{285 - \frac{(45)^2}{10}} = 27.0061$$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}} = \frac{8924 - \frac{(45)(1488)}{10}}{282290 - \frac{(1488)^2}{10}} = 0.0366$$

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{10} = 4.5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{1488}{10} = 148.8$$

The equation of the line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 148.8 = 27.0061\,(x - 4.5)$$

$$y = 27.0061x + 27.2726$$

The equation of the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 4.5 = 0.0366(y - 148.8)$$

$$x = 0.0366\,y - 0.9461$$

At $x = 15$ hours,

$y = 27.0061\ (15) + 27.2726 = 432.3641$

EXAMPLE 17.11

Calculate the regression coefficients and find the two lines of regression from the following data:

x	57	58	59	59	60	61	62	64
y	67	68	65	68	72	72	69	71

Find the value of y when $x = 66$.

Solution:

$$n = 8$$

$$\bar{x} = \frac{\sum x}{n} = \frac{480}{8} = 60$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
57	67	–3	–2	9	4	6
58	68	–2	–1	4	1	2
59	65	–1	–4	1	16	4
59	68	–1	–1	1	1	1
60	72	0	3	0	9	0
61	72	1	3	1	9	3
62	69	2	0	4	0	0
64	71	4	2	16	4	8
$\sum x$ = 480	$\sum y$ = 552	$\sum(x-\bar{x})$ = 0	$\sum(y-\bar{y})$ = 0	$\sum(x-\bar{x})^2$ = 36	$\sum(y-\bar{y})^2$ = 44	$\sum(x-\bar{x})(y-\bar{y})$ = 24

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{24}{36} = 0.667$$

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} = \frac{24}{44} = 0.545$$

The equation of regression line of y on x is

$$y-\bar{y}=b_{yx}(x-\bar{x})$$
$$y-69=0.667(x-60)$$
$$y=0.667x+28.98$$

The equation of regression line of x on y is

$$x-\bar{x}=b_{xy}(y-\bar{y})$$
$$x-60=0.545(y-69)$$
$$x=0.545y+22.395$$

Value of y when $x = 66$ is

$$y=0.667\,(66)+28.98=73.002$$

EXAMPLE 17.12

The following data represents rainfall (x) and yield of paddy per hectare (y) in a particular area. Find the linear regression of x on y.

x	113	102	95	120	140	130	125
y	1.8	1.5	1.3	1.9	1.1	2.0	1.7

Solution: Let $a = 120$ and $b = 1.8$ be the assumed means of x and y series respectively.

$$d_x=x-a=x-120$$
$$d_y=y-b=y-1.8$$
$$n=7$$

x	y	d_x	d_y	d_y^2	$d_x d_y$
113	1.8	–7	0	0	0
102	1.5	–18	–0.3	0.09	5.4
95	1.3	–25	–0.5	0.25	12.5
120	1.9	0	0.1	0.01	0
140	1.1	20	–0.7	0.49	–14
130	2.0	10	0.2	0.04	2.0
125	1.7	5	–0.1	0.01	–0.5
$\Sigma x = 825$	$\Sigma y = 11.3$	$\Sigma d_x = -15$	$\Sigma d_y = -1.3$	$\Sigma d_y^2 = 0.89$	$\Sigma d_x d_y = 5.4$

$$b_{xy}=\frac{\sum d_x d_y-\frac{\sum d_x \sum d_y}{n}}{\sum d_y^2-\frac{\left(\sum d_y\right)^2}{n}}=\frac{5.4-\frac{(-15)(-1.3)}{7}}{0.89-\frac{(-1.3)^2}{7}}=4.03$$

$$\bar{x} = \frac{\sum x}{n} = \frac{825}{7} = 117.86$$

$$\bar{y} = \frac{\sum y}{n} = \frac{11.3}{7} = 1.614$$

The equation of the regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 117.86 = 4.03(y - 1.614)$$

$$x = 4.03y + 111.36$$

EXERCISE 17.3

1. The following are the lines of regression $4y = x + 38$ and $9y = x + 288$. Estimate y when $x = 99$ and x when $y = 30$. Also, find the means of x and y.

[**Ans.:** $y = 43, x = 82, \bar{x} = 162, \bar{y} = 50$]

2. The equations of the two lines of regression are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (i) the means of x and y, and (ii) the coefficient of correlation between x and y.

[**Ans.:** $\bar{x} = 15.79, \bar{y} = 3.74$,
(ii) $r = -0.66, b_{yx} = -0.5, b_{xy} = 0.87$]

3. Given var(x) = 25. The equations of the two lines of regression are $5x - y = 22$ and $64x - 45y = 24$. Find (i) $\bar{x}$ and $\bar{y}$, (ii) r, and (iii) σ_y.

[**Ans.:** $\bar{x} = 6, \bar{y} = 8$, (ii) $r = 1.87$
(iii) $\sigma_y = 0.2$]

4. In a partially destroyed laboratory record of analysis of correlation data the following results are legible. Variance = 9, the equations of the lines of regression $4x - 5y + 33 = 0, 20x - 9y - 107 = 0$. Find (i) the mean values of x and y, (ii) the standard deviation of y, and (iii) the coefficient of correlation between x and y.

[**Ans.:** (i) $\bar{x} = 13, \bar{y} = 17$, (ii) $\sigma_y = 4$,
(iii) $r = 0.6$]

5. From a sample of 200 pairs of observation, the following quantities were calculated:

$$\sum x = 11.34, \sum y = 20.78, \sum x^2 = 12.16,$$
$$\sum y^2 = 84.96, \sum xy = 22.13$$

From the above data, show how to compute the coefficients of the equation $y = a + bx$.

[**Ans.:** $a = 0.0005, b = 1.82$]

6. In the estimation of regression equations of two variables x and y, the following results were obtained:

$$\bar{x} = 90, \bar{y} = 70, n = 10,$$
$$\Sigma(x - \bar{x})^2 = 6360, \Sigma(y - \bar{y})^2 = 2860$$
$$\Sigma(x - \bar{x})(y - \bar{y}) = 3900$$

Obtain the two lines of regression.

[**Ans.:** $x = 1.361y - 5.27$,
$y = 0.613x + 14.812$]

7. Find the likely production corresponding to a rainfall of 40 cm from the following data:

	Rainfall (in cm)	*Output (in quintals)*
mean	30	50
SD	5	10
$r = 0.8$		

[**Ans.:** 66 quintals]

8. The following table gives the age of a car of a certain make and annual maintenance cost. Obtain the equation of the line of regression of cost on age.

Age of a Car	2	4	6	8
Maintenance	1	2	2.5	3

[**Ans.:** $x = 0.325\,y + 0.5$]

9. Obtain the equation of the line of regression of y on x from the following data and estimate y for $x = 73$.

x	70	72	74	76	78	80
y	163	170	179	188	196	220

[**Ans.:** $y = 5.31\,x - 212.57$, $y = 175.37$]

10. The heights in cm of fathers (x) and of the eldest sons (y) are given below:

x	165	160	170	163	173	158	178	168	173	170	175	180
y	173	168	173	165	175	168	173	165	180	170	173	178

Estimate the height of the eldest son if the height of the father is 172 cm and the height of the father if the height of the eldest son is 173 cm. Also, find the coefficient of correlation between the heights of fathers and sons.

[**Ans.:** (i) $y = 1.016\,x - 5.123$
(ii) $x = 0.476\,y + 98.98$
(iii) 169.97, 173.45 (iv) $r = 0.696$]

11. Find (i) the lines of regression, and (ii) coefficient of correlation for the following data:

x	65	66	67	67	68	69	70	72
y	67	68	65	66	72	72	69	71

[**Ans.:** (i) $y = 19.64 + 0.72\,x$, $x = 33.29 + 0.5\,y$, (ii) $r = 0.604$]

12. Find the line of regression for the following data and estimate y corresponding to $x = 15.5$.

x	10	12	13	16	17	20	25
y	19	22	24	27	29	33	37

[**Ans.:** $y = 1.21x + 7.71$, $y = 26.465$]

13. The following data give the heights in inches (x) and weights in lbs (y) of a random sample of 10 students:

x	61	68	68	64	65	70	63	62	64	67
y	112	123	130	115	110	125	100	113	116	126

Estimate the weight of a student of height 59 inches.

[**Ans.:** 126.4 lbs]

14. Find the regression equations of y on x from the data given below taking deviations from actual mean of x and y.

Price in Rupees (x)	10	12	13	12	16	15
Demand (y)	40	38	43	45	37	43

Estimate the demand when the price is ₹20.

[**Ans.:** $y = -0.25\,x + 44.25$, $y = 39.25$]

17.5 PROBABILITY

The concept of probability originated from the analysis of the games of chance. Even today, a large number of problems exist which are based on the games of chance, such as tossing of a coin, throwing of dice, and playing of cards etc. The utility of probability in business and economics is most emphatically revealed in the field of predictions for the future. Probability is a concept which measures the degree of uncertainty and that of certainty as a corollary.

The word *probability* or 'chance' is used commonly in day-to-day life, e.g., 'it may rain today', 'India may win the forthcoming cricket match against Sri Lanka', 'the chances of making profits by investing in shares of Company A are very bright, etc. Each of the above sentences involves

an element of uncertainty. A numerical measure of uncertainty is provided by a very important branch of mathematics called *theory of probability*. Before studying the probability theory in detail, it is appropriate to explain certain terms which are essential for the study of the theory of probability.

17.5.1 Some Important Terms and Concepts

1. Random Experiment If an experiment is conducted, any number of times, under identical conditions, there is a set of all possible outcomes associated with it. If the outcome is not unique but may be any one of the possible outcomes, the experiment is called a random experiment, e.g., tossing a coin, throwing a dice.

2. Outcome The result of a random experiment is called an outcome. For example, consider the following:

(a) Suppose a random experiment is 'a coin is tossed'. This experiment gives two possible outcomes–head or tail.
(b) Suppose a random experiment is 'a dice is thrown'. This experiment gives six possible outcomes–1, 2, 3, 4, 5 or 6–on the uppermost face of a dice.

3. Trial and Event Any particular performance of a random experiment is called a trial and outcome. A combination of outcomes is called an event. For example, consider the following:

(a) Tossing of a coin is a trial, and getting a head or tail is an event.
(b) Throwing of a dice is a trial and getting 1 or 2 or 3 or 4 or 5 or 6 is an event.

4. Exhaustive Event The total number of possible outcomes of a random experiment is called an exhaustive event. For example, consider the following:

(a) In tossing of a coin, there are two exhaustive events, viz., head and tail.
(b) In throwing of a dice, there are six exhaustive events, getting 1 or 2 or 3 or 4 or 5 or 6.

5. Mutually Exclusive Events Events are said to be mutually exclusive if the occurrence of one of them prevents the occurrence of all others in the same trial, i.e., they cannot occur simultaneously. For example, consider the following:

(a) In tossing a coin, the events head or tail are mutually exclusive since both head and tail cannot occur at the same time.
(b) In throwing a dice, all the six events, i.e., getting 1 or 2 or 3 or 4 or 5 or 6 are mutually exclusive events.

6. Equally Likely Events The outcomes of a random experiment are said to be equally likely if the occurrence of none of them is expected in preference to others. For example, consider the following:

(a) In tossing a coin, head or tail are equally likely events.
(b) In throwing a dice, all the six faces are equally likely events.

7. Independent Events Events are said to be independent if the occurrence of an event does not have any effect on the occurrence of other events. For example, consider the following:

(a) In tossing a coin, the event of getting a head in the first toss is independent of getting a head in the second, third, and subsequent tosses.
(b) In throwing a dice, the result of the first throw does not affect the result of the second throw.

8. Favourable Events The favourable events in a random experiment are the number of outcomes which entail the occurrence of the event. For example, consider the following:

In throwing of two dice, the favourable events of getting the sum 5 is (1, 4), (4, 1),(2, 3), (3, 2), i.e., 4.

17.5.2 Definitions of Probability

Classical Definition of Probability Let n be the number of equally likely, mutually exclusive, and exhaustive outcomes of a random experiment. Let m be number of the outcomes which are favourable to the occurrence of an event A. The probability of event A occurring, denoted by $P(A)$, is given by

$$P(A) = \frac{\text{Number of outcomes favourable to } A}{\text{Number of exhaustive outcomes}} = \frac{m}{n}$$

Empirical or Statistical Definition of Probability If an experiment is repeated a large number of times under identical conditions, the limiting value of the ratio of the number of times the event A occurs to the total number of trials of the experiment as the number of trials increase indefinitely is called the probability of occurrence of the event A.

Let $P(A)$ be the probability of occurrence of the event A. Let m be the number of times in which an event A occurs in a series of n trials.

$$P(A) = \lim_{n\to\infty} \frac{m}{n}, \text{ provided the limit is finite and unique.}$$

Axiomatic Definition of Probability Before discussing the axiomatic definition of probability, it is necessary to explain certain concepts that are necessary to its understanding.

1. Sample Space A set of all possible outcomes of a random experiment is called a sample space. Each element of the set is called a *sample point* or a *simple event* or an *elementary event.*
The sample space of a random experiment is denoted by S. For example, consider the following:

(a) In a random experiment of tossing of a coin, the sample space consists of two elementary events.
$S = \{H, T\}$

(b) In a random experiment of throwing of a dice, the sample space consists of six elementary events.
$S = \{1, 2, 3, 4, 5, 6\}$
The elements of S can either be single elements or ordered pairs. If two coins are tossed, each element of the sample space consists of the following ordered pairs:
$S = \{(H, H), (H, T), (T, H), (T, T)\}$

2. Event Any subset of a sample space is called an event. In the experiment of throwing of a dice, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Let A be the event that an odd number appears on the dice. Then $A = \{1, 3, 5\}$ is a subset of S. Similarly, let B be the event of getting a number greater than 3. Then $B = \{4, 5, 6\}$ is another subset of S.

■ **Definition of Probability** Let S be a sample space of an experiment and A be any event of this sample space. The probability $P(A)$ of the event A is defined as the real-value set function which associates a real value corresponding to a subset A of the sample space S. The probability $P(A)$ satisfies the following three axioms.

Axiom I: $P(A) \geq 0$, i.e., the probability of an event is a nonnegative number.

Axiom II: $P(S) = 1$, i.e., the probability of an event that is certain to occur must be equal to unity.

Axiom III: If $A_1, A_2, ..., A_n$ are finite mutually exclusive events then

$$P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = \sum_{i=1}^{n} P(A_i)$$

i.e., the probability of a union of mutually exclusive events is the sum of probabilities of the events themselves.

17.5.3 Theorems on Probability

Theorem 1 The probability of an impossible event is zero, i.e., $P(\phi) = 0$, where ϕ is a null set.

Proof An event which has no sample points is called an impossible event and is denoted by ϕ. For a sample space S of an experiment,

$$S \cup \phi = S$$

Taking probability of both the sides,

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

Since S and ϕ are mutually exclusive events,

$$P(S) + P(\phi) = P(S) \qquad \text{[Using Axiom III]}$$

$$\therefore \quad P(\phi) = 0$$

Theorem 2 The probability of the complementary event $\bar{A}$ of A is

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

Proof Let A be an event in the sample space S.

$$A \cup \bar{A} = S$$

$$P(A \cup \bar{A}) = P(S)$$

Since A and $\bar{A}$ are mutually exclusive events,

$$P(A) + P(\bar{A}) = P(S)$$

$$P(A) + P(\bar{A}) = 1 \qquad [\because \ P(S) = 1]$$

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

Note Since A and $\bar{A}$ are mutually exclusive events,

$$A \cup \bar{A} = S \text{ and } A \cap \bar{A} = \phi$$

Corollary Probability of an event is always less than or equal to one, i.e., $P(A) \leq 1$.

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

$$P(A) \leq 1 \qquad [\because \ P(\bar{A}) \geq 0 \text{ by Axiom I}]$$

■ **De Morgan's Laws** Since an event is a subset of a sample space, De Morgan's laws are applicable to events.

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

HISTORICAL DATA

Augustus De Morgan (1806–1871) was a British mathematician and logician. He formulated De Morgan's laws and introduced the term mathematical induction, making its idea rigorous.

His contributions include De Morgan's laws, De Morgan algebra, Relation algebra and Universal algebra.

Beyond his great mathematical legacy, the headquarters of the London Mathematical Society is called De Morgan House and the student society of the Mathematics Department of University College London is called the August De Morgan Society. The crater De Morgan on the Moon is named after him.

Theorem 3 For any two events A and B in a sample space S,

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

Proof From the Venn diagram (Fig. 17.3),

$$B = (A \cap B) \cup (\bar{A} \cap B)$$
$$P(B) = P\left[(A \cap B) \cup (\bar{A} \cap B)\right]$$

Since $(A \cap B)$ and $(\bar{A} \cap B)$ are mutually exclusive events,

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

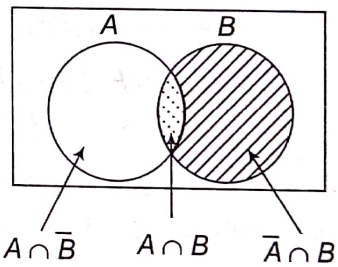


Fig. 17.3 Venn diagram

Similarly, it can be shown that

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

Theorem 4 Additive Law of Probability (Addition Theorem)
The probability that at least one of the events A and B will occur is given by

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

Proof From the Venn diagram (Fig. 17.3),

$$A \cup B = A \cup (\bar{A} \cap B)$$
$$P(A \cup B) = P\left[A \cup (\bar{A} \cap B)\right]$$

Since A and $(\bar{A} \cap B)$ are mutually exclusive events,

$$P(A \cup B) = P(A) + P(\bar{A} \cap B) \quad \text{[Using Axiom III]}$$
$$= P(A) + P(B) - P(A \cap B) \quad \text{[Using Theorem 3]}$$

Remarks

1. If A and B are mutually exclusive events, i.e., $A \cap B = \phi$ then $P(A \cap B) = 0$ according to Theorem 1.

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

2. The event $A \cup B$ (i.e., A or B) denotes the occurrence of either A or B or both. Alternately, it implies the occurrence of at least one of the two events.

   $A \cup B = A + B$

3. The event $A \cap B$ (i.e., A and B) is a compound or joint event that denotes the simultaneous occurrence of the two events.

   $A \cap B = AB$

■ **Corollary 1** From the Venn diagram (Fig. 17.3),

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

where $P(\bar{A} \cap \bar{B})$ is the probability that none of the events A and B occur simultaneously.

■ **Corollary 2** P(Exactly one of A and B occurs) = $P[(A \cap \bar{B}) \cup (\bar{A} \cap B)]$

$$= P(A \cap \bar{B}) + P(\bar{A} \cap B) \qquad [\because \ (A \cap \bar{B}) \cap (\bar{A} \cap B) = \phi]$$
$$= P(A) - P(A \cap B) + P(B) - P(A \cap B) \quad \text{[Using Theorem 3]}$$
$$= P(A) + P(B) - 2P(A \cap B) = P(A \cup B) - P(A \cap B) \quad \text{[Using Theorem 4]}$$

■ **Corollary 3** The addition theorem can be applied for more than two events. If A, B, and C are three events of a sample space S then the probability of occurrence of at least one of them is given by

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

[Applying Theorem 4 on second and third term]

Alternately, the probability of occurrence of at least one of the three events can also be written as

$$P(A \cup B \cup C) = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$$

If A, B, and C are mutually exclusive events,

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

■ **Corollary 4** The probability of occurrence of at least two of the three events is given by

$$P[(A \cap B) \cup (B \cap C) \cup (A \cap C)]$$
$$= P(A \cap B) + P(B \cap C) + P(A \cap C) - 3P(A \cap B \cap C) + P(A \cap B \cap C) \quad \text{[Using Corollary 3]}$$
$$= P(A \cap B) + P(B \cap C) + P(A \cap C) - 2P(A \cap B \cap C)$$

■ **Corollary 5** The probability of occurrence of exactly two of the three events is given by

$$P[(A\cap B\cap \bar{C})\cup(A\cap \bar{B}\cap C)\cup(\bar{A}\cap B\cap C)]$$
$$= P[(A\cap B)\cup(B\cap C)\cup(A\cap C)]-P(A\cap B\cap C) \quad \text{[Using Corollary 2]}$$
$$= P(A\cap B)+P(B\cap C)+P(A\cap C)-3P(A\cap B\cap C) \quad \text{[Using Corollary 4]}$$

■ **Corollary 6** The probability of occurrence of exactly one of the three events is given by

$$P[(A\cap \bar{B}\cap \bar{C})\cup(\bar{A}\cap B\cap \bar{C})\cup(\bar{A}\cap \bar{B}\cap C)] = P(\text{at least one of the three event occur}) - P(\text{at least two of the three events occur})$$
$$= P(A)+P(B)+P(C)-2P(A\cap B)-2P(B\cap C)-2P(A\cap C)+3P(A\cap B\cap C)$$

EXAMPLE 17.13

In a group of 1000 *persons, there are* 650 *who can speak Hindi,* 400 *can speak English, and* 150 *can speak both Hindi and English. If a person is selected at random, what is the probability that he speaks (i) Hindi only, (ii) English only, (iii) only one of the two languages, and (iv) at least one of the two languages?*

Solution: Let A and B be the events that a person selected at random speaks Hindi and English respectively.

$$P(A)=\frac{650}{1000},\quad P(B)=\frac{400}{1000},\quad P(A\cap B)=\frac{150}{1000}$$

(i) Probability that a person selected at random speaks Hindi only

$$P(A\cap \bar{B}) = P(A)-P(A\cap B)=\frac{650}{1000}-\frac{150}{1000}=\frac{1}{2}$$

(ii) Probability that a person selected at random speaks English only

$$P(\bar{A}\cap B) = P(B)-P(A\cap B)=\frac{400}{1000}-\frac{150}{1000}=\frac{1}{4}$$

(iii) Probability that a person selected at random speaks only one of the languages

$$P[(A\cap \bar{B})\cup(\bar{A}\cap B)]=P(A)+P(B)-2P(A\cap B)=\frac{650}{1000}+\frac{400}{1000}-2\left(\frac{150}{1000}\right)=\frac{3}{4}$$

(iv) Probability that a person selected at random speaks at least one of the two languages

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{650}{1000}+\frac{400}{1000}-\frac{150}{1000}=\frac{9}{10}$$

EXAMPLE 17.14

From a city, 3 *newspapers A, B, C are being published. A is read by* 20%, *B is read by* 16%, *C is read by* 14%, *both A and B are read by* 8%, *both A and C are read by* 5%, *both B and C are read by* 4% *and all three A, B, C are read by* 2%. *What is the probability that a randomly chosen person (i) reads at least one of these newspapers, and (ii) reads one of these newspapers?*

Solution: Let A, B, and C be the events that the person reads newspapers A, B, and C respectively.

$$P(A) = 0.2, \quad P(B) = 0.16 \quad P(C) = 0.14$$
$$P(A \cap B) = 0.08, \quad P(A \cap C) = 0.05, \quad P(B \cap C) = 0.04$$
$$P(A \cap B \cap C) = 0.02$$

(i) Probability that the person reads at least one of these newspapers

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$
$$= 0.2 + 0.16 + 0.14 - 0.08 - 0.05 - 0.04 + 0.02 = 0.35$$

(ii) Probability that the person reads none of these newspapers

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = 1 - P(A \cup B \cup C) = 1 - 0.35 = 0.65$$

Alternatively, the problem can be solved by a Venn diagram (Fig. 17.4).

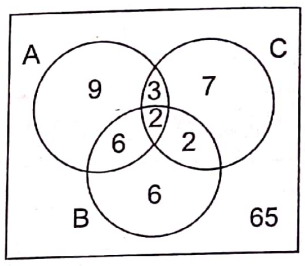


Fig. 17.4

(i) P(the person reads at least one paper) $= 1 - \frac{65}{100} = 0.35$

(ii) P(the person reads none of these papers) $= 0.65$

EXERCISE 17.4

1. The probability that a student passes a Physics test is $\frac{2}{3}$ and the probability that he passes both Physics and English tests is $\frac{14}{45}$. The probability that he passes at least one test is $\frac{4}{5}$. What is the probability that the student passes the English test?

   [**Ans.:** $\frac{4}{9}$]

2. What is the probability of drawing a black card or a king from a well-shuffled pack of playing cards?

   [**Ans.:** $\frac{7}{13}$]

3. A pair of unbiased dice is thrown. Find the probability that (i) the sum of spots is either 5 or 10, and (ii) either there is a doublet or a sum less than 6.

   [**Ans.:** (i) $\frac{7}{36}$ (ii) $\frac{7}{18}$]

4. From a pack of well-shuffled cards, a card is drawn at random. What is the probability that the card drawn is a diamond card or a king card?

   [**Ans.:** $\frac{4}{13}$]

5. A bag contains 6 red, 5 blue, 3 white, and 4 black balls. A ball is drawn at random. Find the probability that the ball is (i) red or black, and(ii) neither red or black.

   [**Ans.:** (i) $\frac{5}{9}$ (ii) $\frac{4}{9}$]

6. There are 100 lottery tickets, numbered from 1 to 100. One of them is drawn at random. What is the probability that the number on it is a multiple of 5 or 7?

   [**Ans.:** $\frac{8}{25}$]

7. From a group of 6 boys and 4 girls, a committee of 3 is to be formed. Find the probability that the committee will include

(i) all three boys or all three girls, (ii) at most two girls, and (iii) at least one girl.

[**Ans.:** (i) $\frac{1}{5}$ (ii) $\frac{29}{30}$ (iii) $\frac{5}{6}$]

8. From a pack of 52 cards, three cards are drawn at random. Find the probability that (i) all three will be aces or all three kings, (ii) all three are pictures or all three are aces, (iii) none is a picture, (iv) at least one is a picture, (v) none is a spade, (vi) at most two are spades, and (vii) at least one is a spade.

[**Ans.:** (i) $\frac{2}{5225}$ (ii) $\frac{56}{5225}$ (iii) $\frac{38}{85}$ (iv) $\frac{47}{85}$ (v) $\frac{703}{1700}$ (vi) $\frac{839}{850}$ (vii) $\frac{997}{1700}$]

9. From a set of 16 cards numbered 1 to 16, one card is drawn at random. Find the probability that (i) the number obtained is divisible by 3 or 7, and (ii) not divisible by 3 and 7.

[**Ans.:** (i) $\frac{7}{16}$ (ii) $\frac{9}{16}$]

10. There are 12 bulbs in a basket of which 4 are working. A person tries to fit them in 3 sockets choosing 3 of the bulbs at random. What is the probability that there will be (i) some light, and (ii) no light in the room?

[**Ans.:** (i) $\frac{41}{55}$ (ii) $\frac{14}{55}$]

Theorem 5 Multiplicative Law or Compound Law of Probability

A compound event is the result of the simultaneous occurrence of two or more event. The probability of a compound event depends upon whether the events are independent or not. Hence, there are two theorems:

(a) Conditional Probability Theorem
(b) Multiplicative Theorem for Independent Events

(a) Conditional Probability Theorem For any two events A and B in a sample space S, the probability of their simultaneous occurrence, i.e., both the events occurrings simultaneously is given by

$$P(A \cap B) = P(A)\, P(B/A) \quad \text{or} \quad P(A \cap B) = P(B)\, P(A/B)$$

where $P(B/A)$ is the conditional probability of B given that A has already occurred. $P(A/B)$ is the conditional probability of A given that B has already occurred.

(b) Multiplicative Theorem for Independent Events If A and B are two independent events, the probability of their simultaneous occurrence is given by

$$P(A \cap B) = P(A)\, P(B)$$
$$P(A \cap B) = P(B)\, P(A/B) \qquad \ldots(1.1)$$

Proof $A = (A \cap B) \cup (A \cap \bar{B})$

Since $(A \cap B)$ and $(A \cap \bar{B})$ are mutually exclusive events,

$$P(A) = P(A \cap B) + P(A \cap \bar{B}) \qquad \text{[Using Axiom III]}$$
$$= P(B)\, P(A/B) + P(\bar{B})\, P(A/\bar{B})$$

If A and B are independent events, the proportion of A's in B is equal to proportion of A's in $\bar{B}$, i.e., $P(A/B) = P(A/\bar{B})$.

$$P(A) = P(A/B)\left[P(B) + P(\bar{B})\right] = P(A/B)$$

Substituting in Eq. (1.1),

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

Remark The additive law is used to find the probability of A or B, i.e., $P(A \cup B)$. The multiplicative law is used to find the probability of A and B, i.e., $P(A \cap B)$.

Corollary 1 If A, B, and C are three events then

$$P(A \cap B \cap C) = P(A)\, P(B/A)\, P\left[C/(A \cap B)\right]$$

If A, B, and C are independent events,

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

Corollary 2 If A and B are independent events then A and $\bar{B}$, $\bar{A}$ and B, $\bar{A}$ and $\bar{B}$ are also independent.

Corollary 3 The probability of occurrence of at least one of the events A, B, C is given by

$$P(A \cup B \cup C) = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$$

If A, B, and C are independent events, their complements will also be independent.

$$P(A \cup B \cup C) = 1 - P(\bar{A})\, P(\bar{B})\, P(\bar{C})$$

Pairwise Independence and Mutual Independence The events A, B, and C are mutually independent if the following conditions are satisfied simultaneously:

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

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

If the last condition is not satisfied, the events are said to be pairwise independent. Hence, mutually independent events are always pairwise independent but not vices versa.

EXAMPLE 17.15

The probability that a student A solves a mathematics problem is $\frac{2}{5}$ and the probability that a student B solves it is $\frac{2}{3}$. What is the probability that (i) the problem is not solved, (ii) the problem is solved, and (iii) both A and B, working independently of each other, solve the problem?

Solution: Let A and B be events that students A and B solve the problem respectively.

$$P(A) = \frac{2}{5}, \quad P(B) = \frac{2}{3}$$

Events A and B are independent.

Probability that the student A does not solve the problem

$$P(\bar{A}) = 1 - P(A) = 1 - \frac{2}{5} = \frac{3}{5}$$

Probability that the student B does not solve the problem

$$P(\bar{B}) = 1 - P(B) = 1 - \frac{2}{3} = \frac{1}{3}$$

(i) Probability that the problem is not solved

$$P(\bar{A} \cap \bar{B}) = P(\bar{A})\, P(\bar{B}) = \frac{3}{5} \times \frac{1}{3} = \frac{1}{5}$$

(ii) Probability that the problem is solved

$$P(A \cup B) = 1 - P(\bar{A} \cap \bar{B}) = 1 - \frac{1}{5} = \frac{4}{5}$$

(iii) Probability that both A and B solve the problem

$$P(A \cap B) = P(A)\, P(B) = \frac{2}{5} \times \frac{2}{3} = \frac{4}{15}$$

EXAMPLE 17.16

A problem in statistics is given to three students A, B, and C, whose chances of solving it independently are $\frac{1}{2}, \frac{1}{3}$ *and* $\frac{1}{4}$ *respectively. Find the probability that*

(i) the problem is solved
(ii) at least two of them are able to solve the problem
(iii) exactly two of them are able to solve the problem
(iv) exactly one of them is able to solve the problem

Solution: Let A, B, and C be the events that students A, B, and C solve the problem respectively.

$$P(A) = \frac{1}{2},\ P(B) = \frac{1}{3},\ P(C) = \frac{1}{4}$$

Events A, B, and C are independent.

(i) Probability that the problem is solved or at least one of them is able to solve the problem is same.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \\ &= P(A) + P(B) + P(C) - P(A)\, P(B) - P(A)\, P(C) - P(B)\, P(C) + P(A)\, P(B)\, P(C) \\ &= \frac{1}{2} + \frac{1}{3} + \frac{1}{4} - \left(\frac{1}{2} \times \frac{1}{3}\right) - \left(\frac{1}{2} \times \frac{1}{4}\right) - \left(\frac{1}{3} \times \frac{1}{4}\right) + \left(\frac{1}{2} \times \frac{1}{3} \times \frac{1}{4}\right) = \frac{3}{4} \end{aligned}$$

(ii) Probability that at least two of them are able to solve the problem

$$\begin{aligned} P[(A \cap B) \cup (B \cap C) \cup (A \cap C)] &= P(A \cap B) + P(B \cap C) + P(A \cap C) - 2P(A \cap B \cap C) \\ &= P(A)\, P(B) + P(B)\, P(C) + P(A)\, P(C) - 2P(A)\, P(B)\, P(C) \\ &= \left(\frac{1}{2} \times \frac{1}{3}\right) + \left(\frac{1}{3} \times \frac{1}{4}\right) + \left(\frac{1}{2} \times \frac{1}{4}\right) - 2\left(\frac{1}{2} \times \frac{1}{3} \times \frac{1}{4}\right) = \frac{7}{24} \end{aligned}$$

(iii) Probability that exactly two of them are able to solve the problem

$$P\left[(A\cap B\cap \bar{C})\cup(A\cap \bar{B}\cap C)\cup(\bar{A}\cap B\cap C)\right]$$
$$= P(A\cap B)+P(B\cap C)+P(A\cap C)-3P(A\cap B\cap C)$$
$$= P(A)\,P(B)+P(B)\,P(C)+P(A)\,P(C)-3P(A)\,P(B)\,P(C)$$
$$=\left(\frac{1}{2}\times\frac{1}{3}\right)+\left(\frac{1}{3}\times\frac{1}{4}\right)+\left(\frac{1}{2}\times\frac{1}{4}\right)-3\left(\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\right)=\frac{1}{4}$$

(iv) Probability that exactly one of them is able to solve the problem

$$P\left[(A\cap \bar{B}\cap \bar{C})\cup(\bar{A}\cap B\cap \bar{C})\cup(\bar{A}\cap \bar{B}\cap C)\right]$$
$$= P(A)+P(B)+P(C)-2P(A\cap B)-2P(B\cap C)-2P(A\cap C)+3P(A\cap B\cap C)$$
$$=\frac{1}{2}+\frac{1}{3}+\frac{1}{4}-2\left(\frac{1}{2}\times\frac{1}{3}\right)-2\left(\frac{1}{3}\times\frac{1}{4}\right)-2\left(\frac{1}{2}\times\frac{1}{4}\right)+3\left(\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\right)=\frac{11}{24}$$

EXAMPLE 17.17

A husband and wife appeared in an interview for two vacancies in an office. The probability of the husband's selection is $\frac{1}{7}$ *and that of the wife's selection is* $\frac{1}{5}$. *Find the probability that (i) both of them are selected, (ii) only one of them is selected, (iii) none of them is selected, and (iv) at least one of them is selected.*

Solution: Let A and B be the events that the husband and wife are selected respectively.

$$P(A)=\frac{1}{7},\qquad P(B)=\frac{1}{5}$$

Events A and B are independent.

(i) Probability that both of them are selected

$$P(A\cap B)=P(A)\,P(B)=\frac{1}{7}\times\frac{1}{5}=\frac{1}{35}$$

(ii) Probability that at least one of them is selected

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{1}{7}+\frac{1}{5}-\frac{1}{35}=\frac{11}{35}$$

(iii) Probability that none of them is selected

$$P(\bar{A}\cap \bar{B})=1-P(A\cup B)=1-\frac{11}{35}=\frac{24}{35}$$

(iv) Probability that only one of them is selected

$$P\left[(A\cap \bar{B})\cup(\bar{A}\cap B)\right]=P(A\cup B)-P(A\cap B)=\frac{11}{35}-\frac{1}{35}=\frac{10}{35}=\frac{2}{7}$$

EXERCISE 17.5

1. Find the probability of drawing 2 red balls in succession from a bag containing 4 red and 5 black balls when the ball that is drawn first is(i) not replaced, and (ii) replaced.

[**Ans.:** (i) $\frac{1}{6}$ (ii) $\frac{16}{81}$]

2. Two aeroplanes bomb a target in succession. The probability of each correctly scoring a hit is 0.3 and 0.2 respectively. The second will bomb only if the first misses the target. Find the probability that (i) the target is hit, and (ii) both fail to score hits.

[**Ans.:** (i) 0.44 (ii) 0.56]

3. Box A contains 5 red and 3 white marbles and Box B contains 2 red and 6 white marbles. If a marble is drawn from each box, what is the probability that they are both of the same colour?

[**Ans.:** 0.109]

4. Two marbles are drawn in succession from a box containing 10 red, 30 white, 20 blue, and 15 orange marbles, with replacement being made after each draw. Find the probability that (i) both are white, and (ii) the first is red and the second is white.

[**Ans.:** (i) $\frac{4}{25}$ (ii) $\frac{4}{75}$]

5. A, B, C are aiming to shoot a balloon. A will succeed 4 times out of 5 attempts. The chance of B to shoot the balloon is 3 out of 4, and that of C is 2 out of 3. If the three aim the balloon simultaneously, find the probability that at least two of them hit the balloon.

[**Ans.:** $\frac{5}{6}$]

6. There are 12 cards numbered 1 to 12 in a box. If two cards are selected, what is the probability that the sum is odd (i) with replacement, and (ii) without replacement?

[**Ans.:** (i) $\frac{1}{2}$ (ii) $\frac{6}{11}$]

7. Two cards are drawn from a well-shuffled pack of 52 cards. Find the probability that they are both aces if the first card is (i) replaced, and (ii) not replaced.

[**Ans.:** (i) $\frac{1}{169}$ (ii) $\frac{1}{221}$]

8. A can hit a target 2 times in 5 shots; B, 3 times in 4 shots; and C, 2 times in 3 shots. They fire a volley. What is the probability that at least 2 shots hit the target?

[**Ans.:** $\frac{2}{3}$]

9. There are two bags. The first bag contains 5 red and 7 white balls and the second bag contains 3 red and 12 white balls. One ball is taken out at random from the first bag and is put in the second bag. Now, a ball is drawn from the second bag. What is the probability that this last ball is red?

[**Ans.:** $\frac{41}{192}$]

10. In a shooting competition, the probability of A hitting the target is $\frac{1}{2}$; of B, is $\frac{2}{3}$; and of C, is $\frac{3}{4}$. If all of them fire at the target, find the probability that (i) none of them hits the target, and (ii) at least one of them hits the target.

[**Ans.:** (i) $\frac{1}{24}$ (ii) $\frac{23}{24}$]

11. The odds against a student X solving a statistics problem are 12 to 10 and the odds in favour of a student Y solving the problem are 6 to 9. What is the probability that the problem will be solved when both try independently of each other?

[Ans.: $\frac{37}{55}$]

12. A bag contains 6 white and 9 black balls. Four balls are drawn at random twice. Find the probability that the first draw will give 4 white balls and the second draw will give 4 black balls if (i) the balls are replaced, and (ii) the balls are not replaced before the second draw.

[Ans.: (i) $\frac{6}{5915}$ (ii) $\frac{3}{715}$]

13. An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. Two balls are transferred from the first urn to the second urn and then one ball is drawn from the latter. What is the probability that the ball drawn is white?

[Ans.: $\frac{5}{26}$]

14. A man wants to marry a girl having the following qualities: fair complexion–the probability of getting such a girl is $\frac{1}{20}$, handsome dowry–the probability is $\frac{1}{50}$, westernized manners and etiquettes–the probability of this is $\frac{1}{100}$. Find the probability of his getting married to such a girl when the possessions of these three attributes are independent.

[Ans.: $\frac{1}{100000}$]

15. A small town has one fire engine and one ambulance available for emergencies. The probability that the fire engine is available when needed is 0.98 and the probability that the ambulance is available when called is 0.92. In the event of an injury resulting from a burning building, find the probability that both the fire engine and ambulance will be available.

[Ans.: 0.9016]

16. In a certain community, 36% of the families own a dog and 22% of the families that own a dog also own a cat. In addition, 30% of the families own a cat. What is the probability that (i) a randomly selected family owns both a dog and a cat, and (ii) a randomly selected family owns a dog given that it owns a cat?

[Ans.: (i) 0.0792 (ii) 0.264]

17.5.4 Bayes' Theorem

Let $A_1, A_2, \ldots, A_n$ be n mutually exclusive and exhaustive events with $P(A_i) \neq 0$ for $i = 1, 2, \ldots, n$ in a sample space S. Let B be an event that can occur in combination with any one of the events $A_1, A_2, \ldots, A_n$ with $P(B) \neq 0$. The probability of the event A_i when the event B has actually occurred is given by

$$P(A_i/B) = \frac{P(A_i)\,P(B/A_i)}{\sum_{i=1}^{n} P(A_i)\,P(B/A_i)}$$

Proof Since $A_1, A_2, \ldots, A_n$ are n mutually exclusive and exhaustive events of the sample space S,

$$S = A_1 \cup A_2 \cup \ldots \cup A_n$$

Since B is another event that can occur in combination with any of the mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_n$,

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_n \cap B)$$

Taking probability of both the sides,

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)$$

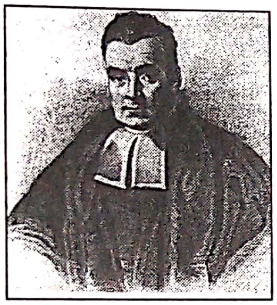
The events $(A_1 \cap B), (A_2 \cap B)$, etc., are mutually exclusive.

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i)\, P(B/A_i)$$

The conditional probability of an event A given that B has already occurred is given by

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)\, P(B/A_i)}{P(B)} = \frac{P(A_i)\, P(B/A_i)}{\sum_{i=1}^{n} P(A_i)\, P(B/A_i)}$$

HISTORICAL DATA



Thomas Bayes (1701–1761) was an English statistician, philosopher and Presbyterian minister who is known for having formulated a specific case of the theorem that bears his name "Bayes' theorem". Bayes never published what would eventually become his most famous accomplishment; his notes were edited and published after his death by Richard Price.

Bayes's solution to a problem of inverse probability was presented in "An Essay towards solving a Problem in the Doctrine of Chances" which was read to the Royal Society in 1763 after Bayes' death. Richard Price shepherded the work through this presentation and its publication in the Philosophical Transactions of the Royal Society of London the following year. This was an argument for using a uniform prior distribution for a binomial parameter and not merely a general postulate. This essay contains a statement of a special case of Bayes's theorem.

EXAMPLE 17.18

In a bolt factory, machines A, B, C manufacture 25%, 35%, and 40% of the total output and out of the total manufacturing, 5%, 4%, and 2% are defective bolts. A bolt is drawn at random from the product and is found to be defective. Find the probabilities that it is manufactured from (i) Machine A, (ii) Machine B, and (iii) Machine C.

Solution: Let A_1, A_2 and A_3 be the events that bolts are manufactured by machines A, B, and C respectively. Let B be the event that the bolt drawn is defective.

$$P(A_1) = \frac{25}{100} = 0.25$$

$$P(A_2) = \frac{35}{100} = 0.35$$

$$P(A_3) = \frac{40}{100} = 0.4$$

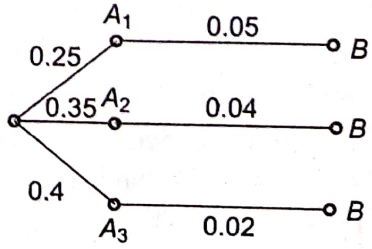


Fig. 17.5

Probability that the bolt drawn is defective given that it is manufactured from Machine A

$$P(B/A_1) = \frac{5}{100} = 0.05$$

Probability that the bolt drawn is defective given that it is manufactured from Machine B

$$P(B/A_2) = \frac{4}{100} = 0.04$$

Probability that the bolt drawn is defective given that it is manufactured from Machine C

$$P(B/A_3) = \frac{2}{100} = 0.02$$

(i) Probability that a bolt is manufactured from Machine A given that it is defective

$$P(A_1/B) = \frac{P(A_1)\,P(B/A_1)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)}$$

$$= \frac{0.25 \times 0.05}{(0.25 \times 0.05) + (0.35 \times 0.04) + (0.4 \times 0.02)} = 0.3622$$

(ii) Probability that a bolt is manufactured from Machine B given that it is defective

$$P(A_2/B) = \frac{P(A_2)\,P(B/A_2)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)}$$

$$= \frac{0.35 \times 0.04}{(0.25 \times 0.05) + (0.35 \times 0.04) + (0.4 \times 0.02)} = 0.4058$$

(iii) Probability that a bolt is manufactured from Machine C given that it is defective

$$P(A_3/B) = \frac{P(A_3)\,P(B/A_3)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)}$$

$$= \frac{0.4 \times 0.02}{(0.25 \times 0.05) + (0.35 \times 0.04) + (0.4 \times 0.02)} = 0.2319$$

EXAMPLE 17.19

Of three persons the chances that a politician, a businessman, or an academician would be appointed the Vice Chancellor (VC) of a university are 0.5, 0.3, 0.2 *respectively. Probabilities that research is promoted by these persons if they are appointed as VC are* 0.3, 0.7, 0.8 *respectively.*

(i) Determine the probability that research is promoted.

(ii) If research is promoted, what is the probability that the VC is an academician?

Solution: Let A_1, A_2, and A_3 be the events that a politician, a businessman or an academician will be appointed as the VC respectively. Let B be the event that research is promoted by these persons if they are appointed as VC.

$$P(A_1) = 0.5$$
$$P(A_2) = 0.3$$
$$P(A_3) = 0.2$$

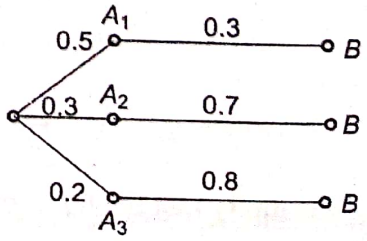


Fig. 17.6

Probability that research is promoted given that a politician is appointed as VC

$$P(B/A_1) = 0.3$$

Probability that research is promoted given that a businessman is appointed as VC

$$P(B/A_2) = 0.7$$

Probability that research is promoted given that an academician is appointed as VC

$$P(B/A_3) = 0.8$$

(i) Probability that research is promoted

$$P(B) = P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)$$
$$= (0.5\times 0.3) + (0.3\times 0.7) + (0.2\times 0.8) = 0.52$$

(ii) Probability that the VC is an academician given that research is promoted by him

$$P(A_3/B) = \frac{P(A_3)\,P(B/A_3)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + P(A_3)\,P(B/A_3)} = \frac{0.2\times 0.8}{0.52} = \frac{4}{13}$$

EXERCISE 17.6

1. There are 4 boys and 2 girls in Room *A* and 5 boys and 3 girls in Room *B*. A girl from one of the two rooms laughed loudly. What is the probability the girl who laughed was from Room *B*?

[**Ans.:** $\frac{9}{17}$]

2. The probability of *X*, *Y*, and *Z* becoming managers are $\frac{4}{9}, \frac{2}{9}$, and $\frac{1}{3}$ respectively. The probabilities that the bonus scheme will be introduced if *X*, *Y*, and *Z* become managers are $\frac{3}{10}, \frac{1}{2}$, and $\frac{4}{5}$ respectively. (i) What is the probability that the bonus scheme will be introduced? (ii) If the bonus scheme has been introduced, what is the probability that the manager appointed was *X*?

[**Ans.:** (i) $\frac{23}{45}$ (ii) $\frac{6}{23}$]

3. A factory has two machines, *A* and *B*. Past records show that the machine *A* produces 30% of the total output and the machine *B*, the remaining 70%. Machine *A* produces 5% defective articles and Machine *B* produces 1% defective items. An item is drawn at random and found to be defective. What is the probability that it was produced (i) by the machine *A*, and (ii) by the Machine *B*?

[**Ans.:** (i) 0.682 (ii) 0.318]

4. A company has two plants to manufacture scooters. Plant I manufactures 80% of the scooters, and Plant II manufactures 20%. At Plant I, 85 out of 100 scooters are rated standard quality or better. At Plant II, only 65 out of 100 scooters are rated standard quality or better. What is the probability that a scooter selected at random came from (i) Plant I, and (ii) Plant II if it is known that the scooter is of standard quality?

[**Ans.:** (i) 0.84 (ii) 0.16]

5. A new pregnancy test was given to 100 pregnant women and 100 nonpregnant

women. The test indicated pregnancy in 92 of the 100 pregnant women and in 12 of the 100 nonpregnant women. If a randomly selected woman takes this test and the test indicates she is pregnant, what is the probability she was not pregnant?

[**Ans.:** $\frac{3}{26}$]

6. An insurance company insured 2000 scooter drivers, 4000 car drivers, and 6000 truck drivers. The probability of an accident is 0.01, 0.03, and 0.15 in the respective category. One of the insured drivers meets with an accident. What is the probability that he is a scooter driver?

[**Ans.:** $\frac{1}{52}$]

7. Consider a population of consumers consisting of two types. The upper-income class of consumers comprise 35% of the population and each member has a probability of 0.8 of purchasing Brand A of a product. Each member of the rest of the population has a probability of 0.3 of purchasing Brand A of the product. A consumer, chosen at random, is found to be the buyer of Brand A. What is the probability that the buyer belongs to the middle-income and lower-income classes of consumers?

[**Ans.:** $\frac{39}{95}$]

8. There are two boxes of identical appearance, each containing 4 spark plugs. It is known that the box I contains only one defective spark plug, while all the four spark plugs of the box II are nondefective. A spark plug drawn at random from a box, selected at random, is found to be nondefective. What is the probability that it came from the box I?

[**Ans.:** $\frac{3}{7}$]

9. Vijay has 5 one-rupee coins and one of them is known to have two heads. He takes out a coin at random and tosses it 5 times–it always falls head upward. What is the probability that it is a coin with two heads?

[**Ans.:** $\frac{8}{9}$]

10. Stores A, B, and C have 50, 75, and 100 employees and, respectively 50, 60, 70 per cent of these are women. Resignations are equally likely among all employees, regardless of sex. One employee resigns and this is a woman. What is the probability that she works in Store C?

[**Ans.:** 0.5]

17.6 RANDOM VARIABLES

A random variable X is a real-valued function of the elements of the sample space of a random experiment. In other words, a variable which takes the real values, depending on the outcome of a random experiment is called a *random variable*. For example, consider the following:

(i) When a fair coin is tossed, $S = \{H, T\}$. If X is the random variable denoting the number of heads, $X(H) = 1$ and $X(T) = 0$
Hence, the random variable X can take values 0 and 1.

(ii) When two fair coins are tossed, $S = \{HH, HT, TH, TT\}$. If X is the random variable denoting the number of heads,
$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0.$
Hence, the random variable X can take values 0, 1, and 2.

(iii) When a fair dice is tossed, $S = \{1, 2, 3, 4, 5, 6\}$.
If X is the random variable denoting the square of the number obtained,
$X(1) = 1, X(2) = 4, X(3) = 9, X(4) = 16, X(5) = 25, X(6) = 36$
Hence, the random variable X can take values 1, 4, 9, 16, 25, and 36.

Types of Random Variables

There are two types of random variables:

(i) Discrete random variables
(ii) Continuous random variables

1. Discrete Random Variables A random variable X is said to be discrete if it takes either finite or countably infinite values. Thus, a discrete random variable takes only isolated values, e.g.,

(i) Number of children in a family
(ii) Number of cars sold by different companies in a year → mass fn
(iii) Number of days of rainfall in a city
(iv) Number of stars in the sky

2. Continuous Random Variables A random variable X is said to be continuous if it takes any values in a given interval. Thus, a continuous random variable takes uncountably infinite values, e.g.,

(i) Height of a person in cm
(ii) Weight of a bag in kg → density fn.
(iii) Temperature of a city in degree Celsius
(iv) Life of an electric bulb in hours

17.7 DISCRETE PROBABILITY DISTRIBUTION

Probability distribution of a random variable is the set of its possible values together with their respective probabilities. Let X be a discrete random variable which takes the values $x_1, x_2, \ldots, x_n$. The probability of each possible outcome x_i is $p_i = p(x_i) = P(X = x_i)$ for $i = 1, 2, \ldots, n$. The number $p(x_i)$, $i = 1,2, \ldots$. must satisfy the following conditions:

(i) $p(x_i) \geq 0$ for all values of i

(ii) $\sum_{i=1}^{\infty} p(x_i) = 1$

The function $p(x_i)$ is called the probability function or probability mass function or probability density function of the random variable X. The set of pairs $\{x, p(x_i)\}$, $i = 1, 2, \ldots, n$ is called the probability distribution of the random variable which can be displayed in the form of a table as shown below:

$X = x_i$	x_1	x_2	x_3	$\ldots x_i$	$\ldots x_n$
$p(x_i) = P(X = x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\ldots p(x_i)$	$\ldots p(x_n)$

17.8 DISCRETE DISTRIBUTION FUNCTION

Let X be a discrete random variable which takes the values $x_1, x_2, \ldots$ such that $x_1 < x_2 < \ldots$ with probabilities $p(x_1), p(x_2) \ldots$ such that $p(x_i) \geq 0$ for all values of i and $\sum_{i=1}^{x} p(x_i) = 1$.

The distribution function $F(x)$ of the discrete random variable X is defined by

$$F(x) = P(X \le x) = \sum_{i=1}^{x} p(x_i)$$

where x is any integer. The function $F(x)$ is also called the cumulative distribution function. The set of pairs $\{x_i, F(x)\}$, $i = 1, 2, \ldots$ is called the cumulative probability distribution.

X	x_1	x_2	...
$F(x)$	$p(x_1)$	$p(x_1) + p(x_2)$	...

EXAMPLE 17.20

A random variable X has the following probability distribution:

X	0	1	2	3	4	5	6	7
$P(X=x)$	a	$4a$	$3a$	$7a$	$8a$	$10a$	$6a$	$9a$

(i) Find the value of a. (ii) Find $P(X < 3)$. (iii) Find the smallest value of m for which $P(X \le m) \ge 0.6$.

Solution:

(i) Since $P(X=x)$ is a probability distribution function,

$$\sum(P(X=x) = 1$$

$$P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6) + P(X=7) = 1$$

$$a + 4a + 3a + 7a + 8a + 10a + 6a + 9a = 1$$

$$a = \frac{1}{48}$$

(ii) $P(X<3) = P(X=0) + P(X=1) + P(X=2) = a + 4a + 3a = 8a = 8\left(\frac{1}{48}\right) = 0.167$

(iii) $P(X \le 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) = a + 4a + 3a + 7a + 8a = 23a$

$$= 23\left(\frac{1}{48}\right) = 0.575$$

$$P(X \le 5) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) = a + 4a + 3a + 7a + 8a + 10a = 33a = 33\left(\frac{1}{48}\right) = 0.69$$

Hence, the smallest value of m for which $P(X \le m) \ge 0.6$ is 5.

EXAMPLE 17.21

A random variable X has the probability function given below:

X	0	1	2
$P(X=x)$	k	$2k$	$3k$

Find (i) k, (ii) $P(X<2)$, $P(X \le 2)$, $P(0<X<2)$, and (iii) the distribution function.

Solution:

(i) Since $P(X=x)$ is a probability density function,

$$\sum(P(X=x)=1$$
$$k+2k+3k=1$$
$$6k=1$$
$$k=\frac{1}{6}$$

Hence, the probability distribution is

X	0	1	2
$P(X=x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

(ii) $P(X<2)=P(X=0)+P(X=1)=\frac{1}{6}+\frac{2}{6}=\frac{1}{2}$

$$P(X\leq 2)=P(X=0)+P(X=1)+P(X=2)=\frac{1}{6}+\frac{2}{6}+\frac{3}{6}=1$$

$$P(0<X<2)=P(X=1)=\frac{1}{3}$$

(iii) Distribution function

X	$P(X=x)$	$F(x)$
0	$\frac{1}{6}$	$\frac{1}{6}$
1	$\frac{2}{6}$	$\frac{1}{6}+\frac{2}{6}=\frac{1}{2}$
2	$\frac{3}{6}$	$\frac{1}{6}+\frac{2}{6}+\frac{3}{6}=1$

EXERCISE 17.7

1. Verify whether the following functions can be considered as probability mass functions:

(i) $P(X=x)=\dfrac{x^2+1}{18}, x=0,1,2,3$

[Ans.: Yes]

(ii) $P(X=x)=\dfrac{x^2-2}{8}, x=1,2,3$

[Ans.: No]

(iii) $P(X=x)=\dfrac{2x+1}{18}, x=0,1,2,3$

[Ans.: No]

2. The probability density function of a random variable X is

X	0	1	2	3	4	5	6
$P(X=x)$	k	$3k$	$5k$	$7k$	$9k$	$11k$	$13k$

Find $P(X<4)$ and $P(3<X\le 6)$.

$\left[\textbf{Ans.:}\ \dfrac{16}{49}, \dfrac{33}{49}\right]$

3. A random variable X has the following probability distribution:

X	1	2	3	4	5	6	7
$P(X=x)$	k	$2k$	$3k$	k^2	k^2+k	$2k^2$	$4k^2$

Find (i) k, (ii) $P(X<5)$, (iii) $P(X>5)$, and (iv) $P(0\le X\le 5)$

$\left[\textbf{Ans.:}\ \dfrac{1}{8}\ \text{(ii)}\ \dfrac{49}{64}\ \text{(iii)}\ \dfrac{3}{32}\ \text{(iv)}\ \dfrac{29}{32}\right]$

4. A discrete random variable X has the following probability distribution:

X	−2	−1	0	1	2	3
$P(X=x)$	0.1	k	0.2	$2k$	0.3	$3k$

Find (i) k, (ii) $P(X\ge 2)$, and (iii) $P(-2<X<2)$.

$\left[\textbf{Ans.:}\ \dfrac{1}{15}\ \text{(ii)}\ \dfrac{1}{2}\ \text{(iii)}\ \dfrac{2}{5}\right]$

5. Given the following probability function of a discrete random variable X:

X	0	1	2	3	4	5	6	7
$P(X=x)$	0	c	$2c$	$2c$	$3c$	c^2	$2c^2$	$7c^2+c$

Find (i) c, (ii) $P(X\ge 6)$, (iii) $P(X<6)$, and (iv) Find k if $P(X\le k)>\dfrac{1}{2}$, where k is a positive integer.

[Ans.: (i) 0.1 (ii) 0.19 (iii) 0.81 (iv) 4]

6. A random variable X assumes four values with probabilities $\dfrac{1+3x}{4}, \dfrac{1-x}{4}, \dfrac{1+2x}{4}$ and $\dfrac{1-4x}{4}$. For what value of x do these values represent the probability distribution of X?

$\left[\textbf{Ans.:}\ -\dfrac{1}{3}\le X\le \dfrac{1}{4}\right]$

7. Let X denote the number of heads in a single toss of 4 fair coins.

Determine (i) $P(X<2)$, and (ii) $P(1<X\le 3)$.

$\left[\textbf{Ans.:}\ \text{(i)}\ \dfrac{5}{16}\ \text{(ii)}\ \dfrac{5}{8}\right]$

8. If 3 cars are selected from a lot of 6 cars containing 2 defective cars, find the probability distribution of the number of defective cars.

Ans.:

X	0	1	2
$P(X=x)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{2}{5}$

9. Five defective bolts are accidentally mixed with 20 good ones. Find the probability distribution of the number of defective bolts, if four bolts are drawn at random from this lot.

Ans.:

X	0	1	2	3	4
$P(X=x)$	$\frac{969}{2530}$	$\frac{1140}{2530}$	$\frac{380}{2530}$	$\frac{40}{2530}$	$\frac{1}{2530}$

10. Two dice are rolled at once. Find the probability distribution of the sum of the numbers on them.

Ans.:

X	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

11. A random variable X takes three values 0, 1, and 2 with probabilities $\frac{1}{3}, \frac{1}{6}$, and $\frac{1}{2}$ respectively. Obtain the distribution function of X.

Ans.: $F(0)=\frac{1}{3}, F(1)=\frac{1}{2}, F(2)=1$

12. A random variable X has the following probability function:

x	0	1	2	3	4
$P(X=x)$	k	$3k$	$5k$	$7k$	$9k$

Find (i) the value of k, (ii) $P(X < 3)$, $P(X \geq 3)$, $P(0 < X < 4)$, and (iii) distribution function of X.

Ans.: (i) $\frac{1}{25}$, (ii) $\frac{9}{25}, \frac{16}{25}, \frac{3}{5}$

(iii) $F(0)=\frac{1}{25}, F(1)=\frac{4}{25}. F(2)=\frac{9}{25}$,

$F(3)=\frac{16}{25}, F(4)=1$

13. A random variable X has the probability function

X	−2	−1	0	1	2	3
$P(X=x)$	0.1	k	0.2	$2k$	0.3	k

Find (i) k, (ii) $P(X \leq 1)$,

(iii) $P(-2 < X < 1)$, and (iv) obtain the distribution function of X.

[Ans.: (i) 0.1 (ii) 0.6 (iii) 0.3]

14. The following is the distribution function $F(x)$ of a discrete random variable X:

X	−3	−2	−1	0	1	2	3
$F(x)$	0.08	0.2	0.4	0.65	0.8	0.9	1

Find (i) the probability distribution of X, (ii) $P(-2 \leq X \leq 1)$, and (iii) $P(X \geq 1)$.

Ans.: (i)

X	−3	−2	−1	0	1	2	3
$P(X=x)$	0.08	0.12	0.2	0.25	0.15	0.1	0.1

(ii) 0.72 (iii) 0.35

17.9 MEASURES OF CENTRAL TENDENCY FOR DISCRETE PROBABILITY DISTRIBUTION

The behaviour of a random variable is completely characterized by the distribution function $F(x)$ or density function $p(x)$. Instead of a function, a more compact description can be made by a single number such as mean, median, mode, variance, and standard deviation known as measures of central tendency of the random variable X.

1. Mean The mean or average value (μ) of the probability distribution of a discrete random variable X is called as expectation and is denoted by $E(X)$.

$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \; p(x_i) = \sum x \; p(x)$$

where $p(x)$ is the probability density function of the discrete random variable X. Expectation of any function $\phi(x)$ of a random variable X is given by

$$E[\phi(x)] = \sum_{i=1}^{\infty} \phi(x_i)\, p(x_i) = \sum \phi(x)\, p(x)$$

Some important results on expectation

(i) $E(k) = k$
(ii) $E(X + k) = E(X) + k$
(iii) $E(aX \pm b) = aE(X) \pm b$
(iv) $E(X + Y) = E(X) + E(Y)$, provided $E(X)$ and $E(Y)$ exists.
(v) $E(XY) = E(X)\, E(Y)$ if X and Y are two independent random variables.

2. Variance Variance characterizes the variability in the distributions since two distributions with same mean can still have different dispersion of data about their means. Variance of the probability distribution of a discrete random variable X is given by

$$\begin{aligned} \text{Var}(X) = \sigma^2 &= E(X-\mu)^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - E(2X\mu) + E(\mu^2) \\ &= E(X^2) - 2\mu\, E(X) + \mu^2 \qquad [\because E(\text{constant}) = (\text{constant})] \\ &= E(X^2) - 2\mu\mu + \mu^2 = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2 \end{aligned}$$

Some important results on variance

(i) Var $(k) = 0$
(ii) Var $(kX) = k^2$ Var (X)
(iii) Var $(X + k) =$ Var (X)
(iv) Var $(aX + b) = a^2$ Var(X)

3. Standard Deviation Standard deviation is the positive square root of the variance.

$$\text{SD} = \sigma = \sqrt{\sum_{i=1}^{\infty} x_i^2 \; p(x_i) - \mu^2} = \sqrt{E(X^2) - \mu^2} = \sqrt{E(X^2) - [E(X)]^2}$$

EXAMPLE 17.22

A random variable X has the following distribution:

X	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Find (i) mean, (ii) variance, and (iii) $P(1 < X < 6)$.

Solution:

(i) $\text{Mean} = \mu = \sum xp(x) = 1\left(\frac{1}{36}\right) + 2\left(\frac{3}{36}\right) + 3\left(\frac{5}{36}\right) + 4\left(\frac{7}{36}\right) + 5\left(\frac{9}{36}\right) + 6\left(\frac{11}{36}\right) = \frac{161}{36} = 4.47$

(ii) $\text{Variance} = \sigma^2 = \sum x^2 p(x) - \mu^2$

$$= 1\left(\frac{1}{36}\right) + 4\left(\frac{3}{36}\right) + 9\left(\frac{5}{36}\right) + 16\left(\frac{7}{36}\right) + 25\left(\frac{9}{36}\right) + 36\left(\frac{11}{36}\right) - (4.47)^2$$

$$= \frac{791}{36} - 19.98 = 1.99$$

(iii) $P(1 < X < 6) = P(X=2) + P(X=3) + P(X=4) + P(X=5) = \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} = \frac{24}{36} = 0.67$

EXAMPLE 17.23

The probability distribution of a random variable X is given below. Find (i) $E(X)$, *(ii)* $\text{Var}(X)$, *(iii)* $E(2X-3)$, *and (iv)* $\text{Var}(2X-3)$

X	−2	−1	0	1	2
$P(X=x)$	0.2	0.1	0.3	0.3	0.1

Solution:

(i) $E(X) = \sum x\, p(x) = -2(0.2) - 1(0.1) + 0 + 1(0.3) + 2(0.1) = 0$

(ii) $\text{Var}(X) = \sum x^2\, p(x) - [E(X)]^2 = 4(0.2) + 1(0.1) + 0 + 1(0.3) + 4(0.1) - 0 = 1.6$

(iii) $E(2X-3) = 2E(X) - 3 = 2(0) - 3 = -3$

(iv) $\text{Var}(2X-3) = (2)^2\, \text{Var}(X) = 4(1.6) = 6.4$

EXAMPLE 17.24

A random variable X has the following probability function:

x	0	1	2	3	4	5	6	7
$p(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2 + k$

(i) *Determine k.* (ii) *Evaluate* $P(X < 6)$, $P(X \geq 6)$, $P(0 < X < 5)$ and $P(0 \leq X \leq 4)$. (iii) *Determine the distribution function of X. (iv) Find the mean. (v) Find the variance.*

Solution:

(i) Since $p(x)$ is a probability mass function,

$$\sum p(x) = 1$$

$$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k - 1 = 1$$

$$(10k - 1)(k + 1) = 0$$

$$k = \frac{1}{10} \text{ or } k = -1$$

$$k = \frac{1}{10} = 0.1 \left[\because \quad p(x) \geq 0, k \neq -1\right]$$

Hence, the probability function is

X	0	1	2	3	4	5	6	7
$P(X = x)$	0	0.1	0.2	0.2	0.3	0.01	0.02	0.17

(ii) $$P(X < 6) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$
$$= 0 + 0.1 + 0.2 + 0.2 + 0.3 + 0.01 = 0.81$$

$$P(X \geq 6) = 1 - P(X < 6) = 1 - 0.81 = 0.19$$

$$P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 0.1 + 0.2 + 0.2 + 0.3 = 0.8$$

$$P(0 \leq X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$
$$= 0 + 0.1 + 0.2 + 0.2 + 0.3 = 0.8$$

(iii) Distribution function of X

x	$p(x)$	$F(x)$
0	0	0
1	0.1	0.1
2	0.2	0.3
3	0.2	0.5
4	0.3	0.8
5	0.01	0.81
6	0.02	0.83
7	0.17	1

(iv) $$\mu = \sum xp(x) = 0 + 1(0.1) + 2(0.2) + 3(0.2) + 4(0.3) + 5(0.01) + 6(0.02) + 7(0.17) = 3.66$$

(v) $$\text{Var}(X) = \sigma^2 = \sum x^2 p(x) - \mu^2 = 0 + 1(0.1) + 4(0.2) + 9(0.2) + 16(0.3) + 25(0.01) + 36(0.02)$$
$$+ 49(0.17) - (3.66)^2 = 3.4044$$

EXERCISE 17.8

1. The probability distribution of a random variable X is given by

X	−2	−1	0	1	2	3
$P(X=x)$	0.1	k	0.2	$2k$	0.3	k

Find k, the mean, and variance.

[**Ans.:** 0.1, 0.8, 2.16]

2. Find the mean and variance of the following distribution:

X	4	5	6	8
$P(X=x)$	0.1	0.3	0.4	0.2

[**Ans.:** 5.9, 1.49]

3. Find the value of k from the following data:

X	0	10	15
$P(X=x)$	$\frac{k-6}{5}$	$\frac{2}{k}$	$\frac{14}{5k}$

Also, find the distribution function and expectation of X.

[**Ans.:** 8,

X	0	10	15
$F(x)$	$\frac{2}{5}$	$\frac{13}{20}$	1

, $\frac{31}{4}$]

4. For the following distribution,

X	−3	−2	−1	0	1	2
$P(X=x)$	0.01	0.1	0.2	0.3	0.2	0.15

Find (i) $P(X \geq 1)$, (ii) $P(X < 0)$, (iii) $E(X)$, and (iv) Var(X)

[**Ans.:** (i) 0.35 (ii) 0.35 (iii) 0.05 (iv) 1.8475]

5. A random variable X has the following probability function:

X	0	1	2	3	4	5	6	7	8
$P(X=x)$	$\frac{k}{45}$	$\frac{k}{15}$	$\frac{k}{9}$	$\frac{k}{5}$	$\frac{2k}{45}$	$\frac{6k}{45}$	$\frac{7k}{45}$	$\frac{8k}{45}$	$\frac{4k}{45}$

Determine (i) k, (ii) mean, (iii) variance, and (iv) SD.

[**Ans.:** (i) 1 (ii) 0.4622 (iii) 4.9971 (iv) 2.24]

6. A fair coin is tossed until a head or five tails appear. Find (i) discrete probability distribution, and (ii) mean of the distribution.

[**Ans.:** (i)

X	1	2	3	4	5
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

(ii) 1.9]

7. Let X denotes the minimum of two numbers that appear when a pair of fair dice is thrown once. Determine (i) probability distribution, (ii) expectation, and (iii) variance.

[**Ans.:** (i)

X	1	2	3	4	5	6
$P(X=x)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

(ii) 2.5278 (iii) 1.9713]

8. For the following probability distribution,

X	−3	−2	−1	0	1	2	3
$P(X=x)$	0.001	0.01	0.1	?	0.1	0.01	0.001

Find (i) missing probability, (ii) mean, and (iii) variance.

[**Ans.:** (i) 0.778 (ii) 0.2 (iii) 0.258]

9. A discrete random variable can take all integer values from 1 to k each with the probability of $\frac{1}{k}$. Show that its mean and variance are $\frac{k+1}{2}$ and $\frac{k^2+1}{2}$ respectively.

$\frac{k^2-1}{12}$

10. An urn contains 6 white and 4 black balls; 3 balls are drawn without replacement. What is the expected number of black balls that will be obtained?

[**Ans.:** $\frac{6}{5}$]

11. A six-faced dice is tossed. If a prime number occurs, Anil wins that number of rupees but if a nonprime number occurs, he loses that number of rupees. Determine whether the game is favourable to the player.

[**Ans.:** The game is favourable to Anil]

12. A man runs an ice-cream parlour at a holiday resort. If the summer is mild, he can sell 2500 cups of ice cream; if it is hot, he can sell 4000 cups; if it is very hot, he can sell 5000 cups. It is known that for any year, the probability of summer to be mild is $\frac{1}{7}$ and to be hot is $\frac{4}{7}$. A cup of ice cream costs ₹2 and is sold for ₹3.50. What is his expected profit?

[**Ans.:** ₹ 6107.14]

13. A player tosses two fair coins. He wins ₹ 1 or ₹ 2 as 1 tail or 2 head appears. On the other hand, he loses ₹ 5 if no head appears. Find the expected gain or loss of the player.

[**Ans.:** Loss of ₹ 0.25]

14. A bag contains 2 white balls and 3 black balls. Four persons A, B, C, D in the order named each draws one ball and does not replace it. The first to draw a white ball receives ₹ 20. Determine their expectations.

[**Ans.:** ₹ 8, ₹ 6, ₹ 4, ₹ 2]

17.10 CONTINUOUS PROBABILITY DISTRIBUTION

For a continuous random variable X, the function $f(x)$ is called the probability density function if it satisfies the following conditions:

(i) $f(x) \geq 0, \quad -\infty < x < \infty$

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

(iii) $P(a < x < b) = \int_{a}^{b} f(x)\,dx$

17.11 CONTINUOUS DISTRIBUTION FUNCTION

If X is a continuous random variable having the probability density function $f(x)$ then the function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx, \quad -\infty < x < \infty$$

is called the distribution function or cumulative distribution function of the random variable X.

Properties of Distribution Function

(i) $F(-\infty) = 0$

(ii) $F(\infty) = 1$

(iii) $0 \leq F(x) \leq 1, \quad -\infty < x < \infty$

(iv) $P(a < X < b) = F(b) - F(a)$

(v) $F'(x) = \frac{d}{dx}F(x) = f(x), \quad f(x) \geq 0$

EXAMPLE 17.25

Find the constant k such that the function

$$f(x) = kx^2, \quad 0 < x < 3$$
$$= 0, \quad \text{otherwise}$$

is a probability density function and compute (i) $P(1 < X < 2)$, (ii) $P(X < 2)$, and (iii) $P(X \geq 2)$.

Solution: Since $f(x)$ is a probability density function,

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

$$\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{3} f(x)\,dx + \int_{3}^{\infty} f(x)\,dx = 1$$

$$0 + \int_{0}^{3} kx^2\,dx + 0 = 1$$

$$k\left|\frac{x^3}{3}\right|_0^3 = 1$$

$$\frac{k}{3}(27 - 0) = 1$$

$$9k = 1$$

$$k = \frac{1}{9}$$

Hence, $f(x) = \frac{1}{9}x^2, \quad 0 < x < 3$

$= 0, \quad$ otherwise

(i) $P(1 < X < 2) = \int_1^2 f(x)\,dx = \int_1^2 \frac{1}{9}x^2\,dx = \frac{1}{9}\left|\frac{x^3}{3}\right|_1^2 = \frac{1}{27}(8-1) = \frac{7}{27}$

(ii) $P(X < 2) = \int_{-\infty}^{2} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_0^2 f(x)\,dx = 0 + \int_0^2 \frac{1}{9}x^2\,dx = \frac{1}{9}\int_0^2 x^2\,dx$

$$= \frac{1}{9}\left|\frac{x^3}{3}\right|_0^2 = \frac{1}{27}(8-0) = \frac{8}{27}$$

(iii) $P(X \geq 2) = 1 - P(X < 2) = 1 - \frac{8}{27} = \frac{19}{27}$

EXAMPLE 17.26

Let X be a continuous random variable with pdf

$$f(x) = kx\,(1 - x),\ 0 \leq x \leq 1$$

Find k and determine a number b such that $P(X \leq b) = P(X \geq b)$.

Solution: Since $f(x)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x) = 1$$

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

$$0 + \int_{0}^{1} kx(1-x)\,dx + 0 = 1$$

$$k\int_{0}^{1} (x - x^2)\,dx = 1$$

$$k\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^1 = 1$$

$$k\left[\left(\frac{1}{2} - \frac{1}{3}\right) - (0-0)\right] = 1$$

$$k\left(\frac{1}{6}\right) = 1$$

$$k = 6$$

Hence, $f(x) = 6(x - x^2), \quad 0 \le x \le 1$

Since total probability is 1 and $P(X \le b) = P(X \ge b)$,

$$P(X \le b) = \frac{1}{2}$$

$$\int_0^b f(x)\,dx = \frac{1}{2}$$

$$6\int_0^b (x - x^2)\,dx = \frac{1}{2}$$

$$6\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^b = \frac{1}{2}$$

$$\frac{b^2}{2} - \frac{b^3}{3} = \frac{1}{12}$$

$$6b^2 - 4b^3 = 1$$

$$4b^3 - 6b^2 + 1 = 0$$

$$(2b-1)(2b^2 - 2b - 1) = 0$$

$$b = \frac{1}{2} \text{ or } b = \frac{1 \pm \sqrt{3}}{2}$$

Since b lies in (0, 1), $b = \frac{1}{2}$

EXAMPLE 17.27 *Verify that the function $F(x)$ is a distribution function.*

$$F(x) = 0, \qquad x < 0$$

$$= 1 - e^{-\frac{x}{4}}, \qquad x \ge 0$$

Also, find the probabilities $P(X \le 4)$, $P(X \ge 8)$, $P(4 \le X \le 8)$.

Solution: For the function $F(x)$,

(i) $F(-\infty) = 0$

(ii) $F(\infty) = 1 - e^{-\infty} = 1 - 0 = 1$

(iii) $0 \le F(x) \le 1, \quad -\infty < x < \infty$

If $f(x)$ is the corresponding probability density function,

$$f(x) = F'(x) = 0, \quad x < 0$$

$$= \frac{1}{4}e^{-\frac{x}{4}}, \quad x \ge 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx = 0 + \int_{0}^{\infty} \frac{1}{4}e^{-\frac{x}{4}}dx = \frac{1}{4}\left|\frac{e^{-\frac{x}{4}}}{-\frac{1}{4}}\right|_0^{\infty} = -\left|e^{-\frac{x}{4}}\right|_0^{\infty} = -(0-1) = 1$$

Hence, $F(x)$ is a distribution function.

$$P(X \le 4) = F(4) = 1 - e^{-1} = 1 - \frac{1}{e} = \frac{e-1}{e}$$

$$P(X \ge 8) = 1 - P(X \le 8) = 1 - F(8) = 1 - (1 - e^{-2}) = e^{-2} = \frac{1}{e^2}$$

$$P(4 \le X \le 8) = F(8) - F(4) = (1 - e^{-2}) - (1 - e^{-1}) = e^{-1} - e^{-2} = \frac{1}{e} - \frac{1}{e^2} = \frac{e-1}{e^2}$$

EXAMPLE 17.28

The probability density function of a continuous random variable X is given by

$$f(x) = \begin{cases} ax & 0 \le x \le 1 \\ a & 1 \le x \le 2 \\ 3a - ax & 2 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

(i) Find the value of a, and (ii) find the cdf of X.

Solution: (i) Since $f(x)$ is a probability density function,

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

$$\int_{-\infty}^{\infty} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{2} f(x)\,dx + \int_{2}^{3} f(x)\,dx = 1$$

$$0 + \int_{0}^{1} ax\,dx + \int_{1}^{2} a\,dx + \int_{2}^{3} (3a - ax)\,dx = 1$$

$$a\left|\frac{x^2}{2}\right|_0^1 + a\left|x\right|_1^2 + \left|3ax - \frac{ax^2}{2}\right|_2^3 = 1$$

$$a\left(\frac{1}{2} - 0\right) + a(2-1) + \left[\left(9a - \frac{9a}{2}\right) - (6a - 2a)\right] = 1$$

$$\frac{1}{2}a + a + \frac{9a}{2} - 4a = 1$$

$$2a = 1$$

$$a = \frac{1}{2}$$

(ii) $F(x)=\int_{-\infty}^{x} f(x)\,dx$

For $0 \le x \le 1$,

$$F(x)=\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{x} f(x)\,dx = 0+\int_{0}^{x} ax\,dx = a\left|\frac{x^2}{2}\right|_0^x = \frac{ax^2}{2}$$

For $1 \le x \le 2$,

$$F(x)=\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{1} f(x)\,dx+\int_{1}^{x} f(x)\,dx = 0+\int_{0}^{1} ax\,dx+\int_{1}^{x} a\,dx$$

$$= a\left|\frac{x^2}{2}\right|_0^1 + a|x|_1^x = a\left(\frac{1}{2}-0\right)+a(x-1)=\frac{a}{2}+ax-a=ax-\frac{a}{2}$$

For $2 \le x \le 3$,

$$F(x)=\int_{-\infty}^{0} f(x)\,dx+\int_{0}^{1} f(x)\,dx+\int_{1}^{2} f(x)\,dx+\int_{2}^{x} f(x)\,dx = 0+\int_{0}^{1} ax\,dx+\int_{1}^{2} a\,dx+\int_{2}^{x} (3a-ax)\,dx$$

$$= a\left|\frac{x^2}{2}\right|_0^1 + a|x|_1^2 + \left|3ax-\frac{ax^2}{2}\right|_2^x = a\left(\frac{1}{2}-0\right)+a(2-1)+\left[\left(3ax-\frac{ax^2}{2}\right)-(6a-2a)\right]$$

$$=\frac{a}{2}+a+3ax-\frac{ax^2}{2}-4a=3ax-\frac{ax^2}{2}-\frac{5a}{2}$$

Hence, $F(x)=\dfrac{ax^2}{2}, \qquad 0 \le x \le 1$

$$= ax-\frac{a}{2}, \qquad 1 \le x \le 2$$

$$= 3ax-\frac{ax^2}{2}-\frac{5a}{2}, \qquad 2 \le x \le 3$$

EXERCISE 17.9

1. Verify whether the following functions are probability density functions:

(i) $f(x)=k\,e^{-kx}, \qquad x \ge 0, k>0$

(ii) $f(x)=\dfrac{1}{2}e^{-|x|}, \qquad -\infty < x < \infty$

(iii) $f(x)=\dfrac{2}{9}x\left(2-\dfrac{x}{2}\right), \qquad 0 \le x \le 3$

[**Ans.:** (i) Yes (ii) Yes (iii) Yes]

2. Find the value of k if the following are probability density functions:

(i) $f(x)=k(1+x), \qquad 2 \le x \le 5$

(ii) $f(x)=k(x-x^2), \qquad 0 \le x \le 1$

(iii) $f(x)=kx\,e^{-4x^2}, \qquad 0 \le x \le \infty$

(iv) $f(x)=kx\,e^{-\frac{x^2}{4}}, \qquad 0 \le x \le \infty$

[**Ans.:** (i) $\dfrac{2}{27}$ (ii) 6 (iii) 8 (iv) $\dfrac{1}{2}$]

3. A function is defined as

$$f(x)=\begin{cases} 0, & x<2 \\ \dfrac{2x+3}{18}, & 2 \le x \le 4 \\ 0, & x>4 \end{cases}$$

Show that $f(x)$ is a probability density function and find $P(2<X<3)$.

[**Ans.:** $\dfrac{4}{9}$]

4. Let X be a continuous random variable with probability distribution

$$f(x)=\begin{cases}\frac{x}{6}+k, & 0\le x\le 3\\ 0, & \text{otherwise}\end{cases}$$

Find k, and $P(1\le X\le 2)$.

[**Ans.:** $1, \frac{1}{3}$]

5. Find the value of k such that $f(x)$ is a probability density function. Find also, $P(X\le 1.5)$.

$$f(x)=\begin{cases}kx, & 0\le x\le 1\\ k, & 1\le x\le 2\\ k(3-x), & 2\le x\le 3\end{cases}$$

[**Ans.:** $\frac{1}{2}, \frac{1}{2}$]

6. If X is a continuous random variable whose probability density function is given by

$$f(x)=k(4x-2x^2), \quad 0<x<2$$
$$=0, \quad \text{otherwise}$$

Find (i) the value of k, and (ii) $P(X>1)$.

[**Ans.:** (i) $\frac{3}{8}$ (ii) $\frac{1}{2}$]

7. If a random variable has the probability density function

$$f(x)=k(x^2-1), \quad -1\le x\le 3$$
$$=0, \quad \text{otherwise}$$

Find (i) the value of k, and (ii) $P\left(\frac{1}{2}\le X\le\frac{5}{2}\right)$.

[**Ans.:** (i) $\frac{3}{28}$ (ii) $\frac{19}{56}$]

8. The probability density function is

$$f(x)=k(3x^2-1), \quad -1\le x\le 2$$
$$=0, \quad \text{otherwise}$$

Find (i) the value of k, and (ii) $P(-1\le X\le 0)$.

[**Ans.:** (i) $\frac{1}{6}$ (ii) 0]

9. Is the function defined by

$$f(x)=0, \quad x<2$$
$$=\frac{1}{18}(2x+3), \quad 2\le x\le 4$$
$$=0, \quad x>4$$

a probability density function? Find the probability that a variate having $f(x)$ as density function will fall in the interval $2\le X\le 3$.

[**Ans.:** Yes, $\frac{4}{9}$]

10. A random variable X gives measurements x between 0 and 1 with a probability function

$$f(x)=12x^3-21x^2+10x, \quad 0\le x\le 1$$
$$=0, \quad \text{otherwise}$$

(i) Find $P\left(X\le\frac{1}{2}\right)$ and $P\left(X>\frac{1}{2}\right)$.

(ii) Find a number k such that

$$P(X\le k)=\frac{1}{2}.$$

[**Ans.:** (i) $\frac{7}{16}$ (ii) 0.452]

11. The distribution function of a random variable X is given by

$$F(x)=\begin{cases}1-e^{-x^2}, & x>0\\ 0, & \text{otherwise}\end{cases}$$

Find the probability density function.

[**Ans.:** $f(x)=2xe^{-x^2}, \quad x>0$
$=0, \quad \text{otherwise}$]

12. The cdf of a continuous random variable X is given by

$$F(x)=\begin{cases}0, & x<0\\ x^2, & 0\le x\le 1\\ 1, & x>1\end{cases}$$

Find the pdf and $P\left(\frac{1}{2}\le X\le\frac{4}{5}\right)$.

[**Ans.:** 0.195]

13. Find the distribution function corresponding to the following probability density functions:

(i) $f(x)=\begin{cases}\frac{1}{2}x^2e^{-x}, & 0\le x<\infty\\ 0, & \text{otherwise}\end{cases}$

(ii) $f(x)=x, \quad 0\le x\le 1$
$=2-x, \quad 1\le x\le 2$
$=0, \quad \text{otherwise}$

(iii) $f(x)=\lambda(x-1)^4, \quad 1\le x\le 3, \lambda>0$
$=0, \quad \text{otherwise}$

Ans.: (i) $F(x)=\begin{cases}1-e^{-x}\left(1+x+\dfrac{x^2}{2}\right), & x\geq 0\\ 0, & \text{otherwise}\end{cases}$

(ii) $F(x)=\begin{cases}0, & x<0\\ \dfrac{x^2}{2}, & 0\leq x\leq 1\\ 2x-0.5x^2-1, & 1\leq x\leq 2\\ 1, & x>2\end{cases}$

(iii) $\lambda=\dfrac{5}{32}$, $F(x)=\begin{cases}0, & x\leq 1\\ \dfrac{5}{32}(x-1)^4, & 1\leq x\leq 3\\ 1, & x\geq 3\end{cases}$

14. A continuous random variable X has the following probability density function

$$f(x)=\frac{a}{x^5},\quad 2\leq x\leq 10$$

Determine the constant a, distribution function of X, and find the probability of the event $4\leq X\leq 7$.

Ans.: $\dfrac{2500}{39}$, $F(x)=\dfrac{625}{39}\left(\dfrac{1}{16}-\dfrac{1}{x^4}\right)$, 0.056

17.12 MEASURES OF CENTRAL TENDENCY FOR CONTINUOUS PROBABILITY DISTRIBUTION

1. Mean The mean or average value (μ) of the probability distribution of a continuous random variable X is called the expectation and is denoted by $E(X)$.

$$\mu=E(X)=\int_{-\infty}^{\infty} x\,f(x)\,dx$$

where $f(x)$ is the probability density function of the continuous random variable.

Expectation of any function $\phi(x)$ of a continuous random variable X is given by

$$E[\phi(x)]=\int_{-\infty}^{\infty}\phi(x)\,f(x)\,dx$$

2. Median The median is the point which divides the entire distribution into two equal parts. Thus, if a continuous random variable X is defined from a to b and M is the median,

$$\int_a^M f(x)\,dx=\int_M^b f(x)\,dx=\frac{1}{2}$$

By solving any one of the equation, the median is obtained.

3. Mode The mode is value of x for which $f(x)$ is maximum. Mode is given by

$$f'(x)=0 \text{ and } f''(x)<0 \text{ for } a<x<b$$

4. Variance The variance of the probability distribution of a continuous random variable X is given by

$$\text{Var}(X)=\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2\,f(x)\,dx=\int_{-\infty}^{\infty}x^2\,f(x)\,dx-\mu^2$$

5. Standard Deviation The standard deviation of the probability distribution of a continuous random variable X is given by

$$\text{SD} = \sqrt{\text{Var}(X)} = \sigma$$

EXAMPLE 17.29

A continuous random variable has the probability density function

$$f(x) = kxe^{-\lambda x}, \quad x \geq 0, \lambda > 0$$
$$= 0, \quad \text{otherwise}$$

Determine (i) k, (ii) mean, and (iii) variance.

Solution: Since $f(x)$ is a probability density function,

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

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

$$0 + \int_{0}^{\infty} k\,x\,e^{-\lambda x}\,dx = 1$$

$$k\int_{0}^{\infty} x\,e^{-\lambda x}\,dx = 1$$

$$k\left|x\frac{e^{-\lambda x}}{-\lambda} - 1\left(\frac{e^{-\lambda x}}{\lambda^2}\right)\right|_0^{\infty} = 1$$

$$k\left[(0-0) - \left(0 - \frac{1}{\lambda^2}\right)\right] = 1$$

$$k = \lambda^2$$

Hence, $f(x) = \lambda^2 x e^{-\lambda x}, \quad x \geq 0, \lambda > 0$

$$= 0, \quad \text{otherwise}$$

(ii) $$\text{Mean} = \mu = \int_{-\infty}^{\infty} x\,f(x)\,dx = \int_{-\infty}^{0} x\,f(x)\,dx + \int_{0}^{\infty} x\,f(x)\,dx = 0 + \int_{0}^{\infty} x\,\lambda^2\,x e^{-\lambda x}\,dx = \lambda^2\int_{0}^{\infty} x^2\,e^{-\lambda x}\,dx$$

$$= \lambda^2\left|x^2\left(\frac{e^{-\lambda x}}{-\lambda}\right) - 2x\left(\frac{e^{-\lambda x}}{\lambda^2}\right) + 2\left(\frac{e^{-\lambda x}}{-\lambda^3}\right)\right|_0^{\infty} = \lambda^2\left[(0-0+0) - \left(0-0-\frac{2}{\lambda^3}\right)\right] = \frac{2}{\lambda}$$

(iii) $$\text{Variance} = \sigma^2 = \int_{-\infty}^{\infty} x^2\,f(x)\,dx - \mu^2 = \int_{-\infty}^{0} x^2\,f(x)\,dx + \int_{0}^{\infty} x^2\,f(x)\,dx - \mu^2$$

$$= 0 + \int_{0}^{\infty} x^2\,\lambda^2\,x e^{-\lambda x}\,dx - \left(\frac{2}{\lambda}\right)^2 = \lambda^2\int_{0}^{\infty} x^3\,e^{-\lambda x}\,dx - \frac{4}{\lambda^2}$$

$$= \lambda^2\left|x^3\left(\frac{e^{-\lambda x}}{-\lambda x}\right) - 3x^2\left(\frac{e^{-\lambda x}}{\lambda^2}\right) + 6x\left(\frac{e^{-\lambda x}}{-\lambda^3}\right) - 6\left(\frac{e^{-\lambda x}}{\lambda^4}\right)\right|_0^{\infty} - \frac{4}{\lambda^2}$$

$$= \lambda^2\left[(0-0+0-0)-\left(0-0+0-\frac{6}{\lambda^4}\right)\right]-\frac{4}{\lambda^2} = \frac{6}{\lambda^2}-\frac{4}{\lambda^2} = \frac{2}{\lambda^2}$$

EXAMPLE 17.30 *A continuous random variable X has the pdf defined by f (x) = A + Bx, $0 \le x \le 1$. If the mean of the distribution is $\frac{1}{3}$, find A and B.*

Solution: Since $f(x)$ is a probability density function,

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

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

$$0 + \int_{0}^{1} (A + Bx)\,dx + 0 = 1$$

$$\left|Ax + \frac{Bx^2}{2}\right|_0^1 = 1$$

$$A + \frac{B}{2} = 1$$

$$2A + B = 2 \qquad \ldots(1)$$

Also,

$$\mu = \frac{1}{3}$$

$$\int_{-\infty}^{\infty} x\,f(x)\,dx = \frac{1}{3}$$

$$\int_{-\infty}^{0} x\,f(x)\,dx + \int_{0}^{1} x\,f(x)\,dx + \int_{1}^{\infty} x\,f(x)\,dx = \frac{1}{3}$$

$$0 + \int_{0}^{1} x\,(A + Bx)\,dx = \frac{1}{3}$$

$$\int_{0}^{1} (Ax + Bx^2)\,dx = \frac{1}{3}$$

$$\left|\frac{Ax^2}{2} + \frac{Bx^3}{3}\right|_0^1 = \frac{1}{3}$$

$$\frac{A}{2} + \frac{B}{3} = \frac{1}{3}$$

$$3A + 2B = 2 \qquad \ldots(2)$$

Solving Eqs (1) and (2),

$$A = 2, \qquad B = -2$$

EXAMPLE 17.31

The probability density function of a random variable X is

$$f(x) = \frac{1}{2}\sin x, \quad 0 \le x \le \pi$$
$$= 0, \qquad \text{otherwise}$$

Find the mean, mode, and median of the distribution and also, find the probability between 0 and $\frac{\pi}{2}$.

Solution:

(i) $\mu = \int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} x f(x)\,dx + \int_{0}^{\pi} x f(x)\,dx + \int_{\pi}^{\infty} x f(x)\,dx = 0 + \int_{0}^{\pi} x\left(\frac{1}{2}\sin x\right)dx + 0$

$= \frac{1}{2}\int_{0}^{\pi} x \sin x\,dx = \frac{1}{2}\left|-x\cos x + \sin x\right|_0^{\pi} = \frac{\pi}{2}$

(ii) Mode is the value of x for which $f(x)$ is maximum. For $f(x)$ to be maximum, $f'(x) = 0$ and $f''(x) < 0$.

$f'(x) = 0$

$\cos x = 0$

$x = \frac{\pi}{2}$

$f''(x) = -\frac{1}{2}\sin x$

At $x = \frac{\pi}{2}$, $f''(x) = -\frac{1}{2} < 0$

Hence, $f(x)$ is maximum of $x = \frac{\pi}{2}$.

Mode $= \frac{\pi}{2}$

(iii) $\int_{a}^{M} f(x)\,dx = \int_{M}^{b} f(x)\,dx = \frac{1}{2}$

$\int_{0}^{M} \frac{1}{2}\sin x\,dx = \int_{M}^{\pi} \frac{1}{2}\sin x\,dx = \frac{1}{2}$

$\int_{0}^{M} \frac{1}{2}\sin x\,dx = \frac{1}{2}$

$-\frac{1}{2}\left|\cos x\right|_0^{M} = \frac{1}{2}$

$-\frac{1}{2}(\cos M - 1) = \frac{1}{2}$

$1 - \cos M = 1$

$\cos M = 0$

$$M = \frac{\pi}{2}$$

Hence, median $M = \frac{\pi}{2}$

(iv) $P\left(0 < X < \frac{\pi}{2}\right) = \int_0^{\frac{\pi}{2}} f(x)\,dx = \int_0^{\frac{\pi}{2}} \frac{1}{2}\sin x\,dx = -\frac{1}{2}|\cos x|_0^{\frac{\pi}{2}} = -\frac{1}{2}(0-1) = \frac{1}{2}$

EXAMPLE 17.32 *A continuous random variable X has the distribution function*

$$F(x) = 0, \quad x \le 1$$
$$= k(x-1)^4, \quad 1 < x \le 3$$
$$= 1, \quad x > 3$$

Determine (i) f(x) (ii) k and (iii) mean.

Solution:

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

$$f(x) = 0, \quad x \le 1$$
$$= 4k(x-1)^3, \quad 1 < x \le 3$$
$$= 0, \quad x > 3$$

(ii) Since $f(x)$ is a probability density function,

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

$$\int_{-\infty}^{1} f(x)\,dx + \int_1^3 f(x)\,dx + \int_3^{\infty} f(x)\,dx = 1$$

$$0 + \int_1^3 4k(x-1)^3\,dx + 0 = 1$$

$$4k\left|\frac{(x-1)^4}{4}\right|_1^3 = 1$$

$$k(16-0) = 1$$

$$k = \frac{1}{16}$$

Hence, $f(x) = 0, \quad x \le 1$

$$= \frac{1}{4}(x-1)^3, \quad 1 < x \le 3$$
$$= 0, \quad x > 3$$

(iii) $\mu = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{1} x f(x)\,dx + \int_1^3 x f(x)\,dx + \int_3^{\infty} x f(x)\,dx$

$$= 0 + \int_1^3 x \cdot \frac{1}{4}(x-1)^3\,dx + 0 = \frac{1}{4}\int_1^3 x(x-1)^3\,dx$$

$$= \frac{1}{4}\int_0^2 (t+1)\, t^3 \, dt \qquad \begin{bmatrix} \text{Putting } x-1=t \\ \text{When } x=1,\ t=0 \\ \text{When } x=3,\ t=2 \end{bmatrix}$$

$$= \frac{1}{4}\int_0^2 (t^4+t^3)\, dt = \frac{1}{4}\left|\frac{t^5}{5}+\frac{t^4}{4}\right|_0^2 = \frac{1}{4}\left[\left(\frac{2^5}{5}+\frac{2^4}{4}\right)-(0)\right] = 2.6$$

EXERCISE 17.10

1. If the probability density function is given by

$$f(x) = kx^2(1-x^3), \quad 0 \le x \le 1$$
$$= 0, \quad \text{otherwise}$$

Find

(i) k, (ii) $P\left(0 < X < \frac{1}{2}\right)$, (iii) $\bar{X}$, and (iv) σ^2.

$$\left[\textbf{Ans.:}\ (i)\ 6\ (ii)\ \frac{15}{64}\ (iii)\ \frac{9}{14}\ (iv)\ \frac{9}{245}\right]$$

2. If the probability density function of a random variable is given by

$$f(x) = kx, \quad 0 \le x \le 2$$
$$= 2k, \quad 2 \le x \le 4$$
$$= 6k - kx, \quad 4 \le x \le 6$$

Find (i) k, (ii) $P(1 \le X \le 3)$, and (iii) $\bar{X}$.

$$\left[\textbf{Ans.:}\ (i)\ \frac{1}{2}\ (ii)\ \frac{1}{3}\ (iii)\ \frac{383}{36}\right]$$

3. If the probability density of a random variable is given by

$$f(x) = kxe^{-\frac{x}{3}}, \quad x > 0$$
$$= 0, \quad x \le 0$$

Find (i) k, (ii) $\bar{X}$, and (iii) σ^2.

$$\left[\textbf{Ans.:}\ (i)\ \frac{1}{9}\ (ii)\ 6\ (iii)\ 18\right]$$

4. A continuous random variable has the probability density function

$$f(x) = 2e^{-2x}, \quad x > 0$$
$$= 0, \quad x \le 0$$

Find (i) $E(X)$, (ii) $E(\bar{X})$, (iii) Var (X), and (iv) SD of X.

$$\left[\textbf{Ans.:}\ (i)\ \frac{1}{2}\ (ii)\ \frac{1}{2}\ (iii)\ \frac{1}{4}\ (iv)\ \frac{1}{2}\right]$$

5. A random variable X has the pdf

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

Determine (i) k, (ii) $P(X \ge 0)$, (iii) mean, and (iv) variance.

$$\left[\textbf{Ans.:}\ (i)\ \frac{1}{\pi}\ (ii)\ \frac{1}{2}\ (iii)\ 0\ (iv)\ \text{does not exist}\right]$$

6. The distribution function of a continuous random variable X is given by $F(x) = 1-(1+x)e^{-x}$, $x \ge 0$. Find (i) pdf, (ii) mean, and (iii) variance.

$$\left[\textbf{Ans.:}\ (i)\ f(x) = xe^{-x},\ x \ge 0\ (ii)\ 2\ (iii)\ 2\right]$$

7. If $f(x)$ is the probability density function of a continuous random variable, find k, mean, and variance.

$$f(x) = kx^2, \quad 0 \le x \le 1$$
$$= (2-x)^2, \quad 1 \le x \le 2$$

$$\left[\textbf{Ans.:}\ 2, \frac{11}{12}, 0.626\right]$$

8. A continuous random variable X has the probability density function given by

$$f(x) = 2ax + b, \quad 0 \le x \le 2$$
$$= 0, \quad \text{otherwise}$$

If the mean of the distribution is 3, find the constants a and b.

$$\left[\textbf{Ans.:}\ \frac{3}{2}, -\frac{5}{2}\right]$$

9. If X is a continuous random variable with probability density function given by

$$f(x) = k(x - x^3), \quad 0 \le x \le 1$$
$$= 0, \quad \text{otherwise}$$

Find (i) k, (ii) mean, (iii) variance, and (iv) median.

$$\left[\textbf{Ans.:}\ \text{(i)}\ \frac{1}{2}\ \text{(ii)}\ 0.06\ \text{(iii)}\ 0.04\ \text{(iv)}\ 2\right]$$

10. The probability density function of a random variable is given by

$$f(x) = 0, \quad x < 2$$
$$= \frac{2x+3}{18}, \quad 2 \le x \le 4$$
$$= 0, \quad x > 4$$

Find the mean and variance.

$$\left[\textbf{Ans.:}\ \text{(i)}\ \frac{83}{27}, 0.33\right]$$

11. A continuous random variable X has the probability density function

$$f(x) = x^3, \quad 0 \le x \le 1$$
$$= (2 - x)^3, \quad 1 \le x \le 2$$
$$= 0, \quad \text{otherwise}$$

Find $P(0.5 \le X \le 1.5)$ and mean of the distribution.

$$\left[\textbf{Ans.:}\ \frac{15}{32}, \frac{1}{2}\right]$$

12. The probability density function of a continuous random variable X is given by

$$f(x) = kx\,(2 - x), \quad 0 \le x \le 2$$

Find k, mean, and variance.

$$\left[\textbf{Ans.:}\ \frac{3}{4}, 1, \frac{1}{5}\right]$$

17.13 BINOMIAL DISTRIBUTION

Consider n independent trials of a random experiments which results in either success or failure. Let p be the probability of success remaining constant every time and $q = 1 - p$ be the probability of failure. The probability of x successes and $n - x$ failures is given by $p^x q^{n-x}$ (multiplication theorem of probability). But these x successes and $n - x$ failures can occur in any of the ${}^nC_x$ ways in each of which the probability is same. Hence, the probability of x successes is ${}^nC_x\, p^x q^{n-x}$.

$$P(X = x) = {}^nC_x\, p^x\, q^{n-x}, \quad x = 0, 1, 2, \ldots, n, \text{ where } p + q = 1$$

A random variable X is said to follow the binomial distribution if the probability of x is given by

$$P(X = x) = p(x) = {}^nC_x\, p^x\, q^{n-x}, \quad x = 0, 1, 2, \ldots, n \text{ and } q = 1 - p.$$

The two constants n and p are called the parameters of the distribution.

17.13.1 Examples of Binomial Distribution

(i) Number of defective bolts in a box containing n bolts.
(ii) Number of post-graduates in a group of n people.
(iii) Number of oil wells yielding natural gas in a group of n wells test drilled.
(iv) Number of machines lying idle in a factory having n machines.

17.13.2 Conditions for Binomial Distribution

The binomial distribution holds under the following conditions:

(i) The number of trials n is finite.
(ii) There are only two possible outcomes, success or failure.
(iii) The trials are independent of each other.
(iv) The probability of success p is constant for each trial.

17.13.3 Constants of the Binomial Distribution

1. Mean of the Binomial Distribution

$$E(X) = \sum x\, p(x) = \sum_{x=1}^{n} x\, {}^nC_x\, p^x\, q^{n-x}$$

$$= 1 \cdot {}^nC_1\, pq^{n-1} + 2 \cdot {}^nC_2\, p^2\, q^{n-2} + \cdots + n\, p^n$$

$$= np\,[q^{n-1} + {}^{(n-1)}C_1\, q^{n-2}\, p + {}^{(n-1)}C_2\, q^{n-3}\, p^2 + \cdots + p^{n-1}]$$

$$= np\,(q+p)^{n-1} = np \qquad [\because\ p+q=1]$$

2. Variance of the Binomial Distribution

$$\text{Var}(X) = E(X^2) - \mu^2$$

$$= \sum_{x=1}^{n} x^2\, p(x) - \mu^2 = \sum_{x=0}^{n} x^2\, {}^nC_x\, p^x\, q^{n-x} - \mu^2$$

$$= \sum_{x=1}^{n} [x + x(x-1)]\, {}^nC_x\, p^x\, q^{n-x} - \mu^2 = \sum_{x=1}^{n} x\, {}^nC_x\, p^x\, q^{n-x} + \sum_{x=2}^{n} x\,(x-1)\, {}^nC_x\, p^x\, q^{n-x} - \mu^2$$

$$= np + \sum_{x=2}^{n} x(x-1)\, \frac{n(n-1)}{x(x-1)} \cdot {}^{(n-2)}C_{x-2} p^x\, q^{n-x} - \mu^2 = np + \sum_{x=2}^{n} n(n-1) \cdot {}^{(n-2)}\, C_{x-2} p^2 p^{x-2} q^{n-x} - \mu^2$$

$$= np + n(n-1)\, p^2 \sum_{x=2}^{n} {}^{(n-2)}C_{x-2}\, p^{x-2}\, q^{n-x} - \mu^2$$

$$= np + n(n-1)p^2 \cdot (q+p)^{n-2} - \mu^2 = np + n(n-1)\, p^2 - \mu^2 \qquad [\because\ p+q=1]$$

$$= np\,[1 + (n-1)p] - \mu^2 = np\,[1 - p + np] - \mu^2 = np\,[q + np] - \mu^2 \qquad [\because\ 1-p=q]$$

$$= np\,(q+np) - (np)^2 = npq$$

3. Standard Deviation of the Binomial Distribution

$$\text{SD} = \sqrt{\text{Variance}} = \sqrt{npq}$$

4. Mode of the Binomial Distribution

Mode of the binomial distribution is the value of x at which $p(x)$ has maximum value.

Mode = integral part of $(n+1)p$, if $(n+1)p$ is not an integer
= $(n+1)\,p$ and $(n+1)\,p - 1$, if $(n+1)\,p$ is an integer.

17.13.4 Recurrence Relation for the Binomial Distribution

For the binomial distribution,

$$P(X=x) = {}^nC_x\, p^x\, q^{n-x}$$

$$P(X=x+1) = {}^nC_{x+1}\, p^{x+1}\, q^{n-x-1}$$

$$\frac{P(X=x+1)}{P(X=x)} = \frac{{}^nC_{x+1}\, p^{x+1}\, q^{n-x-1}}{{}^nC_x\, p^x\, q^{n-x}}$$

$$= \frac{n!}{(x+1)!\,(n-x-1)!} \times \frac{x!\,(n-x)!}{n!} \cdot \frac{p}{q} = \frac{(n-x)\,(n-x-1)!\,x!}{(x+1)\,x!\,(n-x-1)!} \cdot \frac{p}{q} = \frac{n-x}{x+1} \cdot \frac{p}{q}$$

$$P(X=x+1) = \frac{n-x}{x+1} \cdot \frac{p}{q} \cdot P(X=x)$$

17.13.5 Binomial Frequency Distribution

If n independent trials constitute one experiment and this experiment is repeated N times, the frequency of x successes is $N\,P(X=x)$, i.e., $N\,{}^nC_x\, p^x\, q^{n-x}$. This is called expected or theoretical frequency $f(x)$ of a success.

$$\sum_{x=0}^{n} f(x) = N \sum_{x=0}^{n} P(X=x) = N \quad \left[\because \sum_{x=0}^{n} P(X=x) = 1\right]$$

The expected or theoretical frequencies $f(0), f(1), f(2), \ldots, f(n)$ of 0, 1, 2, ..., n, successes are respectively the first, second, third, ..., $(n+1)^{\text{th}}$ term in the expansion of $N(q+p)^n$. The possible number of successes and their frequencies is called a binomial frequency distribution. In practice, the expected frequencies differ from observed frequencies due to chance factor.

EXAMPLE 17.33 *4 coins are tossed simultaneously. What is the probability of getting (i) 2 heads (ii) at least 2 heads (iii) at most 2 heads*

Solution: Let p be the probability of getting a head in the toss of a coin.

$$p = \frac{1}{2}, \quad q = 1-p = 1-\frac{1}{2} = \frac{1}{2}, \quad n = 4$$

The probability of getting x heads when 4 coins are tossed

$$P(X=x) = {}^nC_x\, p^x q^{n-x} = {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}, \quad x = 0, 1, 2, 3, 4$$

(i) Probability of getting 2 heads when 4 coins are tossed

$$P(X=2) = {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

(ii) Probability of getting at least two heads when 4 coins are tossed

$$P(X \geq 2) = P(X=2) + P(X=3) + P(X=4) = \sum_{x=2}^{4} P(X=x)$$

$$= \sum_{x=2}^{4} {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} = \frac{11}{16}$$

(iii) Probability getting at most 2 heads when 4 coins are tossed

$$P(X \le 2) = P(X=0) + P(X=1) + P(X=2)$$

$$= \sum_{x=0}^{2} P(X=x) = \sum_{x=0}^{2} {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} = \frac{11}{16}$$

EXAMPLE 17.34

A multiple-choice test consists of 8 questions with 3 answers to each question (of which only one is correct). A student answers each question by rolling a balanced dice and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4, and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 75% correct answers. If there is no negative making, what is the probability that the student secures a distinction?

Solution: Let p be the probability of getting an answer to a question correctly. There are three answers to each question, out of which only one is correct.

$$p = \frac{1}{3}, \quad q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}, \quad n = 8$$

Probability of getting x correct answers in an 8 questions test

$$P(X=x) = {}^nC_x \; p^x \; q^{n-x} = {}^8C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{8-x}, \quad x = 0, 1, 2, \ldots, 8$$

Probability of securing a distinction, i.e., getting at least 6 correct answers out of the 8 questions

$$P(X \ge 6) = P(X=6) + P(X=7) + P(X=8) = \sum_{x=6}^{8} P(X=x)$$

$$= \sum_{x=6}^{8} {}^8C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{8-x} = \frac{43}{2187} = 0.0197$$

EXAMPLE 17.35

The probability of a man hitting a target is $\frac{1}{3}$. (i) If he fires 5 times, what is the probability of his hitting the target at least twice? (ii) How many times must he fire so that the probability of his hitting the target at least once is more than 90%?

Solution: Let p be probability of hitting a target.

$$p = \frac{1}{3}, \quad q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}, \quad n = 5$$

Probability of hitting the target x times out of 5 times

$$P(X=x) = {}^nC_x \; p^x \; q^{n-x} = {}^5C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{5-x}, \quad x = 0, 1, 2, \ldots, 5$$

(i) Probability of hitting the target at least twice out of 5 times

$$P(X \ge 2) = P(X=2) + P(X=3) + P(X=4) + P(X=5)$$

$$= \sum_{x=2}^{5} P(X=x) = \sum_{x=2}^{5} {}^5C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{5-x} = \frac{131}{243} = 0.5391$$

(ii) Probability of hitting the target at least once out of 5 times

$$P(X \geq 1) > 0.9$$

$$1 - P(X = 0) > 0.9$$

$$1 - {}^{n}C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^n > 0.9$$

$$1 - \left(\frac{2}{3}\right)^n > 0.9$$

For $n = 6$, $1 - \left(\frac{2}{3}\right)^6 = 0.9122$

Hence, the man must fire 6 times so that the probability of hitting the target at least once is more than 90%.

EXAMPLE 17.36

Out of 800 families with 5 children each, how many would you expect to have (i) 3 boys (ii) 5 girls (iii) either 2 or 3 boys (iv) at least one boy? Assume equal probabilities for boys and girls.

Solution: Let p be the probability of having a boy in each family.

$$p = \frac{1}{2}, \quad q = 1 - \frac{1}{2} = \frac{1}{2}, \quad n = 5, \quad N = 800$$

Probability of having x boys out of 5 children in each family

$$P(X = x) = {}^{n}C_x \; p^x \; q^{n-x} = {}^{5}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}, \; x = 0, 1, 2, \ldots, 5$$

(i) Probability of having 3 boys out of 5 children in each family

$$P(X = 3) = {}^{5}C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{5}{16}$$

Expected number of families having 3 boys out of 5 children $= N\,P(X = 3) = 800\left(\frac{5}{16}\right) = 250$

(ii) Probability of having 5 girls, i.e., no boys out of 5 children in each family

$$P(X = 0) = {}^{5}C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

Expected number of families 5 girls out of 5 children $= NP(X = 0) = 800\left(\frac{1}{32}\right) = 25$

(iii) Probability of having either 2 or 3 boys out of 5 children in each family

$$P(X = 2) + P(X = 3) = \sum_{x=2}^{3} P(X = x) = \sum_{x=2}^{3} {}^{5}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x} = \frac{5}{8}$$

Expected number of families having either 2 of 3 boys out of 5 children

$$= N[P(X = 2) + P(X = 3)] = 800\left(\frac{5}{8}\right) = 500$$

(iv) Probability of having at least one boy out of 5 children in each family

$$P(X \geq 1) = P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)$$

$$= \sum_{x=1}^{5} P(X=x) = \sum_{x=1}^{5} {}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x} = \frac{31}{32}$$

Expected number of families having at least-one boy out of 5 children

$$= NP(X \geq 1) = 800\left(\frac{31}{32}\right) = 775$$

EXAMPLE 17.37

Seven unbiased coins are tossed 128 times and the number of heads obtained is noted as given below:

No. of Heads	0	1	2	3	4	5	6	7
Frequency	7	6	19	35	30	23	7	1

Fit a binomial distribution to the data.

Solution: Since the coin is unbiased,

$$p = \frac{1}{2}, \quad q = \frac{1}{2}, n = 7, N = 128$$

For binomial distribution,

$$P(X=x) = {}^nC_x\ p^x\ q^{n-x} = {}^7C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{7-x}, x = 0, 1, 2, ..., 7$$

Theoretical or expected frequency $f(x) = N\,P(X=x)$

$$f(x) = 128\ {}^7C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{7-x} = 128\ {}^7C_x \left(\frac{1}{2}\right)^7$$

$$f(0) = 128\ {}^7C_0 \left(\frac{1}{2}\right)^7 = 1$$

$$f(1) = 128\ {}^7C_1 \left(\frac{1}{2}\right)^7 = 7$$

$$f(2) = 128\ {}^7C_2 \left(\frac{1}{2}\right)^7 = 21$$

$$f(3) = 128\ {}^7C_3 \left(\frac{1}{2}\right)^7 = 35$$

$$f(4) = 128\ {}^7C_4 \left(\frac{1}{2}\right)^7 = 35$$

$$f(5) = 128\ {}^7C_5 \left(\frac{1}{2}\right)^7 = 21$$

$$f(6) = 128\ {}^7C_6 \left(\frac{1}{2}\right)^7 = 7$$

$$f(7) = 128\ {}^{7}C_{7}\left(\frac{1}{2}\right)^{7} = 1$$

Binomial distribution

No. of Heads x	0	1	2	3	4	5	6	7
Expected Binomial Frequency $f(x)$	1	7	21	35	35	21	7	1

EXERCISE 17.11

1. Find the fallacy if any in the following statements:

(a) The mean of a binomial distribution is 6 and SD is 4.

(b) The mean of a binomial distribution is 9 and its SD is 4.

[**Ans.:** (a) False, $q = \frac{8}{3}$ is impossible

(b) False, $q = \frac{19}{9}$ is impossible]

2. The mean and variance of a binomial distribution are 3 and 1.2 respectively. Find n, p, and $P(X < 4)$.

[**Ans.:** 5, 0.6, $\frac{2068}{3125}$]

3. Find the binomial distribution if the mean is 5 and the variance is $\frac{10}{3}$. Find $P(X = 2)$.

[**Ans.:** $P(X = x) = {}^{25}C_{x}\left(\frac{1}{3}\right)^{x}\left(\frac{2}{3}\right)^{25-x}$, 0.003]

4. In a binomial distribution, the mean and variance are 4 and 3 respectively. Find $P(X \geq 1)$.

[**Ans.:** 0.9899]

5. The odds in favour of X winning a game against Y are 4:3. Find the probability of Y winning 3 games out of 7 played.

[**Ans.:** 0.0929]

6. On an average, 3 out of 10 students fail in an examination. What is the probability that out of 10 students that appear for the examination none will fail?

[**Ans.:** 0.0282]

7. If on the average rain falls on 10 days in every thirty, find the probability (i) that the first three days of a week will be fine and remaining wet, and (ii) that rain will fall on just three days of a week.

[**Ans.:** (i) $\frac{8}{2187}$ (ii) $\frac{280}{2187}$]

8. Two unbiased dice are thrown three times. Find the probability that the sum nine would be obtained (i) once, and (ii) twice.

[**Ans.:** (i) 0.26 (ii) 0.03]

9. For special security in a certain protected area, it was decided to put three lightbulbs on each pole. If each bulb has probability p of burning out in the first 100 hours of service, calculate the probability that at least one of them is still good after 100 hours. If $p = 0.3$, how many bulbs would be needed on each pole to ensure with 99% safety that at least one is good after 100 hours?

[**Ans.:** (i) $1 - p^3$ (ii) 4]

10. It is known from past records that 80% of the students in a school do their homework. Find the probability that during a random check of 10 students, (i) all have done their homework, (ii) at the most two have not done their homework, and (iii) at least one has not done the homework.

[**Ans.:** (i) 0.1074 (ii) 0.6778 (iii) 0.8926]

11. An insurance salesman sells policies to 5 men, all of identical age and good health. According to the actuarial tables, the probability that a man of this particular age will be alive 30 years hence is $\frac{2}{3}$. Find the probability that 30 years hence (i) at least 1 man will be alive, (ii) at least 3 men will be alive, and (iii) all 5 men will be alive.

$\left[\textbf{Ans.:}\ (i)\ \frac{242}{243}\ (ii)\ \frac{64}{81}\ (iii)\ \frac{32}{243}\right]$

12. A company has appointed 10 new secretaries out of which 7 are trained. If a particular executive is to get three secretaries selected at random, what is the chance that at least one of them will be untrained?

[**Ans.:** 0.7083]

13. The overall pass rate in a university examination is 70%. Four candidates take up such an examination. What is the probability that (i) at least one of them will pass? (ii) all of them will pass the examination?

[**Ans.:** (i) 0.9919 (ii) 0.7599]

14. The normal rate of infection of a certain disease in animals is known to be 25%. In an experiment with a new vaccine, it was observed that none of the animals caught the infection. Calculate the probability of the observed result.

$\left[\textbf{Ans.:}\ \frac{729}{4096}\right]$

15. Suppose that weather records show that on the average, 5 out of 31 days in October are rainy days. Assuming a binomial distribution with each day of October as an independent trial, find the probability that the next October will have at most three rainy days.

[**Ans.:** 0.2403]

16. Assuming that half the population of a village is female and assuming that 100 samples each of 10 individuals are taken, how many samples would you expect to have 3 or less females?

[**Ans.:** 17]

17. Assuming that half the population of a town is vegetarian so that the chance of an individual being vegetarian is $\frac{1}{2}$, and assuming that 100 investigators can take a sample of 10 individuals to see whether they are vegetarians, how many investigators would you expect to report that three people or less in the sample were vegetarians?

[**Ans.:** 17]

18. The probability of failure in a physics practical examination is 20%. If 25 batches of 6 students each take the examination, in how many batches of 4 or more students would pass?

[**Ans.:** 23]

19. A lot contains 1% defective items. What should be the number of items in a lot so that the probability of finding at least one defective item in it is at least 0.95?

[**Ans.:** 299]

20. The probability that a bomb will hit the target is 0.2. Two bombs are required to destroy the target. If six bombs are used, find the probability that the target will be destroyed.

[**Ans.:** 0.3447]

21. Out of 1000 families with 4 children each, how many would you expect to have (i) 2 boys and 2 girls? (ii) at least one boy? (iii) no girl? (iv) at most 2 girls?

[**Ans.:** (i) 375 (ii) 938 (iii) 63 (iv) 69]

22. In a sampling of a large number of parts produced by a machine, the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples, how many samples would you expect to contain at least 3 defectives?

[**Ans.:** 323]

23. Five pair coins are tossed 3200 times, find the frequency distribution of the number of heads obtained. Also, find the mean and SD.

[**Ans.:** (i) 100, 500, 1000, 1000, 500, 100
(ii) 1600 (iii) 28.28]

24. Fit a binomial distribution to the following data:

x	0	1	2	3	4
f	12	66	109	59	10

[**Ans.:** 17, 67, 96, 61, 15]

17.14 POISSON DISTRIBUTION

Poisson distribution is a limiting case of binomial distribution under the following conditions:

(i) The number of trials should be infinitely large, i.e., $n \to \infty$.
(ii) The probability of successes p for each trial should be very small, i.e., $p \to 0$.
(iii) $np = \lambda$ should be finite, where λ is a constant.

The binomial distribution is

$$P(X = x) = {}^nC_x\, p^x\, q^{n-x} = {}^nC_x \left(\frac{p}{q}\right)^x q^n = {}^nC_x \left(\frac{p}{1-p}\right)^x (1-p)^n$$

Putting $p = \dfrac{\lambda}{n}$,

$$P(X = x) = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \left(\frac{\frac{\lambda}{n}}{1-\frac{\lambda}{n}}\right)^x \left(1-\frac{\lambda}{n}\right)^n$$

$$= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \frac{1}{\left(1-\frac{\lambda}{n}\right)^x} \left(1-\frac{\lambda}{n}\right)^n$$

$$= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x} = \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left[1-\left(\frac{x-1}{n}\right)\right]}{x!} \lambda^x \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

Since $\lim\limits_{n\to\infty} \left(1-\dfrac{\lambda}{n}\right)^n = e^{-\lambda}$

and $\lim\limits_{n\to\infty} \left(1-\dfrac{1}{n}\right) = \lim\limits_{n\to\infty} \left(1-\dfrac{2}{n}\right) = 1$

Taking the limits of both the sides as $n \to \infty$,

$$P(X = x) = \frac{\lambda^x\, e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \infty$$

A random variable X is said to follow poisson distribution if the probability of x is given by

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

where λ is called the *parameter of the distribution.*

17.14.1 Examples of Poisson Distribution

(i) Number of defective bulbs produced by a reputed company
(ii) Number of telephone calls per minute at a switchboard
(iii) Number of cars passing a certain point in one minute
(iv) Number of printing mistakes per page in a large text
(v) Number of persons born blind per year in a large city

17.14.2 Conditions of Poisson Distribution

The Poisson distribution holds under the following conditions:

(i) The random variable X should be discrete.
(ii) The numbers of trials n is very large.
(iii) The probability of success p is very small (very close to zero).
(iv) $\lambda = np$ is finite.
(v) The occurrences are rare.

17.14.3 Constants of the Poisson Distribution

1. Mean of the Poisson Distribution

$$E(X) = \sum_{x=0}^{\infty} x\, p(x) = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{x e^{-\lambda}\lambda\lambda^{x-1}}{x!} = e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{x\lambda^{x-1}}{x!}$$

$$= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \qquad \left[\because \frac{x}{x!} = \frac{1}{(x-1)!}\right]$$

$$= \lambda e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \cdots\right) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

2. Variance of the Poisson Distribution

$$\text{Var}(X) = E(X^2) - \mu^2 = \sum_{x=1}^{\infty} x^2\, p(x) - \mu^2 = \sum_{x=1}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} - \lambda^2$$

$$= \sum_{x=1}^{\infty} x[(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!} - \lambda^2 = \sum_{x=2}^{\infty} \frac{x(x-1)e^{-\lambda}\lambda^x}{x!} + \sum_{x=1}^{\infty} \frac{x e^{-\lambda}\lambda^x}{x!} - \lambda^2$$

$$= \sum_{x=2}^{\infty} \frac{x(x-1)e^{-\lambda}\lambda^{x-2}\lambda^2}{x(x-1)(x-2)\cdots 1} + \lambda - \lambda^2 = e^{-\lambda}\lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda - \lambda^2$$

$$= e^{-\lambda}\lambda^2\left(1 + \lambda + \frac{\lambda^2}{2!} + \cdots\right) + \lambda - \lambda^2 = e^{-\lambda}\lambda^2 e^{\lambda} + \lambda - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

3. Standard Deviation of the Poisson Distribution

$$\text{SD} = \sqrt{\text{Variance}} = \sqrt{\lambda}$$

4. Mode of the Poisson Distribution

Mode is the value of x for which the probability $p(x)$ is maximum.

$$p(x) \geq p(x+1) \text{ and } p(x) \geq p(x-1)$$

When $p(x) \geq p(x+1)$,

$$\frac{e^{-\lambda}\,\lambda^{x}}{x!} \geq \frac{e^{-\lambda}\,\lambda^{x+1}}{(x+1)!}$$

$$1 \geq \frac{\lambda}{x+1}$$

$$(x+1) \geq \lambda$$

$$x \geq \lambda - 1 \qquad \ldots(17.1)$$

Similarly, for $p(x) \geq p(x-1)$,

$$x \leq \lambda \qquad \ldots(17.2)$$

Combining Eqs (17.1) and (17.2),

$$\lambda - 1 \leq x \leq \lambda$$

Hence, the mode of the Poisson distribution lies between $\lambda - 1$ and λ.

Case I If λ is an integer then $\lambda - 1$ is also an integer. The distribution is bimodal and the two modes are $\lambda - 1$ and λ.

Case II If λ is not an integer, the distribution is unimodal and the mode of the Poisson distribution is an integral part of λ. The mode is the integer between $\lambda - 1$ and λ.

17.14.4 Recurrence Relation for the Poisson Distribution

For the Poisson distribution,

$$p(x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

$$p(x+1) = \frac{e^{-\lambda}\,\lambda^{x+1}}{(x+1)!}$$

$$\frac{p(x+1)}{p(x)} = \frac{e^{-\lambda}\,\lambda^{x+1}}{(x+1)!} \cdot \frac{x!}{e^{-\lambda}\,\lambda^{x}} = \frac{\lambda}{x+1}$$

$$p(x+1) = \frac{\lambda}{x+1}\,p(x)$$

HISTORICAL DATA

Siméon Denis Poisson (1781–1840) was a French mathematician, geometer, and physicist. He obtained many important results, but within the elite Académie des Sciences he also was the final leading opponent of the wave theory of light and was proven wrong on that matter by Augustin-Jean Fresnel.

The Poisson distribution in probability theory and statistics has been named after him.

He contributed significantly for introducing the Poisson process, Poisson equation, Poisson kernel, Poisson distribution, Poisson bracket, Poisson algebra, Poisson regression, Poisson summation formula, Poisson's spot, Poisson's ratio, Poisson zeros, Conway–Maxwell–Poisson distribution and Euler–Poisson–Darboux equation.

EXAMPLE 17.38

If 2% of lightbulbs are defective, find the probability that (i) at least one is defective, and (ii) exactly 7 are defective. Also, find $P(1 < X < 8)$ in a sample of 100.

Solution: Let p be the probability of defective bulb.

$$p = 2\% = 0.02$$
$$n = 100$$

Since p is very small and n is large, Poisson distribution is used.

$$\lambda = np = 100(0.02) = 2$$

Let X be the random variable which denotes the number of defective bulbs in a sample of 100.

Probability of x defective bulb in a sample of 100

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-2}2^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) Probability that at least one bulb is defective

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-2}2^0}{0!} = 0.8647$$

(ii) Probability that exactly 7 bulbs are defective

$$P(X = 7) = \frac{e^{-2}2^7}{7!} = 0.0034$$

(iii) $P(1 < X < 8) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)$

$$= \sum_{x=2}^{7} P(X = x) = \sum_{x=2}^{7} \frac{e^{-2}2^x}{x!} = 0.5929$$

EXAMPLE 17.39

A manufacturer, who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of bottles. Using Poisson distribution, find how many boxes will contain (i) no defective bottles and (ii) at least 2 defective bottles.

Solution: Let p be the probability of deflective bottles.

$$p = 0.1\% = 0.001$$
$$n = 500$$
$$\lambda = np = 500(0.001) = 0.5$$

Let X be the random variable which denotes the number of defective bottles in a box of 500.

Probability of x defective bottles in a box of 500

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.5}0.5^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) Probability of no defective bottles in a box

$$P(X=0)=\frac{e^{-0.5}\,0.5^0}{0!}=0.6065$$

Number of boxes containing no defective bottles

$f(x)=N\,P(x=0)=100(0.6065)\approx 61$

(ii) Probability of at least 2 defective bottles

$$P(X\geq 2)=1-P(X<2)=1-[P(X=0)+P(X=1)]$$

$$=1\sum_{x=0}^{1}P(X=x)=1-\sum_{x=0}^{1}\frac{e^{-0.5}\,0.5^x}{x!}=0.0902$$

Number of boxes containing at least 2 defective bottles

$f(x)=N\,P(X\geq 2)=100\,(0.0902)\approx 9$

EXAMPLE 17.40

In a certain factory turning out blades, there is a small chance of $\frac{1}{500}$ *for any blade to be defective. The blades are supplied in packets of 10. Use the Poisson distribution to calculate the approximate number of packets containing no defective, one defective, and two defective blades in a consignment of 10000 packets.*

Solution: Let p be the probability of defective blades in a packet.

$$p=\frac{1}{500},\qquad n=10,\qquad N=10000$$

$$\lambda=np=10\left(\frac{1}{500}\right)=0.02$$

Let X be the random variable which denotes the number of defective blades in a packet.

Probability of x defective blades in a packet

$$P(X=x)=\frac{e^{-\lambda}\,\lambda^x}{x!}=\frac{e^{-0.02}\,0.02^x}{x!},\qquad x=0,1,2,\ldots$$

(i) Probability of no defective blades in a packet

$$P(X=0)=\frac{e^{-0.02}\,0.02^0}{0!}=0.9802$$

Number of packets with no defective blades

$f(x)=N\,P(X=0)=10000(0.9802)=9802$

(ii) Probability of one defective blade in a packet

$$P(X=1)=\frac{e^{-0.02}\,0.02^1}{1!}=0.0196$$

Number of packets with one defective blade

$f(x)=N\,P(X=1)=10000\,(0.0196)=196$

(iii) Probability of two defective blades in a packet

$$P(X=2)=\frac{e^{-0.02}\,0.02^2}{2!}=1.96\times10^{-4}$$

Number of packets with 2 defective blades

$$f(x)=N\,P(X=2)=10000\,(1.96\times10^{-4})=1.96\approx 2$$

EXAMPLE 17.41 *Fit a Poisson distribution to the following data:*

Number of Deaths (x)	0	1	2	3	4
Frequency (f)	122	60	15	2	1

Solution:

$$\text{Mean}=\frac{\sum fx}{\sum f}=\frac{122(0)+60(1)+15(2)+2(3)+1(4)}{122+60+15+2+1}=\frac{100}{200}=0.5$$

For a Poisson distribution,

$$\lambda=0.5$$

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-0.5}\,0.5^x}{x!},\quad x=0,1,2,3,4$$

$$N=\sum f=100$$

Theoretical or expected frequency $f(x)=N\,P(X=x)$

$$f(x)=\frac{200\,e^{-0.5}\,0.5^x}{x!}$$

$$f(0)=\frac{200\,e^{-0.5}\,0.5^0}{0!}=121.31\approx 121$$

$$f(1)=\frac{200\,e^{-0.5}\,0.5^1}{1!}=60.65\approx 61$$

$$f(2)=\frac{200\,e^{-0.5}\,0.5^2}{2!}=15.16\approx 15$$

$$f(3)=\frac{200\,e^{-0.5}\,0.5^3}{3!}=2.53\approx 3$$

$$f(4)=\frac{200\,e^{-0.5}\,0.5^4}{4!}=0.32\approx 0$$

Poisson Distribution

Number of Deaths (x)	0	1	2	3	4
Expected Poisson Frequency f(x)	121	61	15	3	0

EXERCISE 17.12

1. The mean and variance of a probability distribution is 2. Write down the distribution.

[**Ans.:** $P(X=x)=\dfrac{e^{-2}\,2^x}{x!},\ x=0,1,2,\ldots$]

2. In a Poisson distribution, the probability $P(X=0)$ is 20 percent. Find the mean of the distribution.

[**Ans.:** 2.9957]

3. If X is a Poisson variate and $P(X=0)=6\,P(X=3)$, find $P(X=2)$.

[**Ans.:** 0.1839]

4. The standard deviation of a Poisson distribution is 3. Find the probability of getting 3 successes.

[**Ans.:** 0.0149]

5. The probability that a Poisson variable X takes a positive value is $1-e^{-1.5}$. Find the variance and the probability that X lies between –1.5 and 1.5.

[**Ans.:** 1.5, 0.5578]

6. If 2 per cent bulbs are known to be defective bulbs, find the probability that in a lot of 300 bulbs, there will be 2 or 3 defective bulbs using Poisson distribution.

[**Ans.:** 0.1338]

7. In a certain manufacturing process, 5% of the tools produced turn out to be defective. Find the probability that in a sample of 40 tools, at most 2 will be defective.

[**Ans.:** 0.675]

8. If the probability that an individual suffers a bad reaction from a particular injection is 0.001, determine the probability that out of 2000 individuals (i) exactly three, and (ii) more than two individuals suffer a bad reaction.

[**Ans.:** (i) 0.1804 (ii) 0.3233]

9. It is known from past experience that in a certain plant, there are on the average 4 industrial accidents per year. Find the probability that in a given year, there will be less than 4 accidents. Assume Poisson distribution.

[**Ans.:** 0.43]

10. Find the probability that at most 5 defective fuses will be found in a box of 200 fuses, if experience shows that 2% of such fuses are defective.

[**Ans.:** 0.7851]

11. Assume that the probability of an individual coal minor being killed in a mine accident during a year is $\dfrac{1}{2400}$. Use appropriate statistical distribution to calculate the probability that in a mine employing 200 miners, there will be at least one fatal accident every year.

[**Ans.:** 0.07]

12. Between the hours of 2 and 4 p.m., the average number of phone calls per minute coming into the switchboard of a company is 2.5. Find the probability that during a particular minute, there will be (i) no phone call at all, (ii) 4 or less calls, and (iii) more than 6 calls.

[**Ans.:** (i) 0.0821 (ii) 0.8909 (iii) 0.0145]

13. Suppose that a local appliances shop has found from experience that the demand for tubelights is roughly distributed as Poisson with a mean of 4 tubelights per week. If the shop keeps 6 tubelights during a particular week, what is the probability that the demand will exceed the supply during that week?

[**Ans.:** 0.1106]

14. The distribution of the number of road accidents per day in a city is Poisson with

a mean of 4. Find the number of days out of 100 days when there will be (i) no accident, (i) at least 2 accidents, and (iii) at most 3 accidents.

[**Ans.:** (i) 2 (ii) 91 (iii) 44]

15. A manufacturer of electric bulbs sends out 500 lots each consisting of 100 bulbs. If 5% bulbs are defective, in how many lot can we expect (i) 97 or more good bulbs? (ii) less than 96 good bulbs?

[**Ans.:** (i) 62 (ii) 132]

16. A firm produces articles, 0.1 per cent of which are defective. It packs them in cases containing 500 articles. If a wholesaler purchases 100 such cases, how many cases can be expectecd (i) to be free from defects? (ii) to have one defective article?

[**Ans.:** (i) 16 (ii) 30]

17. In a certain factory producing certain articles, the probability that an article is defective is $\frac{1}{500}$. The articles are supplied in packets of 20. Find approximately the number of packets containing no defective, one defective, two defectives in a consignment of 20000 packets.

[**Ans.:** 19200, 768, 15]

18. In a certain factory manufacturing razor blades, there is a small chance, $\frac{1}{50}$ for any blade to be defective. The blades are placed in packets, each containing 10 blades. Using the Poisson distribution, calculate the approximate number of packets containing not more than 2 defective blades in a consignment of 10000 packets.

[**Ans.:** 9988]

19. It is known that 0.5% of ballpen refills produced by a factory are defective. These refills are dispatched in packaging of equal numbers. Using a Poisson distribution, determine the number of refills in a packing to be sure that at least 95% of them contain no defective refills.

[**Ans.:** 10]

20. A manufacturer finds that the average demand per day for the mechanics to repair his new product is 1.5 over a period of one year and the demand per day is distributed as a Poisson variate. He employs two mechanics. On how many days in one year (i) would both mechanics would be free? (ii) some demand is refused?

[**Ans.:** (i) 81.4 days (ii) 69.8 days]

21. Fit a Poisson distribution to the following data:

X	0	1	2	3	4
f	211	90	19	5	0

[**Ans.:** $\lambda = 0.44$, Frequencies: 209, 92, 20, 3, 1]

22. Fit a Poisson distribution to the following data:

No. of Defects Per Piece	0	1	2	3	4
No. of Pieces	43	40	25	10	2

[**Ans.:** Frequencies: 42, 44, 24, 8, 2]

23. Fit a Poisson distribution to the following data:

X	0	1	2	3	4	5
f	142	156	69	27	5	1

[**Ans.:** Frequencies: 147, 147, 74, 24, 6, 2]

24. Fit a Poisson distribution to the following data:

X	0	1	2	3	4	5	6	7	8
f	56	156	132	92	37	22	4	0	1

[**Ans.:** Frequency: 70, 137, 135, 89, 44, 17, 6, 2, 0]

17.15 NORMAL DISTRIBUTION

A continuous random variable X is said to follow normal distribution with mean μ and variance σ^2, if its probability function is given by

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

$$-\infty < X < \infty, -\infty < \mu < \infty, \sigma > 0$$

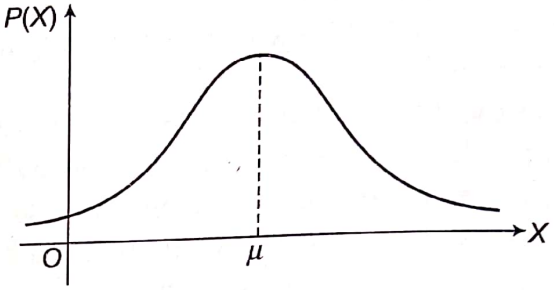


Fig. 17.7 Normal Distribution Curve

where μ and σ are called parameters of the normal distribution. The curve representing the normal distribution is called the normal curve (Fig. 17.7).

17.15.1 Properties of the Normal Distribution

A normal probability curve, or normal curve, has the following properties:

(i) It is a bell-shaped symmetrical curve about the ordinate $X = \mu$. The ordinate is maximum at $X = \mu$.

(ii) It is a unimodal curve and its tails extend infinitely in both the directions, i.e., the curve is asymptotic to X-axis in both the directions.

(iii) All the three measures of central tendency coincide, i.e., mean = median = mode

(iv) The total area under the curve gives the total probability of the random variable X taking values between $-\infty$ to ∞. Mathematically,

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

(v) The line at $X = \mu$ divides the area under the normal curve into two equal parts, i.e.,

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

(vi) The value of $f(x)$ is always nonnegative for all values of X, i.e., the whole curve lies above the X-axis.

(vii) The area under the normal curve (Fig. 17.8) is distributed as follows:

(a) The area between the lines at $X = \mu - \sigma$ and $X = \mu + \sigma$ is 68.27%

(b) The area between the lines at $X = \mu - 2\sigma$ and $X = \mu + 2\sigma$ is 95.45%

(c) The area between the lines at $X = \mu - 3\sigma$ and $X = \mu + 3\sigma$ is 99.74%

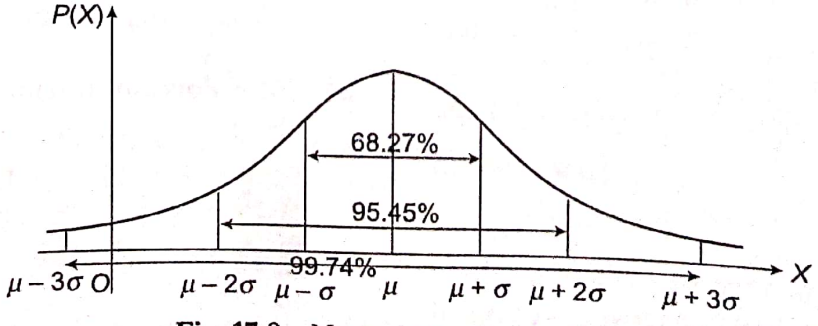


Fig. 17.8 Normal Distribution Curve

17.15.2 Constants of the Normal Distribution

1. Mean of the Normal Distribution

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Putting $\frac{x-\mu}{\sigma} = t$, $dx = \sigma\, dt$

$$E(X) = \int_{-\infty}^{\infty} (\mu + \sigma t) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt + \int_{-\infty}^{\infty} \sigma \frac{t}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

Putting $t^2 = u$ in the second integral,

$2t\, dt = du$

When $t \to \infty$, $u \to \infty$

When $t \to -\infty$, $u \to \infty$

$$E(X) = \mu \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} + \int_{\infty}^{\infty} \sigma \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u} \frac{du}{2} \qquad \left[\because \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2} dt = \sqrt{2\pi}\right]$$

$$= \mu + 0 \qquad [\because \text{ the limits of integration are same}]$$

$$= \mu$$

2. Variance of the Normal Distribution

$\text{Var}(X) = E(X-\mu)^2$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Putting $\frac{x-\mu}{\sigma} = t$, $dx = \sigma\, dt$

$$\text{Var}(X) = \int_{-\infty}^{\infty} \sigma^2 t^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{1}{2}t^2} dt$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} t^2 e^{-\frac{1}{2}t^2} dt \qquad [\because \text{ integral is an even function}]$$

Putting $\frac{t^2}{2} = u$,

$t = \sqrt{2u}$

$$dt = \sqrt{2} \frac{1}{2\sqrt{u}} du = \frac{1}{\sqrt{2u}} du$$

When $t = 0$, $u = 0$

When $t = \infty$, $u = \infty$

$$\text{Var}(X) = \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^\infty 2u e^{-u} \frac{1}{\sqrt{2u}}\, du = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty e^{-u}\, u^{\frac{1}{2}}\, du = \frac{2\sigma^2}{\sqrt{\pi}} \left\lceil \frac{3}{2} \right. \quad \left[\because \int_0^\infty e^{-x} x^{n-1}\, dx = \lceil n \right]$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \left\lceil \frac{1}{2} = \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \sqrt{\pi} = \sigma^2 \right.$$

3. Standard Deviation of the Normal Distribution

$$\text{SD} = \sigma$$

4. Mode of the Normal Distribution

Mode is the value of x for which $f(x)$ is maximum. Mode is given by

$$f'(x) = 0 \quad \text{and} \quad f''(x) < 0$$

For normal distribution,

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

Differentiating w.r.t. x,

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left[-\left(\frac{x-\mu}{\sigma^2} \right) \right] = -\frac{x-\mu}{\sigma^2} f(x)$$

When $f'(x) = 0$, $\quad x - \mu = 0$

$$x = \mu$$

$$f''(x) = -\frac{1}{\sigma^2}[(x-\mu) f'(x) + f(x)]$$

$$= -\frac{1}{\sigma^2}\left[(x-\mu)\left\{ -\frac{(x-\mu)}{\sigma^2} f(x) \right\} + f(x) \right] = -\frac{1}{\sigma^2} f(x) \left[1 - \frac{(x-\mu)^2}{\sigma^2} \right]$$

At $x = \mu$,

$$f''(x) = -\frac{f(x)}{\sigma^2} = -\frac{1}{\sigma^3\sqrt{2\pi}} < 0$$

Hence, $x = \mu$ is the mode of the normal distribution.

5. Median of the Normal Distribution

If M is median of the normal distribution,

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

$$\int_{-\infty}^{M} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2} \qquad \ldots(17.3)$$

Putting $\frac{x-\mu}{\sigma} = t$ in the first integral,

$dx = \sigma\, dt$

When $x \to -\infty$, $t \to -\infty$

When $x = \mu$, $t = 0$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{1}{2}t^2} \sigma\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{1}{2}t^2} dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{1}{2}t^2} dt \quad \text{[By symmetry]}$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} = \frac{1}{2} \qquad \ldots(17.4)$$

From Eqs (17.3) and (17.4),

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2}$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0$$

$$\int_{\mu}^{M} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0$$

$$\mu = M \left[\because \quad \text{if } \int_{a}^{b} f(x)\, dx = 0 \text{ then } a = b, \text{ where } f(x) > 0 \right]$$

Hence, mean = median for the normal distribution.

Note For normal distribution,
mean = median = mode = μ
Hence, the normal distribution is symmetrical.

17.15.3 Probability of a Normal Random Variable in an Interval

Let X be a normal random variable with mean μ and standard deviation σ. The probability of X lying in the interval (x_1, x_2) (Fig. 17.9) is given by

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Hence, the probability is equal to the area under the normal curve between the lines $X = x_1$ and $X = x_2$ respectively. $P(x_1 < X < x_2)$ can be evaluated easily by converting a normal random variable into another random variable.

Let $Z = \frac{X-\mu}{\sigma}$ be a new random variable.

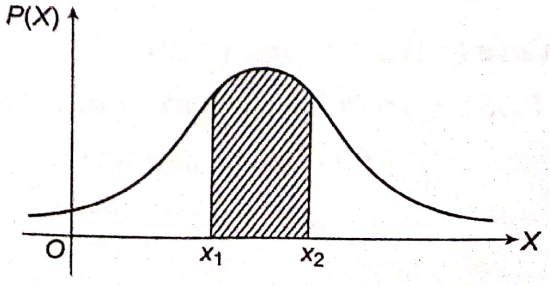


Fig. 17.9 Normal Distribution Curve

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(X)-\mu] = 0$$

$$\text{Var}(Z) = \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X-\mu) = \frac{1}{\sigma^2}\text{Var}(X) = 1$$

The distribution of Z is also normal. Thus, if X is a normal random variable with mean μ and standard deviation σ then $Z = \dfrac{X-\mu}{\sigma}$ is a normal random variable with mean 0 and standard deviation 1. Since the parameters of the distribution of Z are fixed, it is a known distribution and is termed *standard normal distribution*. Further, Z is termed as a *standard normal variate*. Thus, the distribution of any normal variate X can always be transformed into the distribution of the standard normal variate Z.

$$P(x_1 \le X \le x_2) = P\left[\left(\frac{x_1-\mu}{\sigma}\right) \le \left(\frac{X-\mu}{\sigma}\right) \le \left(\frac{x_2-\mu}{\sigma}\right)\right] = P(z_1 \le Z \le z_2)$$

where $z_1 = \dfrac{x_1-\mu}{\sigma}$ and $z_2 = \dfrac{x_2-\mu}{\sigma}$

This probability is equal to the area under the standard normal curve between the lines at $Z = z_1$ and $Z = z_2$.

Case I If both z_1 and z_2 are positive (or both negative) (Fig. 17.10),

$$P(x_1 \le X \le x_2) = P(z_1 \le Z \le z_2) = P(0 \le Z \le z_2) - P(0 \le Z \le z_1)$$

= (Area under the normal curve from 0 to z_2)

−(Area under the normal curve from 0 to z_1)

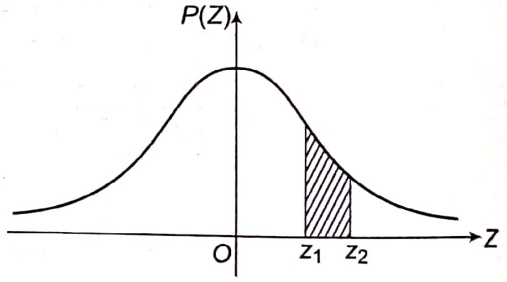


Fig. 17.10 Normal Distribution Curve

Case II If $z_1 < 0$ and $z_2 > 0$ (Fig. 17.11),

$$P(x_1 \le X \le x_2) = P(-z_1 \le Z \le z_2) = P(-z_1 \le Z \le 0) + P(0 \le Z \le z_2)$$

$= P(0 \le Z \le z_1) + P(0 \le Z \le z_2)$ [By symmetry]

= (Area under the normal curve from 0 to z_1)

+(Area under the normal curve from 0 to z_2)

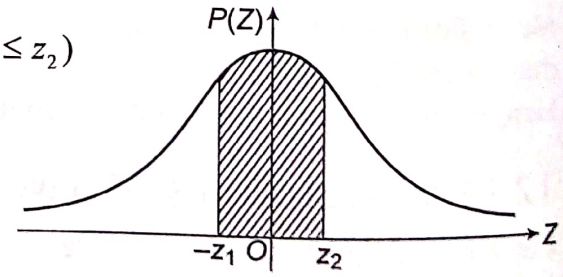


Fig. 17.11 Normal Distribution Curve

When $X > x_1$, $Z > z_1$, the probability $P(Z > z_1)$ can be found for two cases as follows:

Case I If $z_1 > 0$ (Fig. 17.12),

$$P(X > x_1) = P(Z > z_1) = 0.5 - P(0 \le Z \le z_1)$$

$= 0.5$ – (Area under the curve from 0 to z_1)

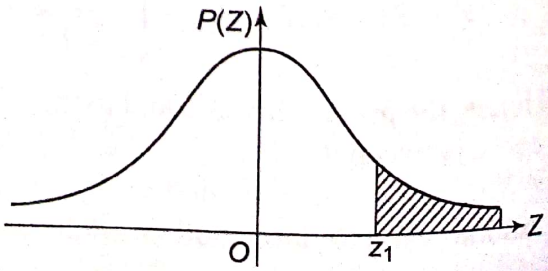


Fig. 17.12 Normal Distribution Curve

Case II If $z_1 < 0$ (Fig. 17.13),

$$P(X > x_1) = P(Z > -z_1) = 0.5 + P(-z_1 < Z < 0)$$

$$= 0.5 + P(0 < Z < z_1) \quad \text{[By symmetry]}$$

$$= 0.5 + (\text{Area under the curve from 0 to } z_1)$$

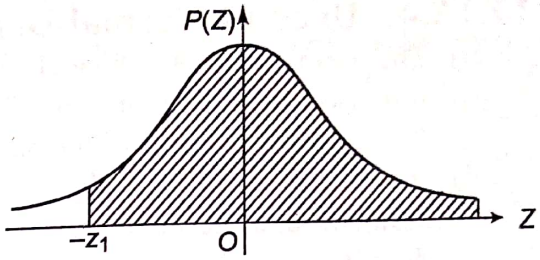


Fig. 17.13 Normal Distribution Curve

When $X < x_1$, $Z < z_1$, the probability $P(Z < z_1)$ can be found for two cases as follows:

Case I If $z_1 > 0$ (Fig. 17.14),

$$P(X < x_1) = P(Z < z_1) = 1 - P(Z \geq z_1)$$

$$= 1 - [0.5 - P(0 < Z < z_1)] = 0.5 + P(0 < Z < z_1)$$

$$= 0.5 + (\text{Area under the curve from 0 to } z_1)$$

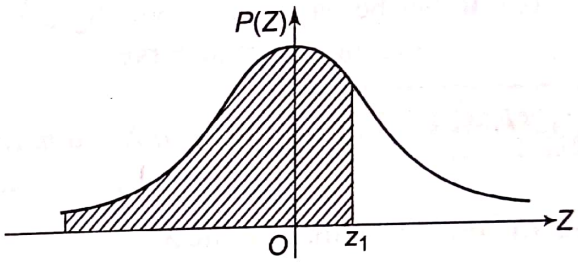


Fig. 17.14 Normal Distribution Curve

Case II If $z_1 < 0$ (Fig. 17.15),

$$P(X < x_1) = P(Z < -z_1) = 1 - P(Z \geq -z_1)$$

$$= 1 - [0.5 + P(-z_1 \leq Z \leq 0)]$$

$$= 1 - [0.5 + P(0 \leq Z \leq z_1)] \quad \text{[By symmetry]}$$

$$= 0.5 - P(0 \leq Z \leq z_1)$$

$$= 0.5 - (\text{Area under the curve from 0 to } z_1)$$

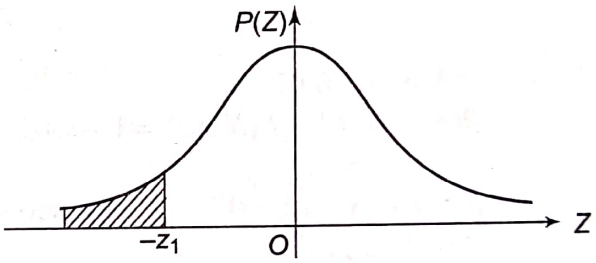


Fig. 17.15 Normal Distribution Curve

Notes

(i) $P(X < x_1) = F(x_1) = \int_{-\infty}^{x_1} f(x)\,dx$

Hence, $P(X < x_1)$ represents the area under the curve from $X = -\infty$ to $X = x_1$.

(ii) If $P(X < x_1) < 0.5$, the point x_1 lies to the left of $X = \mu$ and the corresponding value of standard normal variate will be negative (Fig. 17.16).

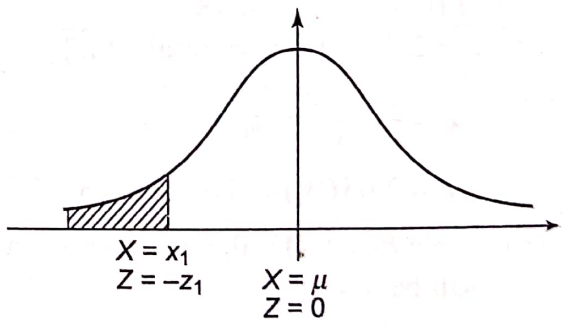


Fig. 17.16 Normal Distribution Curve

(iii) If $P(X < x_1) > 0.5$, the point x_1 lies to the right of $x = \mu$ and the corresponding value of standard normal variate will be positive (Fig. 17.17).

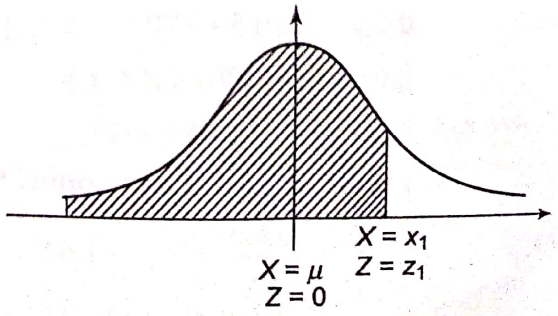


Fig. 17.17 Normal Distribution Curve

17.15.4 Uses of Normal Distribution

(i) The normal distribution can be used to approximate binomial and Poisson distributions.

(ii) It is used extensively in sampling theory. It helps to estimate parameters from statistics and to find confidence limits of the parameter.

(iii) It is widely used in testing statistical hypothesis and tests of significance in which it is always assumed that the population from which the samples have been drawn should have normal distribution.

(iv) It serves as a guiding instrument in the analysis and interpretation of statistical data.

(v) It can be used for smoothing and graduating a distribution which is not normal simply by contracting a normal curve.

EXAMPLE 17.42 *If X is a normal variate with a mean of 120 and a standard deviation of 10, find c such that (i) $P(X > c) = 0.02$, and (ii) $P(X < c) = 0.05$.*

Solution: For normal variate X,

$$\mu = 120, \quad \sigma = 10$$

$$Z = \frac{X - \mu}{\sigma}$$

(i) $P(X > c) = 0.02$

$P(X < c) = 1 - P(X \geq c) = 1 - 0.02$

$= 0.98$

Since $P(X < c) > 0.5$, the corresponding value of Z will be positive.

$P(X > c) = P(Z > z_1)$ (Fig. 17.18)

$0.02 = 0.5 - P(0 \leq Z \leq z_1)$

$P(0 \leq Z \leq z_1) = 0.48$

$\therefore$ $z_1 = 2.05$ [From normal table]

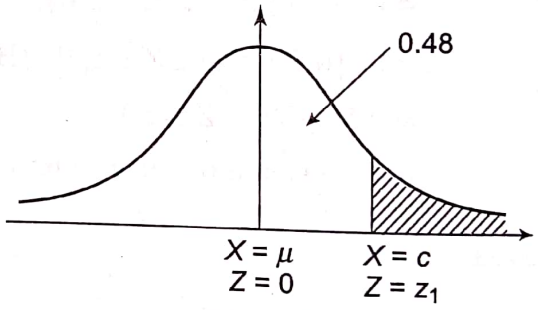


Fig. 17.18

$$Z = \frac{c - 120}{10} = z_1 = 2.05$$

$$c = 2.05(10) + 120 = 140.05$$

(ii) Since $P(X < c) < 0.5$, the corresponding value of Z will be negative.

$P(X < c) = P(Z < -z_1)$ (Fig. 17.19)

$0.05 = 1 - P(Z \geq -z_1)$

$0.05 = 1 - [0.5 + P(-z_1 \leq Z \leq 0)]$

$0.05 = 1 - [0.5 + P(0 \leq Z \leq z_1)]$ [By symmetry]

$0.05 = 0.5 - P(0 \leq Z \leq z_1)$

$P(0 \leq Z \leq z_1) = 0.5 - 0.05 = 0.45$

$\therefore$ $z_1 = -1.64$ [From normal table]

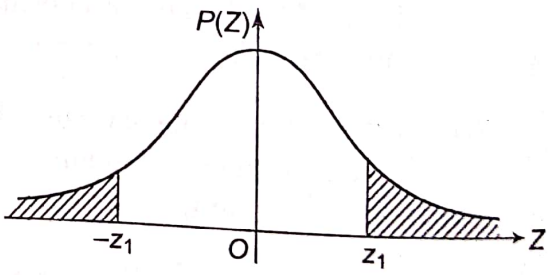


Fig. 17.19

$$Z = \frac{c - 120}{10} = z_1 = -1.64$$

$$c = 10(-1.64) + 120 = 103.6$$

EXAMPLE 17.43

The lifetime of a certain kind of batteries has a mean life of 400 hours and the standard deviation as 45 hours. Assuming the distribution of lifetime to be normal, find (i) the percentage of batteries with a lifetime of at least 470 hours, (ii) the proportion of batteries with a lifetime between 385 and 415 hours, and (iii) the minimum life of the best 5% of batteries.

Solution: Let X be the random variable which denotes the lifetime of a certain kind of batteries.

$$\mu = 400, \quad \sigma = 45$$

$$Z = \frac{X - \mu}{\sigma}$$

(i) When $X = 470$,

$$Z = \frac{470 - 400}{45} = 1.56$$

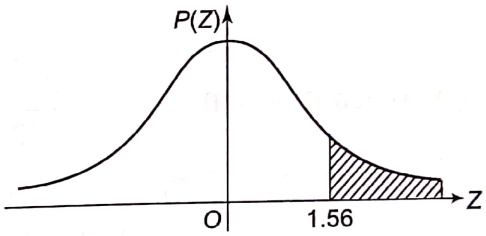


Fig. 17.20

$P(X \geq 470) = P(Z \geq 1.56)$ (Fig. 17.20)

$= 0.5 - P(0 < Z < 1.56) = 0.5 - 0.4406 = 0.0594$

Hence, the percentage of batteries with a lifetime of at least 470 hours = 5.94%.

(ii) When $X = 385$,

$$Z = \frac{385 - 400}{45} = -0.33$$

When $X = 415$,

$$Z = \frac{415 - 400}{45} = 0.33$$

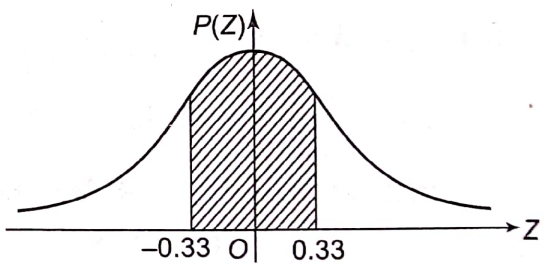


Fig. 17.21

$P(385 < X < 415) = P(-0.33 < Z < 0.33)$ (Fig. 17.21)

$= P(-0.33 < Z < 0) + P(0 < Z < 0.33)$

$= P(0 < Z < 0.33) + P(0 < Z < 0.33)$ [By symmetry]

$= 2P(0 < Z < 0.33) = 2(0.1293) = 0.2586$

Hence, the proportion of batteries with a lifetime between 385 and 415 hours = 25.86%.

(iii) $P(X > x_1) = 0.05$ (Fig. 17.22)

$P(X > x_1) = P(Z > z_1)$

$0.05 = 0.5 - P(0 \leq Z \leq z_1)$

$P(0 \leq Z \leq z_1) = 0.5 - 0.05 = 0.45$

$\therefore z_1 = 1.65$ [From normal table]

$$Z = \frac{x_1 - 400}{45} = z_1 = 1.65$$

$\therefore x_1 = 1.65(45) + 400 = 474.25$ hours

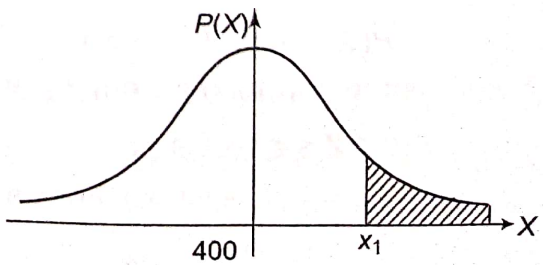


Fig. 17.22

EXAMPLE 17.44

The mean yield for a one-acre plot is 662 kg with an SD of 32 kg. Assuming normal distribution, how many one-acre plots in a batch of 1000 plots would you expect to have yields (i) over 700 kg? (ii) below 650 kg? (iii) What is the lowest yield of the best 100 plots?

Solution: Let X be the random variable which denotes the yield for the one-acre plot.

$$\mu = 662, \quad \sigma = 32, \quad N = 1000$$

$$Z = \frac{X - \mu}{\sigma}$$

(i) When $X = 700$, $Z = \dfrac{700 - 662}{32} = 1.19$

$$P(X > 700) = P(Z > 1.19) \text{ (Fig. 17.23)}$$
$$= 0.5 - P(0 \le Z \le 1.19) = 0.5 - 0.3830 = 0.1170$$

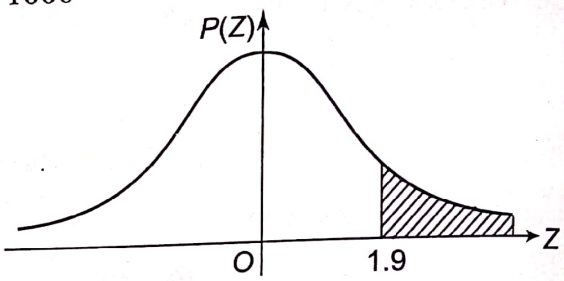


Fig. 17.23

Expected number of plots with yields over 700 kg $= N\,P(X > 700)$
$= 1000(0.1170)$
$= 117$

(ii) When $X = 650$,

$$Z = \frac{650 - 662}{32} = -0.38$$

$$P(X < 650) = P(Z < -0.38) \text{ (Fig. 17.24)}$$
$$= P(Z > 0.38) \quad \text{[By symmetry]}$$
$$= 0.5 - P(0 \le Z \le 0.38) = 0.5 - 0.1480 = 0.352$$

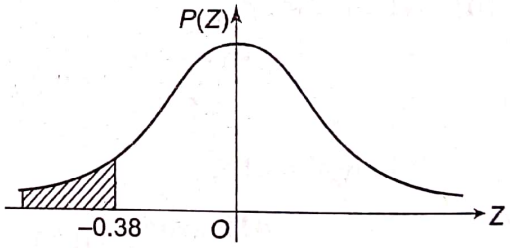


Fig. 17.24

Expected number of plots with yields below 650 kg $= N\,P(X < 650) = 1000(0.352) = 352$

(iii) The lowest yield, say, x_1 of the best 100 plots is given by

$$P(X > x_1) = \frac{100}{1000} = 0.1$$

When $X = x_1$, $Z = \dfrac{x_1 - 662}{32} = z_1$

$$P(X > x_1) = P(Z > z_1)$$
$$0.1 = 0.5 - P(0 \le Z \le z_1)$$
$$P(0 \le Z \le z_1) = 0.4$$
$$\therefore \quad z_1 = 1.28 \text{ (approx.) [From normal table]}$$
$$\frac{x_1 - 662}{32} = 1.28$$
$$x_1 = 702.96$$

Hence, the best 100 plots have yields over 702.96 kg.

EXAMPLE 17.45 *Find the mean and standard deviation in which 7% of items are under 35 and 89% are under 63.*

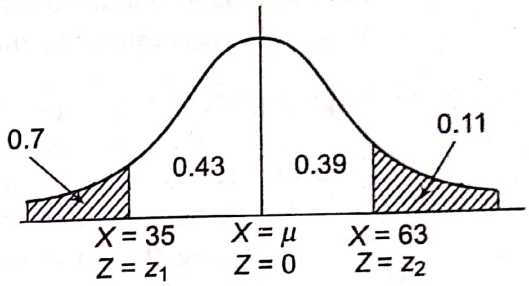


Fig. 17.25

Solution: Let μ be the mean and σ be standard deviation of the normal curve.

$$P(X < 35) = 0.07$$
$$P(X < 63) = 0.89$$
$$P(X > 63) = 1 - P(X < 63) = 1 - 0.89 = 0.11$$
$$Z = \frac{X - \mu}{\sigma}$$

Since $P(X < 35) < 0.5$, the corresponding value of Z will be negative.

When $X = 35$, $Z = \frac{35 - \mu}{\sigma} = -z_1$ (say)

Since $P(X < 63) > 0.5$, the corresponding value of Z will be positive.

When $X = 63$, $Z = \frac{63 - \mu}{\sigma} = z_2$ (say)

From Fig. 17.25,

$$P(Z < -z_1) = 0.07$$
$$P(Z > z_2) = 0.11$$
$$P(0 < Z < z_1) = P(-z_1 < Z < 0) = 0.5 - P(Z \le -z_1) = 0.5 - 0.07 = 0.43$$
$$z_1 = 1.48 \quad \text{[From normal table]}$$
$$P(0 < Z < z_2) = 0.5 - P(Z \ge z_2) = 0.5 - 0.11 = 0.39$$
$$z_2 = 1.23 \quad \text{[From normal table]}$$

Hence, $\frac{35 - \mu}{\sigma} = -1.48$

$$-1.48\,\sigma + \mu = 35 \quad \ldots(1)$$

and $\frac{63 - \mu}{\sigma} = 1.23$

$$1.23\,\sigma + \mu = 63 \quad \ldots(2)$$

Solving Eqs (1) and (2),

$$\mu = 50.29, \qquad \sigma = 10.33$$

17.15.5 Fitting a Normal Distribution

Fitting a normal distribution or a normal curve to the data means to find the equation of the curve in the form $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ which will be as close as possible to the points given. There are two purposes of fitting a normal curve:

(i) To judge the whether the normal curve is the best fit to the sample data.
(ii) To use the normal curve to estimate the characteristics of a population.

The area method for fitting a normal curve is given by the following steps:

(i) Find the mean μ and standard deviation σ for the given data if not given.

(ii) Write the class intervals and lower limits X of class intervals in two columns.

(iii) Find $Z = \dfrac{X - \mu}{\sigma}$ for each class interval.

(iv) Find the area corresponding to each Z from the normal table.

(v) Find the area under the normal curve between the successive values of Z. These are obtained by subtracting the successive areas when the corresponding Z's have the same sign and adding them when the corresponding Z's have the same sign and adding them when the corresponding Z's opposite sign.

(vi) Find the expected frequencies by multiplying the relative frequencies by the number of observations.

EXAMPLE 17.46

Fit a normal curve from the following distribution. It is given that the mean of the distribution is 43.7 *and its standard distribution is* 14.8.

Class interval	11–20	21–30	31–40	41–50	51–60	61–70	71–80
Frequency	20	28	40	60	32	20	8

Solution

$\mu = 43.7, \quad \sigma = 14.8 \quad N = \Sigma f = 200$

The series is converted into an inclusive series.

Class Interval	*Lower class*	$Z = \dfrac{X - \mu}{\sigma}$	*Area from 0 to Z*	*Area in class Interval*	*Expected Frequencies*
10.5–20.5	10.5	–2.24	0.4875	0.0457	$9.14 \approx 9$
20.5–30.5	20.5	–1.57	0.4418	0.1285	$25.7 \approx 26$
30.5–40.5	30.5	–0.89	0.3133	0.2262	$45.24 \approx 45$
40.5–50.5	40.5	–0.22	0.0871	0.2643	$52.86 \approx 53$
50.5–60.5	50.5	0.46	0.1772	0.1957	$39.14 \approx 39$
60.5–70.5	60.5	1.14	0.3729	0.092	$18.4 \approx 18$
70.5–80.5	70.5	1.81	0.4649	0.0287	$5.74 \approx 65$
	80.5	2.49	0.4936		

EXERCISE 17.13

1. If X is normally distributed with a mean and standard deviation of 4, find (i) $P(5 \le X \le 10)$, (ii) $P(X \ge 15)$, (iii) $P(10 \le X \le 15)$, and (iv) $P(X \le 5)$.

[**Ans.:** (i) 0.3345 (ii) 0.003 (iii) 0.0638 (iv) 0.4013]

2. A normal distribution has a mean of 5 and a standard deviation of 3. What is the probability that the deviation from the mean of an item taken at random will be negative?

[**Ans.:** 0.0575]

3. If X is a normal variate with a mean of 30 and an SD of 6, find the value of $X = x_1$ such that $P(X \ge x_1) = 0.05$.

[**Ans.:** 39.84]

4. If X is a normal variate with a mean of 25 and SD of 5, find the value of $X = x_1$ such that $P(X \le x_1) = 0.01$.

[**Ans.:** 11.02]

5. The weights of 4000 students are found to be normally distributed with a mean of 50 kg and an SD of 5 kg. Find the probability that a student selected at random will have weight (i) less than 45 kg, and (ii) between 45 and 60 kg.

[**Ans.:** (i) 0.1587 (ii) 0.8185]

6. The daily sales of a firm are normally distributed with a mean of ₹ 8000 and a variance of ₹ 10000. (i) What is the probability that on a certain day the sales will be less than ₹ 8210? (ii) What is the percentage of days on which the sales will be between ₹ 8100 and ₹ 8200?

[**Ans.:** (i) 0.482 (ii) 14%]

7. The mean height of Indian soldiers is 68.22″ with a variance of 10.8″. Find the expected number of soldiers in a regiment of 1000 whose height will be more than 6 feet.

[**Ans.:** 125]

8. The life of army shoes is normally distributed with a mean of 8 months and a standard deviation of 2 months. If 5000 pairs are issued, how many pairs would be expected to need replacement after 12 months?

[**Ans.:** 2386]

9. In an intelligence test administered to 1000 students, the average was 42 and the standard deviation was 24. Find the number of students (i) exceeding 50, (ii) between 30 and 54, and (iii) the least score of top 1000 students.

[**Ans.:** (i) 129 (ii) 383 (iii) 72.72]

10. In a test of 2000 electric bulbs, it was found that the life of a particular make was normally distributed with an average of life of 2040 hours and a standard deviation of 60 hours. Estimate the number of bulbs likely to burn for (i) more than 2150 hours, and (ii) less than 1950 hours.

[**Ans.:** (i) 67 (ii) 184]

11. The marks of 1000 students of a university are found to be normally distributed with a mean of 70 and a standard of deviation 5. Estimate the number of students whose marks will be (i) between 60 and 75, (ii) more than 75, and (iii) less than 68.

[**Ans.:** (i) 910 (ii) 23 (iii) 37]

12. In a normal distribution, 31% items are under 45 and 8% are over 64. Find the

mean and standard deviation. Find also, the percentage of items lying between 30 and 75.

[**Ans.:** 50, 10, 0.957]

13. Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming a normal distribution, find the mean and standard deviation of distribution.

[**Ans.:** 65.42, 3.27]

14. The marks obtained by students in an examination follow a normal distribution. If 30% of the students got marks below 35 and 10% got marks above 60, find the mean and percentage of students who got marks between 40 and 50.

[**Ans.:** 42.23, 13.88, 28%]

15. Fit a normal distribution to the following data:

Class	60–65	65–70	70–75	75–80	80–85	85–90	90–95	95–100
Frequency	3	21	150	335	326	135	26	4

[Ans.: Expected frequency: 3, 31, 148, 322, 319, 144, 30, 3]

17.16 TESTS OF HYPOTHESIS

The main purpose behind the sampling theory is the study of the Tests of Hypothesis or Tests of significance. In many situations, assumptions are made about the population parameters involved in order to arrive at decisions related to population on the basis of sample information. Such an assumption is called statistical hypothesis which may or may not be true. The procedure which enables us to decide on the basis of sample results whether a hypothesis is true or not, is called test of hypothesis or test of significance.

17.16.1 Terms Related to Tests of Hypothesis

(1) *Parameters*: The statistical constants of population such as mean (μ), standard deviation (σ), correlation coefficient (ρ), population proportion (P) etc. are called the parameters. Greek letters are used to denote the population parameters.

(2) *Statistic*: The statistical constants for the sample drawn from the given population such as mean ($\bar{x}$), standard deviation (s), correlation coefficient (r), sample proportion (p) etc., are called the statistic. Roman letters are used to denote the sample statistic.

(3) *Sampling Distribution*: Consider all possible samples of size 'n' which can be drawn from a population of size 'N'. These samples will give different values of a statistic. The means of the samples will not be identical. If these different means are arranged according to their frequencies, the frequency distribution formed is called sampling distribution of mean. Similarly, the sampling distribution of other statistics can be defined.

(4) *Standard Error*: The standard deviation of the sampling distribution of a statistic is known as its standard error SE. Standard error plays a very important role in the large sample theory and forms the basis of the testing of hypothesis.

(5) *Null Hypothesis*: Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. It is denoted by H_0. It asserts that there is no significant difference between the statistic and the population parameter and whatever observed difference exists, is merely due to the fluctuations in sampling from the same population.

(6) *Alternative Hypothesis*: Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is denoted by H_1. It is set in such a way that the rejection of null hypothesis implies the acceptance of alternative hypothesis. For example, if the null hypothesis is that the average height of the students of a college is 166 cm. i.e., $\mu_0 = 166$ cm, say then the null hypothesis is

$$H_0 : \mu = 166 (= \mu_0)$$

and the alternative hypothesis could be

(i) $H_1 : \mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)
(ii) $H_1 : \mu > \mu_0$
(iii) $H_1 : \mu < \mu_0$

Thus, there can be more than one alternative hypothesis.

(7) *Test Statistic*: After setting up the null hypothesis and alternative hypothesis, test statistic is calculated. The test statistic is a statistic based on appropriate probability distribution. It is used to test whether the null hypothesis should be accepted or rejected. Different probability distribution values are used in appropriate cases while testing the null hypothesis.
For Z-distribution under normal curve for large samples ($n > 30$), the Z-statistic is defined by

$$Z = \frac{t - E(t)}{SE(t)}$$

(8) *Errors in Hypothesis Testing*: The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. There is every chance that a decision regarding a null hypothesis may be correct or may not be correct. There are two types of errors.
(i) *Type I error*: It is the error of rejecting the null hypothesis H_0, when it is true. It occurs when a null hypothesis is true, but the difference of means is significant and the hypothesis is rejected. If the probability of making a type I error is denoted by α, the level of significance, then the probability of making a correct decision is $(1 - \alpha)$.
(ii) *Type II error*: It is the error of accepting the null hypothesis H_0, when it is false. It occurs when a null hypothesis is false, but the difference of means is insignificant and the hypothesis is accepted. The probability of making a type II error is denoted by β.

(9) *Level of Significance*: The level of significance is the maximum probability of making a type I error and is denoted by α, i.e., P[Rejecting H_0 when H_0 is true] = α. The commonly used level of significance in practice are 5% (0.05) and 1% (0.01). For 5% level of significance ($\alpha = 0.05$), the probability of making type I error is 0.05 or 5% i.e., P[Rejecting H_0 when H_0 is true] = 0.05. This means that there is a probability of making 5 out of 100 type I error. Similarly, 1% level of significance ($\alpha = 0.01$) means that there is a probability of making 1 error out of 100. If no level of significance is given, α is taken as 0.05.

(10) *Critical Region*: The critical region or rejection region is the region of the standard normal curve corresponding to a predetermined level of significance α. The region under the normal curve which is not covered by the rejection region is known as acceptance region. Thus, the statistic which leads to rejection of null hypothesis H_0 gives rejection region or critical region. The value of the test statistic calculated to test the null hypothesis H_0 is known as critical value. Thus, the critical value separates the rejection region from the acceptance region.

(11) *Two Tailed Test and One Tailed Test:* When the test of hypothesis is made on the basis of rejection region represented by both the sides of the standard normal curve, it is called a two tailed test. A test of statistical hypothesis, where the alternative hypothesis H_1 is two sided or two tailed such as:

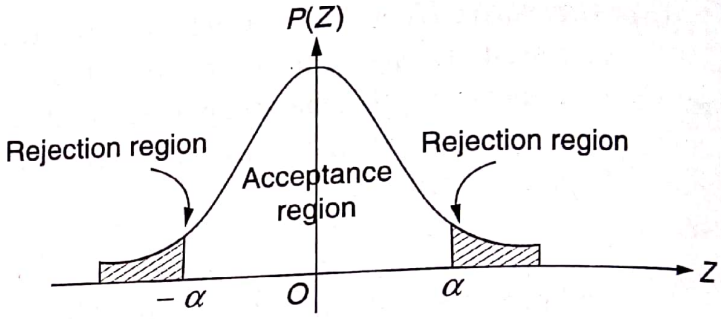


Fig 17.26 Two tailed test

Null Hypothesis $H_0 : \mu = \mu_0$

Alternative Hypothesis $H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$),

is called two tailed test or two sided test.

A test of statistical hypothesis, where the alternative hypothesis is one sided is called one tailed test or one sided test. There are two types of one tailed tests.

(i) Right Tailed Test: In the right tailed test, the rejection region or critical region lies entirely on the right tail of the normal curve (Fig 17.27).

(ii) Left Tailed Test: In the left tailed test, the rejection region or critical region lies entirely on the left tail of the normal curve.

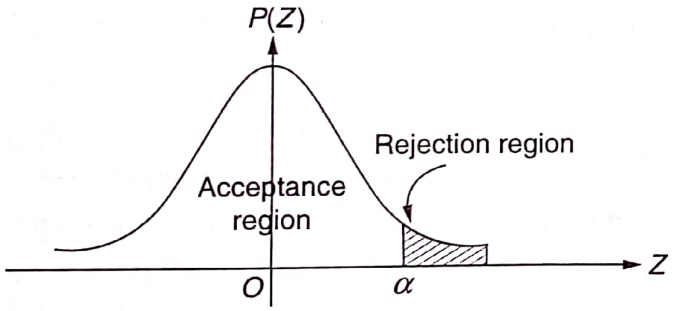


Fig 17.27 Right tailed test

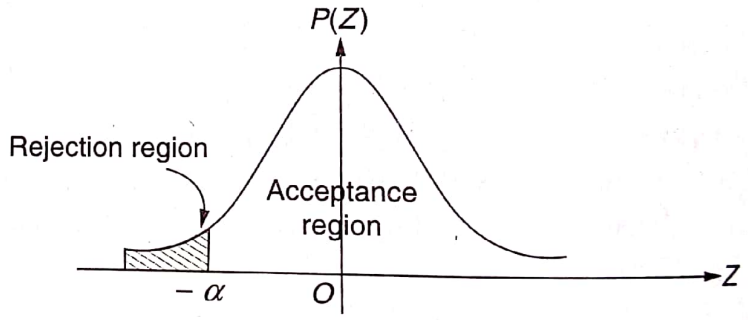


Fig 17.28 Left tailed test

For example, in a test for testing the mean (μ) of the population

Null Hypothesis $H_0 : \mu = \mu_0$

Alternative Hypothesis $H_1 : \mu > \mu_0$ (Right tailed)

$\mu < \mu_0$ (Left tailed)

A two tailed test is applied in such cases when the difference between the sample mean and population mean is tending to reject the null hypothesis H_0, the difference may be positive or negative.

A one tailed test is applied in such cases when the population mean is at least as large as some specified value of the mean (right tailed test) or at least as small as some specified value of the mean (left tailed test).

Critical value (Z_α)	*Level of significance α*		
	1%	5%	10%
Two tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

(12) *Confidence Limits:* The limits within which a hypothesis should lie with specified probability are called confidence limits or fiducial limits. Generally, the confidence limits are set up with 5% or 1% level of significance. If the sample value lies between the confidence limits, the hypothesis is accepted, if it does not, then the hypothesis is rejected at the specified level of significance. Suppose that the sampling distribution of a statistic S is normal with mean μ and standard deviation σ. The sample statistic S can be expected to lie in the interval $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ for 95% times (Fig 17.29). Because of this, $(S - 1.96\sigma, S + 1.96\sigma)$ is called the 95% confidence interval for estimation of μ. The ends of this interval, i.e., $S \pm 1.96\sigma$ are called 95% confidence limits for S. Similarly, $S \pm 2.58\sigma$ are 99% confidence limits. The numbers 1.96, 2.58 etc. are called confidence coefficients.

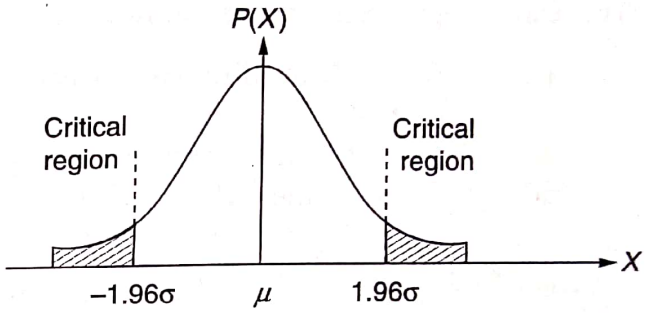


Fig 17.29 Confidence Limits

17.16.2 Procedure for Testing of Hypothesis

The various steps in testing of a statistical hypothesis are as follows:

(i) *Null Hypothesis*: Set up the Null Hypothesis H_0

(ii) *Alternative Hypothesis*: Set up the Alternative Hypothesis H_1.
This will decide the use of single-tailed (right or left) or two-tailed test.

(iii) *Level of Significance*; Select the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. If no level of significance is given α is selected as 0.05.

(iv) *Test Statistic*: Calculate the test statistic

$$Z = \frac{t - E(t)}{SE(t)} \text{ under } H_0$$

(v) *Critical Value*: Find the significant value (tabulated value) Z_α of Z at the given level of significance α.

(vi) *Decision*: Compare the calculated value of Z with the tabulated value Z_α. If $|Z| < Z_\alpha$ i.e., if the calculated value of Z is less than tabulated value Z_α at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ i.e., if the calculated value of Z is more than tabulated value Z_α at the level of significance α, the null hypothesis is rejected.

17.17 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

If a sample consists of more than 30 items, i.e., $n > 30$, it is considered as large sample. The following assumptions are applied for significance tests of large samples:

(i) The random sampling distribution of statistic has the properties of the normal curve.

(ii) Values (i.e., statistic) given by the samples are sufficiently close to the populate values (i.e., parameters) and can be used in its place for calculating the standard error (SE) of the estimate.

For example, if SD of the population is not known, SE can be calculated by SD of the sample.

Suppose the hypothesis to be tested is that the probability of success in such trail is p. Assuming it to be true, the mean μ and the standard deviation σ of the sampling distribution of the number of successes

are np and $\sqrt{npq}$ respectively as the sampling distribution of number of successes follows a binomial probability distribution.

If x is the observed number of successes in the sample and Z is the standard normal variate then

$$Z = \frac{x - \mu}{\sigma}$$

The tests of significance are as follows:

(i) If $|Z| < 1.96$, the difference between the observed and expected number of successes is not significant.
(ii) If $|Z| > 1.96$, the difference is significant at 5% level of significance.
(iii) If $|Z| > 2.58$, the difference is significant at 1% level of significance.

EXAMPLE 17.47

A coin was tossed 960 times and returned heads 183 times. Test the hypothesis that the coin is unbiased. Use a 0.05 level of significance.

Solution:

$$n = 960$$

$$p = \text{probability of getting head} = \frac{1}{2}$$

$$q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$$

$$\mu = np = 960\left(\frac{1}{2}\right) = 480$$

$$\sigma = \sqrt{npq} = \sqrt{960 \times \frac{1}{2} \times \frac{1}{2}} = 15.49$$

$$x = \text{number of sucesses} = 183$$

(i) Null Hypothesis H_0: The coin is unbiased.
(ii) Alternative Hypothesis H_1: The coin is biased.
(iii) Level of significance: $\alpha = 0.05$
(iv) Test statistic: $Z = \dfrac{x - \mu}{\sigma} = \dfrac{183 - 480}{15.49} = -19.17$

$$|Z| = 19.17$$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$
(vi) Decision: Since $|Z| > 1.96$, the null hypothesis H_0 is rejected at 5% level of significance i.e., the coin is biased.

EXAMPLE 17.48

A dice is tossed 960 times and it falls with 5 upwards 184 times. Is the dice unbiased at a level of significance of 0.01?

Solution:

$$n = 960$$

$$p = \text{Probability of throwing 5 with one die} = \frac{1}{6}$$

$$q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}$$

$$\mu = np = 960\left(\frac{1}{6}\right) = 160$$

$$\sigma = \sqrt{npq} = \sqrt{960 \times \frac{1}{6} \times \frac{5}{6}} = 11.55$$

x = number of successes = 184

(i) Null Hypothesis H_0: The dice is unbiased

(ii) Alternative Hypothesis H_1: The dice is biased

(iii) Level of significance: $\alpha = 0.01$

(iv) Test statistic: $Z = \dfrac{x - \mu}{\sigma} = \dfrac{184 - 160}{11.55} = 2.08$

(v) Critical value: Tabulated value of Z at 1% level of significance is $Z_{0.05} = 2.58$

(vi) Decision: Since $|Z| < 2.58$, the null hypothesis H_0 is accepted at 1% level of significance, i.e., the dice is unbiased.

17.17.1 Test of Significance of a Single Mean-Large Samples

Let a random sample size n ($n > 30$) has the sample mean $\bar{x}$ and population has the mean μ.

Working Rule:

(i) Null Hypothesis $H_0 : \bar{x} = \mu$, i.e., there is no significance difference between the sample mean and population mean or the sample has been drawn from the parent population.

(ii) Alternative Hypothesis $H_1 : \bar{x} \neq \mu$.

(iii) Level of significance: Select the level of significance α

(iv) Test statistic: There are two cases for calculating a test statistic Z

(a) When the standard deviation σ of population is known

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(b) When the standard deviation σ of population is not known

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where s is the sample SD.

(v) Critical value: Find the critical value (tabulated value) Z_α of Z at the given level of significance α.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

EXAMPLE 17.49

A random sample of 100 Indians has an average life span of 71.8 years with standard deviation of 8.9 years. Can it be concluded that the average life span of an Indian is 70 years?

Solution: $n = 100$, $\bar{x} = 71.8$, $\mu = 70$, $s = 8.9$

(i) Null Hypothesis H_0: $\mu = 70$ years i.e., the average life span of an Indian is 70 years.

(ii) Alternative Hypothesis H_1: $\mu \neq 70$

(iii) Level of Significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}} = \dfrac{71.8 - 70}{\dfrac{8.9}{\sqrt{100}}} = 2.02$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $|Z| > 1.96$, the null hypothesis is rejected at 5% level of significance, i.e, the average life span of an Indian is not 70 years.

EXAMPLE 17.50

A sample of 900 members has a mean of 3.4 cm and SD 2.61 cm. Is the sample from a large population of mean 3.25 cm and SD 2.61 cm? If the population is normal and its mean is unknown, find the 95% and 98% fiducial limits of its true mean.

Solution: $n = 900$, $\bar{x} = 3.4$ cm, $s = 2.61$ cm, $\mu = 3.25$ cm, $\sigma = 2.61$ cm

(i) Null Hypothesis H_0: The sample has been drawn from the population with mean $\mu = 3.25$ cm and $\sigma = 2.61$ cm.

(ii) Alternative Hypothesis H_1: $\mu \neq 3.25$ cm

(iii) Level of significance : $\alpha = 0.05$

(iv) Test statistic : $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{3.4 - 3.25}{\dfrac{2.61}{\sqrt{900}}} = 1.72$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $Z < 1.96$, the null hypothesis H_0 is accepted at 5% level of significance i.e., the sample has been drawn from the population with mean $\mu = 3.25$ cm.

95% fudicial limits:

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right) = 3.4 \pm 1.96\left(\frac{2.61}{\sqrt{900}}\right) = 3.4 \pm 0.1705,$$

i.e., 3.5705 and 3.2295

98% fudicial limits:

$$\bar{x} \pm 2.33\left(\frac{\sigma}{\sqrt{n}}\right) = 3.4 \pm 2.33\left(\frac{2.61}{\sqrt{900}}\right) = 3.4 \pm 0.2027,$$

i.e., 3.6027 and 3.1973.

17.17.2 Test of Significance of Difference between Two Means – Large Samples

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means of two independent large random samples with sizes n_1 and n_2 ($n_1 > 30$, $n_2 > 30$) drawn from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2.

Working Rule

(i) Null Hypothesis H_0: $\mu_1 = \mu_2$, i.e., the two samples have been drawn from two different populations having the same means and equal standard deviations or $H_0 : \bar{x}_1 = \bar{x}_2$, i.e., the two samples have been drawn from the same parent population.

(ii) Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ (two tailed test)
or H_1: $\mu_1 < \mu_2$ (one tailed test)
or H_1: $\mu_1 > \mu_2$ (one tailed test)

(iii) Level of significance: Select the level of significance α

(iv) Test statistic: These are two cases for calculating test statistic.

(a) When the population standard deviations σ_1 and σ_2 are known

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(b) When the population standard deviations σ_1 and σ_2 are not known

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1 and s_2 are sample standard deviations.

(v) Critical Value: Find the critical value (tabulated value) Z_α of Z at the given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

EXAMPLE 17.51

Test the significance of the difference between the means of two normal population with the same standard deviation from the following data:

	Size	Mean	SD
Sample I	100	64	6
Sample II	200	67	8

Solution: $n_1 = 100, \quad n_2 = 200, \quad \bar{x}_1 = 64, \quad \bar{x}_2 = 67 \quad s_1 = 6, \quad s_2 = 8$

(i) Null Hypothesis $H_0 : \mu_1 = \mu_2$

(ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \dfrac{64 - 67}{\sqrt{\frac{36}{200} + \frac{64}{100}}} = -3.31$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $|Z| = 3.31 > 1.16$, the null hypothesis H_0 is rejected at 5% level of significance. Hence, the samples do not support the hypothesis that the two population have the same mean although they may have the same standard deviation.

EXAMPLE 17.52

The means of simple samples of sizes 1000 and 2000 are 67.5 and 68 cm respectively. Can the samples be regarded as drawn from the same population of S.D. 2.5 cm.

Solution: $n_1 = 1000, n_2 = 2000, \bar{x}_1 = 67.5$ cm, $\bar{x}_2 = 68$ cm, $\sigma = 2.5$ cm

(i) Null Hypothesis H_0: The samples have been from the same population of S.D. 2.5 cm, i.e., $\mu_1 = \mu_2$ and $\sigma = 2.5$ cm

(ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}} = \dfrac{67.5 - 68}{\sqrt{\dfrac{(2.5)^2}{1000} + \dfrac{(2.5)^2}{2000}}} = -5.16$

$|Z| = 5.16$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(iv) Decision: Since $|Z| = 5.16 > 1.96$, the null hypothesis H_0 is rejected at 5% level of significance. Hence, the samples cannot be regarded as drawn from the same population of SD 2.5 cm.

EXAMPLE 17.53

The mean life of a sample of 10 electric bulbs was found to be 1456 hours with SD of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 with SD of 398 hours. Is there a significant difference between the means of two batches?

Solution: $n_1 = 10$, $n_2 = 17$, $\bar{x}_1 = 1456$, $\bar{x}_2 = 1280$ $s_1 = 423$, $s_2 = 398$

(i) Null Hypothesis $H_0 : \mu_1 = \mu_2$, i.e., there is no significant difference between the means of two batches.

(ii) Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

(iii) Level of significance: $\alpha = 0.05$ (assumption)

(iv) Test statistic: $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{1456 - 1280}{\sqrt{\dfrac{(423)^2}{10} + \dfrac{(398)^2}{17}}} = 1.07$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $Z < 1.96$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference between the means of two batches.

EXERCISE 17.14

1. A random sample of 100 students gave a mean weight of 58 kg with a SD of 4 kg. Test the hypothesis that the mean weight in the population is 60 kg.

[**Ans.:** The mean weight in the population is not 60 kg]

2. A sample of 400 items is taken from a normal population whose mean is 4 and whose variance is also 4. If the sample mean is 4.45, can the sample be regarded as truly random sample?

[**Ans.:** Sample cannot be regarded as truly random sample]

3. The mean IQ of a sample of 1600 children was 99. Is it likely that this was a random sample from a population with mean IQ 100 and SD 15?

[**Ans.:** Sample was not drawn from a population with mean 100 and SD 15]

4. In a random sample of 60 workers, the average time taken by them to get to work

is 33.8 minutes with a standard deviation of 6.1 minutes. Can we reject the null hypothesis $\mu = 32.6$ minutes in favour of alternative hypothesis $\mu > 32.6$ at $\alpha = 0.025$ level of significance

[**Ans.:** The null hypothesis is accepted]

5. It is claimed that a random sample of 49 types has a mean life of 15200 km. This sample was drawn from a population whose mean is 15150 km and a standard deviation of 1200 km. Test the significance at 0.05 level.

[**Ans.:** The null hypothesis is accepted]

6. An ambulance service claims that it takes on the average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the claim at 0.05 level of significance.

[**Ans.:** The null hypothesis is accepted]

7. Samples of students were drawn from two universities and from their weights in kilograms, the mean and standard deviations are calculated. Make a large sample test to test the significance of the difference between the means.

	Mean	*SD*	*Size of the sample*
University A	55	10	400
University B	57	15	100

[**Ans.:** There is no significant difference between the means]

8. A researchers wants to know the intelligence of students in a school. He selected two groups of students. In the first group, there are 150 students having mean IQ of 75 with a SD of 15. In the second group there are 250 students having mean IQ of 70 with SD of 20. Test the significance that the groups have come from same population.

[**Ans.:** The groups have not come from same population]

9. Random samples drawn from two places gave the following data relating to the heights of children:

	Mean height in cm	*SD in cm*	*No. of children in sample*
Place A	68.50	2.5	1200
Place B	68.58	3.0	1500

Test at 5% level of significance that the mean height is the same for children at two places.

[**Ans.:** The mean height is same for children at two places]

10. The mean life of a sample of 10 electric bulbs was found to be 1456 hours with SD of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with SD of 398 hours. Is there a significant difference between the means of two batches?

[**Ans.:** There is no difference between the mean life of two batches]

17.17.3 Test of Significance for Single Proportion – Large Samples

Let p be the sample proportion in a large random sample of size n drawn from a population having proportion P.

Working Rule

(i) Null Hypothesis $H_0: P = P_0$

(ii) Alternative Hypothesis $H_1: P \neq P_0$ (i.e., $P > P_0$ or $P < P_0$)
or $H_1: P > P_0$
or $H_1: P < P_0$

(iii) Level of significance: Select the level of significance α

(iv) Test statistic: $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$, where $Q = 1 - P$

(v) Critical Value: Find the critical value (tabulated value) Z_α of Z at the given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance α, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance α, the null hypothesis is rejected.

EXAMPLE 17.54

A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specification. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance.

Solution: $n = 200$

Number of pieces conforming to specification = 200 – 18 = 182

p = Sample proportion of pieces conforming to specification $= \dfrac{182}{200} = 0.91$

P = Population proportion of pieces conforming to specification $= \dfrac{95}{100} = 0.95$

$Q = 1 - P = 1 - 0.95 = 0.05$

(i) *Null Hypothesis* H_0: $P = 0.95$ i.e., the proportion of pieces conforming to proportion is 95%

(ii) *Alternating Hypothesis* H_1: $P < 0.95$ (left tailed test)

(iii) *Level of significance*: $\alpha = 0.05$

(iv) Test statistic: $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.91 - 0.95}{\sqrt{\dfrac{(0.95)(0.05)}{200}}} = -2.59$

$|Z| = 2.59$

(v) Critical value: Tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.645$

(vi) Decision: Since $|Z| > 1.645$, the null hypothesis is rejected at 5% level of significance i.e., the manufacturer's claim is rejected.

EXAMPLE 17.55

In a hospital 480 female and 520 male babies were born in a week. Do these figures confirm the hypothesis that males and females were born in equal numbers?

Solution: n = Total number of births = 480 + 520 = 1000

p = Sample proportion of females born $= \dfrac{480}{100} = 0.48$

P = Population proportion of females born = 0.5

$Q = 1 - P = 1 - 0.5 = 0.5$

(i) Null Hypothesis H_0: $P = 0.5$ i.e., the males and females were born in equal numbers.

(ii) Alternative Hypothesis H_1: $P \neq 0.5$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0.48 - 0.5}{\sqrt{\dfrac{(0.5)(0.5)}{1000}}} = -1.265$

$|Z| = 1.265$

(v) Critical value: The tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $|Z| < 1.96$, the null hypothesis is accepted at 5% level of significance, i.e., males and females were born in equal proportions.

17.17.4 Test of Significance of Difference between Two Sample Proportions – Large Samples

Let p_1 and p_2 be the sample proportions in two large samples of sizes n_1 and n_2 drawn from two populations having proportions P_1 and P_2.

Working Rule

(i) Null Hypothesis H_0: $P_1 = P_2$

(ii) Alternative Hypothesis H_1: $P_1 \neq P_2$
or H_1: $P_1 > P_2$
or H_1: $P_1 < P_2$

(iii) Level of significance: Select level of significance α

(iv) Test statistic: There are two cases:

(a) When the population proportions P_1 and P_2 are known

$$Z = \frac{P_1 - P_2}{\sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}}$$

(b) When the population proportions P_1 and P_2 are not known but sample proportions p_1 and p_2 are known

There are two methods to estimate P_1 and P_2.
Method of Substitution: In this method, sample proportions p_1 and p_2 are substituted for P_1 and P_2.

$$Z = \frac{p_1 - p_2}{\sqrt{\dfrac{p_1q_1}{n_1} + \dfrac{p_2q_2}{n_2}}}$$

Method of pooling: In this method, the estimated value of two population proportions is obtained by pooling the two sample proportions p_1 and p_2 into a single proportion p.

$$p = \frac{n_1p + n_2p_2}{n_1 + n_2}$$

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

(v) Critical value: Find the critical value (tabulated value) of Z at given level of significance.

(vi) Decision: If $|Z| < Z_\alpha$ at the level of significance, the null hypothesis is accepted. If $|Z| > Z_\alpha$ at the level of significance, the null hypothesis is rejected.

EXAMPLE 17.56

Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence ·200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal are same at 5% level of significance.

Solution: $n_1 = 400, n_2 = 600$

p_1 = Proportion of men $= \frac{200}{400} = 0.5$

p_2 = Proportion of women $= \frac{325}{600} = 0.541$

(i) Null Hypothesis H_0: $p_1 = p_2$, i.e., there is no significant difference in proportion of men and women in favour of the proposal.

(ii) Alternative Hypothesis is H_1: $p_1 \neq p_2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{400(0.5) + 600(0.541)}{400 + 600} = 0.525$

$$q = 1 - p = 1 - 0.525 = 0.475$$

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.5 - 0.541}{\sqrt{(0.525)(0.475)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.28$$

$$|Z| = 1.28$$

(v) Critical value: The tabulated value of Z at 5% level of significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $|Z| < 1.96$, the null hypothesis is accepted at 5% level of significance, i.e., there is no significant difference of opinion between men and women in favour of the proposal.

EXAMPLE 17.57

In two large populations, there are 30% and 25% fair haired people respectively. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution: $n_1 = 1200, n_2 = 900$

P_1 = Proportion of faired people in the first population $= \frac{30}{100} = 0.3$

$Q_1 = 1 - P_1 = 1 - 0.3 = 0.7$

P_2 = Proportion of faired people in the second population $= \frac{25}{100} = 0.25$

$Q_2 = 1 - P_2 = 1 - 0.25 = 0.75$

(i) Null Hypothesis H_0: $P_1 = P_2$, i.e., the difference in population proportions is likely to be hidden in sampling.

(ii) Alternative Hypothesis H_1: $P_1 \neq P_2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} = \frac{0.3 - 0.25}{\sqrt{\frac{(0.3)(0.7)}{1200} + \frac{(0.25)(0.75)}{900}}} = 2.56$$

(v) Critical value: The tabulated value of Z at 5% level of Significance is $Z_{0.05} = 1.96$

(vi) Decision: Since $|Z| > 1.96$, the null hypothesis is rejected at 5% level of significance, i.e., the difference in population proportions is not likely to be hidden in sampling.

EXERCISE 17.15

1. A manufacturer claims at least 95% of the items he produces are failure free. Examinations of a random sample of 600 items showed 39 to be defective. Test the claim at a significance level of 0.05.

[**Ans.:** Claim is rejected]

2. In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however, claim that only 5% of their product is defective. Is the claim tenable?

[**Ans.:** Claim is rejected]

3. A sample of 600 persons selected at random from a large city shows that the percentage of male in the sample is 53%. It is believed that male to the total population ratio in the city is $\frac{1}{2}$. Test whether this belief is confirmed by the observation.

[**Ans.:** Belief is confirmed by the observation]

4. In a sample of 1000 people in Karnataka, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

[**Ans.:** Both rice and wheat are equally popular in state]

5. In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

[**Ans.:** Majority of men in the city are smokers]

6. A dice was thrown 400 times and 'six' resulted 80 times. Do the data justify the hypothesis of an unbiased dice.

[**Ans.:** The dice is unbiased]

7. In a random sample of 125 cold drinkers, 68 said they prefer 'Thumsup' to Pepsi'. Test the null hypothesis $P = 0.5$ against the alternative hypothesis $P > 0.5$.

[**Ans.:** Null hypothesis is accepted]

8. A social worker believes that fewer than 25% of the couples in a certain area have ever used any form of birth control. A random sample of 120 couples was contacted. Twenty of them said they have used. Test the belief of the social worker at 0.05 level.

[**Ans.:** Belief of the social worker is true]

9. 20 people were attacked by a disease and only 18 survived. Will you reject the hypothesis that the survival rate is attacked by this disease is 85% in favour of the hypothesis that is more at 5% level?

[**Ans.:** The hypothesis is accepted]

10. In a study designed to investigate whether certain detonators used with explosives in coal mining meet the requirement that at least 90% will ignite the explosive when charged. It is found that 174 of the 200 detonators function properly. Test the null hypothesis $P = 0.9$ against the alternative hypothesis $P \neq 0.9$ at the 0.05 level of significance.

[**Ans.:** The null hypothesis is accepted]

11. A manufacturer of electronic equipment subjects samples of two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of second kind fail the test, what can be conclude at the level of significance $\alpha = 0.05$ about the difference between the corresponding sample proportion?

[**Ans.:** The difference between the proportions is not significant]

12. On the basis of their total scores, 200 candidates of a civil service examination are divided into two groups, the upper 30% and the remaining 70%. Consider the first question of the examination. Among the first group, 40 had the correct answer, whereas among the second group, 80 had the correct answer. On the basis of these results, can one conclude that the first question is not good at discriminating ability of the type being examined here?

[**Ans.:** The first question is good enough at discriminating ability of the type being examined]

13. A company wanted to introduce a new plan of work and a survey was conducted for this purpose. Out of sample of 500 workers in one group, 62% favoured the new plan and another group of sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level of significance?

[**Ans.:** There is no significant difference between the two groups in their attitude towards the new plan]

14. In a random sample of 1000 persons from town A, 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town A and town B, so far as the proportion of wheat consumers is concerned?

[**Ans.:** There is significant difference between town A and town B as the proportion of wheat consumers is concerned]

15. 100 articles from a factory are examined and 10 are found to be defective. Out of 500 similar articles from a second factory 15 are found to be defective. Test the significance between the difference of two proportions at 5% level.

[**Ans.:** There is a significant difference between the two proportions]

17.18 SMALL SAMPLE TESTS

If the samples are large ($n > 30$) then the sampling distribution of a statistic is normal. But if the samples are small ($n \leq 30$) then the above result does not hold good. For estimation of the parameter as well as for testing a hypothesis, following distributions are used:

(i) Student's t-distribution
(ii) Chi-square (χ^2) distribution
(iii) Snedecer's F-distribution
(iv) Fisher's z-distribution

17.19 STUDENT'S *t*-DISTRIBUTION

The theory of small or exact sample was developed by Irish statistician William S. Gosset who used to write under pen-name of student. The quantity t is defined as

$$t = \frac{\text{Difference of population parameter and the corresponding statistic}}{\text{Standard error of statistic}}$$

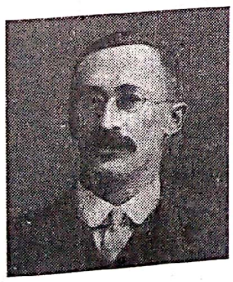
with $(n-1)$ degrees of freedom if the sample size is n.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n ($n \le 30$) drawn from a normal population with mean μ and SD σ. The student's t statistic is defined by

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{or} \quad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

where $\bar{x}$ is sample mean and $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ is an unbiased estimate of σ^2. The test statistic $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$ is a random variable having t-distribution with $v = n - 1$ degrees of freedom and with probability density function $f(t) = c\left(1 + \frac{t^2}{v}\right)^{\frac{-(v+1)}{2}}$, where $v = n-1$ and c is a constant required to make the area under the curve unity, i.e., $\int_{-\infty}^{\infty} f(t)dt = 1$. The t-distribution is used when (i) the sample size is less than or equal to 30, and (ii) population standard deviation is not known.

HISTORICAL DATA



William Sealy Gosset (1876–1937) was an English statistician. He published under the pen name Student, and developed the Student's *t*-distribution. He was born in Canterbury, England to Agnes Sealy Vidal and Colonel Frederic Gosset. He attended Winchester College before reading chemistry and mathematics at New College, Oxford. Upon graduating in 1899, he joined the brewery of Arthur Guinness & Son in Dublin, Ireland. He died in 1937 Beaconsfield, England, of a heart attack.

17.19.1 Assumptions for *t*-test

(i) Samples are drawn from normal population and are random.
(ii) The population standard deviation may not be known.
(iii) For testing the equality of two population mean, the population variances are regarded as equal.
(iv) In case of two samples, some adjustments in degrees of freedom for t are made.

17.19.2 Properties of *t*-distribution

(i) The t-distribution is asymptotic to the x- axis, i.e., it extends to infinity on either side.
(ii) The t-distribution is symmetrical about the mean.
(iii) The shape of the curve varies with the degrees of freedom.
(iv) The larger the number of degrees of freedom, the more closely t-distribution resembles standard normal distribution.

(v) Sampling distribution of t does not depend on population parameter but it depends only on degree of freedom v, i.e., on the sample size.

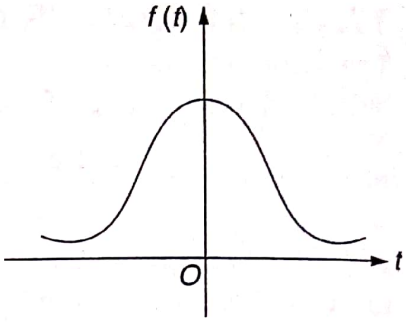


Fig 17.30 t-distribution curve

17.19.3 Applications of t-distribution

The t-distribution has following applications in testing of hypotheses for small samples:

(i) To test the significance of the sample mean, when the population variance σ is not known

(ii) To test the significance of the mean of the sample i.e., to test if the sample mean differs significantly from the population mean

(iii) To test the significance of the difference between two sample means, the population variances being equal and unknown

(iv) To test the significance of an observed sample correlation coefficient

17.19.4 Test of Significance of a Single Mean

If $x_1, x_2, \ldots, x_n$ is a random sample of size n ($n \leq 30$) drawn from a normal population with mean μ and SD σ and if the sample mean $\bar{x}$ differs significantly from the population mean μ then the student's t statistic is given by

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}, \quad \text{where} \quad s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

EXAMPLE 17.58

A machinist is making engine parts with axle diameter of 0.7 cm. A random sample of 10 parts shows a mean diameter of 0.742 cm with a standard deviation of 0.04 cm. Compute the statistic you would use to test whether work is meeting the specification at 0.05 level of significance.

Solution: $n = 10$, $\bar{x} = 0.742$ cm, $s = 0.04$ cm $\mu = 0.7$ cm

(i) Null Hypothesis H_0: $\mu = 0.7$ cm, i.e., the product is meeting the specification.

(ii) Alternative Hypothesis H_1: $\mu \neq 0.7$ cm

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $t = \dfrac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n-1}}\right)} = \dfrac{0.742 - 0.7}{\left(\frac{0.04}{\sqrt{10-1}}\right)} = 3.15$

Calculated value of $t = 3.15$

(v) Critical value: Degree of freedom $v = n - 1 = 10 - 1 = 9$
Tabulated value of t at 0.05 level of significance for 9 degrees of freedom is $t_{0.05} = 2.262$

(vi) Decision: Since calculated value of t is more than tabulated value of t, the null hypothesis is rejected at 0.05 level of significance i.e., the product is not meeting the specification.

EXAMPLE 17.59

Ten objects are chosen at random from a large population and their weights are found to be in grams: 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Discuss the suggestion that the mean weight is 65 g.

Solution: $n = 10$, $\mu = 65$ g

$$\bar{x} = \frac{\Sigma x}{n} = \frac{670}{10} = 67$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
63	–4	16
63	–4	16
64	–3	9
65	–2	4
66	–1	1
69	2	4
69	2	4
70	3	9
70	3	9
71	4	16
$\Sigma x = 670$		$\Sigma(x-\bar{x})^2 = 88$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{88}{10}} = 2.966$$

(i) Null Hypothesis H_0: $\mu = 65$, i.e., there is no significant difference in the mean weight of sample and population.

(ii) Alternate Hypothesis H_1: $\mu \neq 65$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $t = \dfrac{\bar{x}-\mu}{\left(\dfrac{s}{\sqrt{n-1}}\right)} = \dfrac{67-65}{\left(\dfrac{2.966}{\sqrt{10-1}}\right)} = 2.023$

Calculated value of $t = 2.023$

(v) Critical value: Degree of freedom $v = n - 1 = 10 - 1 = 9$
Tabulated value of t at $\alpha = 0.05$ for 9 degrees of freedom is $t_{0.05} = 2.262$

(vi) Decision: Since the calculated value of t is less than tabulated value of t, the null hypothesis is accepted at 5% level of significance, i.e., the mean weight is 65.

17.19.5 Test of Significance for Difference of Means

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be two independent samples of sizes n_1 and n_2 ($n_1 \leq 30$, $n_2 \leq 30$) with means $\bar{x}$ and $\bar{y}$ and standard deviations s_1 and s_2 from a normal population with means μ_1 and μ_2 and same standard deviations. The student's t statistic is given by

$$t = \frac{\bar{x}-\bar{y}}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$$

$$\bar{x} = \frac{\Sigma\bar{x}}{n_1}$$

where

$$\bar{y} = \frac{\Sigma \bar{y}}{n_2}$$

and

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2}{n_1 + n_2 - 2}}$$

In terms of standard deviations s_1 and s_2.

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$$

and

$$s_1 = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n_1}}$$

$$s_2 = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n_2}}$$

EXAMPLE 17.60

Ten soldiers participated in a shooting completion in the first week. After intensive training they participated in the competition in the second week. Their scores before and after training are given as follows:

Scores before training	67	24	57	55	63	54	56	68	33	43
Scores after training	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soldiers have been benefited by the training?

Solution: $n_1 = 10, n_2 = 10$

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{520}{10} = 52$$

$$\bar{y} = \frac{\Sigma y}{n_2} = \frac{570}{10} = 57$$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
67	15	225	70	13	169
24	−28	784	38	−19	361
57	5	25	58	1	1
55	3	9	58	1	1
63	11	121	56	−1	1
54	2	4	67	10	100
56	4	16	68	11	121
68	16	256	75	18	324
33	−19	361	42	−15	225
43	−9	81	38	−19	361
$\Sigma x = 520$		$\Sigma(x-\bar{x})^2 = 1882$	$\Sigma y = 570$		$\Sigma(y-\bar{y})^2 = 1664$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{1882+1664}{10+10-2}} = 14.0357$$

(i) Null Hypothesis $H_0 : \mu_1 = \mu_2$, i.e., there is no benefit of training.

(ii) Alternative Hypothesis $H_1 : \mu_1 < \mu_2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{52-57}{14.0357\sqrt{\dfrac{1}{10} + \dfrac{1}{10}}} = -0.796$

Calculated value of $|t| = 0.796$

(v) Critical value: Degrees of freedom $\nu = n_1 + n_2 - 2 = 18$

Tabulated value of t at 5% level of significance at 18 degrees of freedom is $t_{0.05} = -1.734$

$$|t_{0.05}| = 1.734$$

(vi) Decision: Since calculated value of $|t|$ is less than tabulated value of $|t_{0.05}|$, the null hypothesis is accepted at 5% level of significance, i.e., the soldiers have been benefited by the training.

EXAMPLE 17.61

A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kg. Second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kg. Do you agree with the claim that medicine B increases the weight significantly?

Solution: $n_1 = 5, n_2 = 7$

$$\bar{x} = \frac{\Sigma x}{n_1} = \frac{230}{5} = 46$$

$$\bar{y} = \frac{\Sigma y}{n_2} = \frac{399}{7} = 57$$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
42	−4	16	38	−19	361
39	−7	49	42	−15	225
48	2	4	56	−1	1
60	14	196	64	7	49
41	−5	25	68	11	21
			69	12	44
			62	5	25
$\Sigma x = 230$		$\Sigma(x-\bar{x})^2 = 290$	$\Sigma y = 399$		$\Sigma(y-\bar{y})^2 = 926$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{290+926}{5+7-2}} = 11.03$$

(i) Null Hypothesis $H_0 : \mu_1 = \mu_2$, i.e., there is no significant difference between the medicines A and B as regards their effect on the increase in weight

(ii) Alternative Hypothesis $H_1 : \mu_1 > \mu_2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{46 - 57}{11.03\sqrt{\dfrac{1}{5} + \dfrac{1}{7}}} = -1.7$

Calculated value of $|t| = 1.7$

(v) Critical value: Degrees of freedom $v = n_1 + n_2 - 2 = 5 + 7 - 2 = 10$
Tabulated value of t at 5% level of significance at 10 degrees of freedom is $t_{0.05} = 1.812$.

(vi) Decision: Since calculated value of $|t|$ is less than tabulated value of t, the null hypothesis is accepted at 5% level of significance, i.e., the medicines A and B do not differ significantly as regards their effect on increase in weight.

EXERCISE 17.16

1. A sample of 26 bulbs gives a mean life of 990 hours with a SD of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not up to standard?
[**Ans.:** The sample is not up to the standard]

2. The average breaking strength of the steel rods is specified to be 18.5 thousand pounds. To test this, sample of 14 rods were tested. The mean and SD obtained were 17.85 and 1.955 respectively. Is the result of experiment significant?
[**Ans.:** The result of experiment is not significant]

3. A random sample of six steel beams has a mean compressive strength of 58392 psi (pounds per square inch) with a SD of 648 psi. Use this information and level of significance $\alpha = 0.05$ to test whether the true average compressive strength of the steel from which this sample came is 58000 psi. Assume normality.
[**Ans.:** The average compressive strength of the steel beam is not equal to 58000 psi]

4. A sample of 155 members has a mean of 67 and SD of 52. Is this sample has been taken from a large population of mean 70?
[**Ans.:** The sample has not been taken from the given population]

5. The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level assuming that for 9 degrees of freedom $t = 1.833$ at $\alpha = 0.05$.
[**Ans.:** The average height is greater than 64 inches]

6. A random sample from a company's very extensive files shows that the orders for a certain kind of machinery were filled respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Use the level of significance $\alpha = 0.01$ to test the claim that on the average such orders are filled in 10.5 days. Choose the alternative hypothesis so that rejection of null hypothesis $\mu = 10.5$ days implies that it takes longer than indicated.
[**Ans.:** The orders on average are filled in more than 10.5 days]

7. Producer of gutkha claims that the nicotine content in his gutkha on the average is 1.83 mg. Can this claim be accepted if a random sample of 8 gutkha of this type have the nicotine contents of 2, 1.7, 2.1, 1.9, 2.2, 2.1, 2, 1.6 mg? Use a 0.05 level of significance.

[**Ans.:** The null hypothesis is accepted]

8. Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results:

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

[**Ans.:** The two horses do not have the same running capacity]

9. To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test which measures the IQ. The results are as follows:

Husbands	117	105	97	105	123	109	86	78	103	107
Wives	106	98	87	104	116	95	90	69	108	85

Test the hypothesis with a reasonable test at the level of significance of 0.05.

[**Ans.:** These is no significant difference in IQs]

10. Two independent samples of 8 and 7 items respectively had the following values:

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	–

Is the difference between the means of samples significant?

[**Ans.:** The difference between the mean of samples is not significant]

11. Random samples of specimens of coal from two mines A and B are drawn and their heat-producing capacity (in millions of calories/ton) were measured yielding the following results:

Mine A	8350	8070	8340	8130	8260	–
Mine B	7900	8140	7920	7840	7890	7950

Is there significant difference between the means of these two samples at 0.01 level of significance?

[**Ans.:** There is significant difference between the means of two samples]

12. The table gives the biological values of protein from 6 cow's milk and 6 buffalo's milk. Examine whether the differences are significant.

Cow's milk	1.8	2	1.9	1.6	1.8	1.5
Buffalow's milk	2	1.8	1.8	2	2.1	1.9

[**Ans.:** There is no significant difference in their means]

17.20 CHI-SQUARE (χ^2) TEST

The chi-square (χ^2) test is a useful measure of comparing experimentally obtained results with those expected theoretically and based on hypothesis. It is used as a test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared. The magnitude of discrepancy between observed and theoretical frequencies is given by the quantity χ^2 (pronounced as chi-square). If $\chi^2 = 0$, the observed and expected frequencies completely coincide. As the value of χ^2 increases, the discrepancy between the observed and theoretical frequency decreases.

If $O_1, O_2, \ldots, O_n$ be a set of observed frequencies and $E_1, E_2, \ldots, E_n$ be the corresponding set of expected (or theoretical) frequencies then χ^2 is defined by

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots + \frac{(O_n - E_n)^2}{E_n} = \sum \frac{(O - E)^2}{E}$$

with $n - 1$ degrees of freedom

Note If the data is given in a series of n numbers then degrees of freedom $v = n - 1$
In case of binomial distribution, $v = n - 1$
In case of Poisson distribution, $v = n - 2$
In case of normal distribution, $v = n - 3$

17.20.1 Chi-Square Distribution

If $x_1, x_2, \ldots, x_n$ are n independent normal variates with mean zero and standard deviation unity then $x_1^2 + x_2^2 + \ldots + x_n^2$ is a random variate having χ^2 distribution with probability density function given by

$$P(\chi^2) = y_0(\chi^2)^{\frac{v-1}{2}} e^{-\frac{\chi^2}{2}}$$

where v = degrees of freedom $= n - 1$ and y_0 = constant depending on the degrees of freedom

17.20.2 Properties of χ^2-Distribution

(i) Chi-Square test is always positively skewed.
(ii) The mean of chi-square distribution is the number of degrees of freedom.
(iii) The standard deviation of chi-square distribution $= \sqrt{2v}$.
(iv) Chi-square values increases with the increase in degrees of freedom.
(v) The value of χ^2 lies between zero and infinity.
(vi) For different values of degrees of freedom, the shape of the curve will be different.

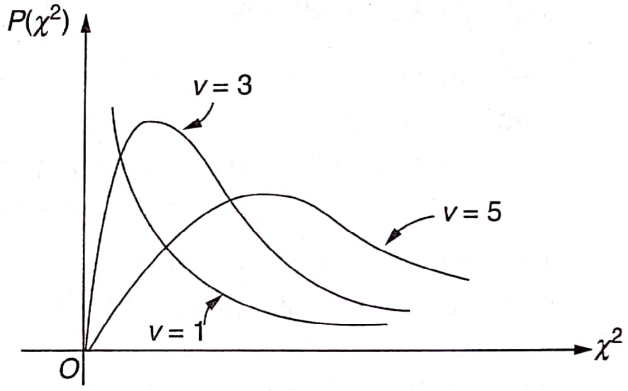


Fig 17.31 Chi-square distribution curve

17.20.3 Chi-square Test as a Test of Goodness of Fit

The values of χ^2 is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not. It is also used to fit a set of observations to a given distribution. Hence, chi-square test provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution.

17.20.4 Test of Significance

Let $O_1, O_2, \ldots, O_n$ be a set of observed frequencies and $E_1, E_2, \ldots, E_n$ be the corresponding set of expected or theoretical frequencies. The χ^2 statistic is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Working Rule

(i) Set up a null hypothesis and calculate χ^2.
(ii) Set up an alternative hypothesis.
(iii) Set a level of significance α.
(iv) Find the degree of freedom and find the corresponding value of χ^2 at given level of significance α.
(v) If the calculated value of χ^2 is less than tabulated value of χ^2 at the level of significance α, the null hypothesis is accepted. If calculated value of χ^2 is more than tabulated value of χ^2 at the level of significance α, the null hypothesis is rejected.

EXAMPLE 17.62

The following mistakes per page were observed in a book:

No. of mistakes per page	0	1	2	3	4
No. of pages	211	90	19	5	0

Fit a Poisson distribution and test the goodness of fit.

Solution:

(i) Null Hypothesis H_0: The mistakes follow Poisson distribution and Poisson distribution can be fitted to the data.

(ii) Alternative Hypothesis H_1: The mistakes do not follow Poisson distribution

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: The expected frequencies by Poisson distribution are given by

$$\text{Expected frequency } f = Np = N\left(\frac{e^{-\lambda}\lambda^x}{x!}\right), x = 0,1,2,3,4$$

$$\lambda = \frac{\Sigma fx}{N} = \frac{211(0)+90(1)+19(2)+5(3)+0(4)}{211+90+19+5+0} = 0.44$$

$$f = Np = 325\left(\frac{e^{-0.44}0.44^x}{x!}\right), \quad x = 0,1,2,3,4$$

Expected or Theoretical frequency

x	0	1	2	3	4
f	209.31	92.1	20.26	2.97	0.33

When expected frequencies are less than 10, classes are grouped together.

No. of mistakes	*Observed frequency* O	*Expected frequency* E	$O-E$	$\frac{(O-E)^2}{E}$
0	211	209.31	1.69	0.014
1	90	92.10	–2.1	0.048
2	19	20.26	0.44	0.008
3	5	2.97		
4	0	0.33		

$$\chi^2 = \sum\frac{(O-E)^2}{E} = 0.07$$

(v) Critical value: The number of degrees of freedom is 1 for each class. There are 5 classes originally. Hence, the degrees of freedom originally is 5. Since the classes are reduced by 2, the degrees of freedom is reduced by 2. Further, while calculating the parameter λ, two sums Σfx and Σf are used. Hence, the degrees of freedom is again reduced by 2.
Hence, the number of degrees of freedom $\gamma = 5 - (2 + 2) = 1$
Tabulated value of χ^2 at $\alpha = 0.05$ for 1 degree of freedom is $\chi^2_{0.05} = 3.84$

(vi) Decision: Since the calculated value of χ^2 is less than the tabulated value of $\chi^2_{0.05}$, the null hypothesis is accepted at 5% level of significance, i.e., the mistakes follow Poisson's distribution.

EXAMPLE 17.63

A set of five similar coins is tossed 320 times and result is obtained as follows:

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution.

Solution:

(i) Null Hypothesis H_0: The data follow a binomial distribution.

(ii) Alternative Hypothesis H_1: The data do not follow binomial distribution.

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: Probability of getting a head $p = \frac{1}{2}$

Probability of getting a tail $q = \frac{1}{2}$

By binomial distribution,

$$p(x) = {}^nC_x p^x q^{n-x} = {}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}, \quad x = 0,1,2,3,4,5$$

$$N = 320$$

Expected frequency $f = Np(x) = 320\left[{}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}\right], x = 0,1,2,3,4,5$

Expected or theoretical frequency

x	0	1	2	3	4	5
f	10	50	100	100	50	10

No. of heads	Observed frequency O	Expected frequency E	$O-E$	$\frac{(O-E)^2}{E}$
0	6	10	−4	1.6
1	27	50	−23	10.58
2	72	100	−28	7.84
3	112	100	12	1.44
4	71	50	21	8.82
5	32	10	22	48.4

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 78.68$$

(v) Critical value: Degrees of freedom $\nu = n - 1 = 6 - 1 = 5$
Tabulated value of χ^2 at $\alpha = 0.05$ for 5 degrees of freedom is $\chi^2_{0.05} = 11.07$

(vi) Decision: Since the calculated value of χ^2 is more than tabulated value of $\chi^2_{0.05}$ at 5% level of significance, the null hypothesis is rejected, i.e., the data do not follow the binomial distribution.

EXAMPLE 17.64

Fit the equation of the best fitting normal curve to the following data:

x	135	145	155	165	175	185	195	205	Total
f	2	14	22	25	19	13	3	2	100

Compare the theoretical and observed frequencies. Using χ^2 test find goodness of fit. Given that $\mu = 165.6$ and $\sigma = 15.02$.

Solution: $\mu = 165.6$, $\sigma = 15.02$, $N = \Sigma f = 100$

The data is first converted into class intervals with inclusive series

Class interval	Lower class X	$Z = \frac{X-\mu}{\sigma}$	Area from O to Z	Area in class interval	Expected frequencies
130–140	130	–2.37	0.4911	0.0357	$3.57 \simeq 4$
140–150	140	–1.70	0.4554	0.1046	$10.46 \simeq 11$
150–160	150	–1.04	0.3508	0.2065	$20.65 \simeq 21$
160–170	160	–0.37	0.1443	0.2584	$25.84 \simeq 26$
170–180	170	0.29	0.1141	0.2174	$21.74 \simeq 21$
180–190	180	0.96	0.3315	0.1159	$11.59 \simeq 12$
190–200	190	1.62	0.4474	0.0416	$4.16 \simeq 4$
200–210	200	2.29	0.4890	0.0095	$0.95 \simeq 1$
210–220	210	2.96	0.4985		

Calculation of χ^2

When expected frequencies are less than 10, classes are grouped together.

x	Observed frequency O	Expected frequency E	$O-E$	$\frac{(O-E)^2}{E}$
135	2	4	1	0.067
145	14	11		
155	22	21	1	0.048
165	25	26	–1	0.038
175	19	21	–2	0.19
185	13	12		
195	3	4	1	0.0588
205	2	1		

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.4018$$

(v) Critical value: There are 5 frequencies. While calculating mean and standard deviation, three sums Σf, Σfx, and Σfx^2 are used. Hence, the number of degrees of freedom $\nu = 5 - 3 = 2$

Tabulated value of χ^2 at $\alpha = 0.05$ for 2 degrees of freedom is $\chi^2_{0.05} = 5.99$

Since the calculated value of χ^2 is less than the tabulated value of $\chi^2_{0.05}$ at 5% level of significance, the fit is good and the distribution is nearly normal.

EXERCISE 17.17

1. A dice is thrown 264 times with the following results: Show that the dice is biased [Given $\chi^2_{0.05} = 11.07$ for 5 df]

No. appeared on the dice	1	2	3	4	5	6
Frequency	40	32	28	58	54	52

2. A pair of dice are thrown 360 times and frequency of each sum is given below:

Sum	2	3	4	5	6	7	8	9	10	11	12
Frequency	8	24	35	37	44	65	51	42	26	14	14

would you say that the dice are fair on the basis of the chi-square test at 0.05 level of significance? [**Ans.:** The dice are fair]

3. 4 coins are tossed 160 times and the following results were obtained:

No. of heads	0	1	2	3	4
Observed frequencies	17	52	54	31	6

Under the assumption that coins are balanced, find the expected frequencies of 0, 1, 2, 3 or 4 heads, and test the goodness of fit ($\alpha = 0.05$).

[**Ans.:** Expected frequencies: 10, 40, 60, 40, 10, the data do not follow binomial distribution]

4. Fit a poisson distribution to the following data and for its goodness of fit at level of significance 0.05:

x	0	1	2	3	4
f	419	352	154	56	19

5. The following table gives the number of accidents in a city during a week. Find whether the accidents are uniformly distributed over a week.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents	13	15	9	11	12	10	14	84

[**Ans.:** The accidents are uniformly distributed over a week]

6. Weights in kilograms of 10 students are given below: 38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of the normal distribution from which the above sample is drawn is 20 kg?

[**Ans.:** The sample is drawn from the normal population with variance 20]

7. Five dice are thrown 192 times and the number of times 4, 5 or 6 are obtained are as follows:

No. of dice showing 4, 5, 6	5	4	3	2	1	0
Frequency	6	46	70	48	20	2

Calculate χ^2. [**Ans.:** 16.94]

8. The distribution of defects in printed circuit board is hypothesised to follow Poisson distribution. A random sample of 60 printed boards shows the following data:

No. of defects	0	1	2	3
Observed frequency	32	15	9	4

Does the hypothesis of Poisson distribution appropriate?

[**Ans.:** The defects follow Poisson distribution]

17.21 SNEDECOR'S *F*-DISTRIBUTION

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ be the values of two independent random samples of sizes n_1 and n_2 ($n_1 \leq 30$, $n_2 \leq 30$) with means $\bar{x}$ and $\bar{y}$ drawn from the normal population with mean μ and standard deviation σ. The Snedecer's F-distribution is given by

$$F = \frac{s_1^2}{s_2^2}$$

with $(n_1 - 1)$ degrees of freedom for the numerator and $(n_2 - 1)$ degrees of freedom for the denominator where s_1 and s_2 are standard deviations of samples.

$$s_1^2 = \frac{\Sigma(x - \bar{x})^2}{n_1 - 1}$$

$$s_2^2 = \frac{\Sigma(y - \bar{y})^2}{n_2 - 1}$$

$$\bar{x} = \frac{\Sigma x}{n_1}$$

$$\bar{y} = \frac{\Sigma y}{n_2}$$

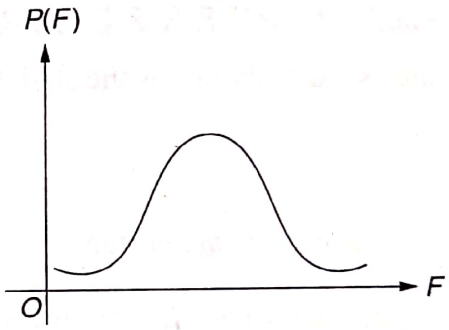


Fig. 17.32 *F*-distribution curve

The Snedecor's F-distribution is defined by

$$P(F) = cF^{\left(\frac{\nu_1 - 2}{2}\right)}\left(1 + \frac{\nu_1}{\nu_2}F\right)^{-\left(\frac{\nu_1 + \nu_2}{2}\right)}$$

where the constant c depends on ν_1 and ν_2. It is so chosen that the area under the curve is unity.

17.21.1 Properties of *F*-distribution

(i) F-distribution curve lies entirely in the first quadrant and is unimodal.

(ii) F-distribution is independent of the population variance σ^2 and depends on ν_1 and ν_2 only.

(iii) The mode of F-distribution is less than unity.

(iv) $F_{1-\alpha}(\nu_1, \nu_2) = \dfrac{1}{F_\alpha(\nu_2, \nu_1)}$

where $F_\alpha(\nu_2, \nu_1)$ is the value of F with ν_2 and ν_1 degrees of freedom such that the area under the F-distribution curve right of F_α is α.

17.21.2 Test of Significance

Significant test is performed by means of Snedecer's F-table which provides 5% and 1% of points of significance for F. 5% points of F means that the area under the F-curve, to the right of the ordinate at a value of F, is 0.05. Further, F-table gives only single tail test. F-distribution is very useful for testing the equality of population means by comparing sample variances.

HISTORICAL DATA

George Waddel Snedecor (1881–1974) was an American mathematician and statistician. He was born in Memphis, Tennessee, into a socially prominent and politically powerful, southern Democratic, Presbyterian family line. Snedecor contributed to the foundations of analysis of variance, data analysis, experimental design, and statistical methodology. Snedecor's F distribution and the George W. Snedecor Award of the American Statistical Association are named after him. He was awarded honorary doctorates in science by North Carolina State University in 1956 and by Iowa State University in 1958. He died on 15 February 1974 in Amherst, Massachusetts, United States.

17.22 FISHER'S *Z*-DISTRIBUTION

Fisher's z-distribution is the statistical distribution of half the logarithm of F-distribution variate.

$$z = \frac{1}{2}\log F$$

Significance tests are performed from the z-table in a way similar to that of F-distribution.

Properties of Fisher's z-distribution

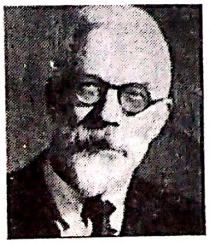
(i) It is more symmetrical than F-distribution.

(ii) z-distribution is a family of distributions for different values of ν_1 and ν_2.

(iii) The mean of z-distribution is $\frac{1}{2}\left(\frac{1}{\nu_2} - \frac{1}{\nu_1}\right)$.

(iv) The variance of z-distribution is $\frac{1}{2}\left(\frac{1}{\nu_1} + \frac{1}{\nu_2} + \frac{1}{\nu_1^2} + \frac{1}{\nu_2^2}\right)$.

HISTORICAL DATA



Sir Ronald Aylmer Fisher (1890 –1962) was born in East Finchley in London, England. Fisher who published as R.A. Fisher was an English statistician and biologist. He gained a scholarship to study Mathematics at the University of Cambridge in 1909, gained a First in Astronomy in 1912. He developed Fisher's z-distribution a new statistical method, commonly used decades later as the F distribution. He pioneered the principles of the design of experiments and the statistics of small samples and the analysis of real data. In 1957, a retired Fisher emigrated to Australia where he spent time as a senior research fellow at the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Adelaide, where he died in 1962, with his remains interred within St Peter's Cathedral.

EXAMPLE 17.65

In two independent samples of sizes 8 and 10, the sum of squares of deviations of the sample values from the respective means were 84.4 and 102.6. Test whether the difference of variances of the population is significant or not. Use a 0.05 level of significance.

Solution:

$$n_1 = 8, \quad n_2 = 10$$
$$\Sigma(x-\bar{x})^2 = 84.4, \quad \Sigma(y-\bar{y})^2 = 102.6,$$

$$s_1^2 = \frac{\Sigma(x-\bar{x})^2}{n_1 - 1} = \frac{84.4}{8-1} = 12.057$$

$$s_2^2 = \frac{\Sigma(y-\bar{y})^2}{n_2 - 1} = \frac{102.6}{10-1} = 11.4$$

(i) Null Hypothesis $H_0 : s_1^2 = s_2^2$

(ii) Alternative Hypothesis $H_1 : s_1^2 \neq s_2^2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $F = \frac{s_1^2}{s_2^2} = \frac{12.057}{11.4} = 1.057$

Calculated $F = 1.057$

(v) Critical value: Degrees of freedom

$$\nu_1 = n_1 - 1 = 8 - 1 = 7$$
$$\nu_2 = n_2 - 1 = 10 - 1 = 9$$

Tabulated value of F at 0.05 level of significance for 7 and 9 degrees of freedom is $F_{0.05} = 3.29$

(vi) Decision: Since calculated F is less than tabulated F, the null hypothesis is accepted at 0.05 level of significance, i.e., there is no significant difference in variances of the population.

EXAMPLE 17.66

In a test given to two groups of students drawn from two normal populations, the marks obtained were as follows:

Group A	18	20	36	50	49	36	34	49	41
Group B	29	28	26	35	30	44	46		

Examine at 5% level, whether the two populations have the same variance.

Solution: $n_A = 9, n_B = 7$

$$\bar{x} = \frac{\Sigma x}{n_A} = \frac{333}{9} = 37$$

$$\bar{y} = \frac{\Sigma y}{n_B} = \frac{238}{7} = 34$$

Group A			Group B		
x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
18	−19	361	29	−5	25
20	−17	289	28	−6	36
36	−1	1	26	−8	64
50	13	169	35	1	1
49	12	144	30	−4	16
36	−1	1	44	10	100
34	−3	9	46	12	144
49	12	144			
41	4	16			
$\Sigma x = 333$		$\Sigma(x-\bar{x})^2 = 1134$	$\Sigma y = 238$		$\Sigma(y-\bar{y})^2 = 386$

$$s_A^2 = \frac{\Sigma(x-\bar{x})^2}{n_A - 1} = \frac{1134}{9-1} = 141.75$$

$$s_B^2 = \frac{\Sigma(y-\bar{y})^2}{n_B - 1} = \frac{386}{7-1} = 64.33$$

(i) Null Hypothesis $H_0 : s_A^2 = s_B^2$

(ii) Alternative Hypothesis $H_1 : s_A^2 \neq s_B^2$

(iii) Level of significance: $\alpha = 0.05$

(iv) Test statistic: $F = \dfrac{s_A^2}{s_B^2} \quad \left(\because s_A^2 > s_B^2\right)$

$$= \frac{141.75}{64.33} = 2.203$$

Calculated $F = 2.203$

(v) Critical value: Degrees of freedom

$$\nu_A = n_A - 1 = 9 - 1 = 8$$

$$\nu_B = n_B - 1 = 7 - 1 = 6$$

Tabulated value of F at 5% level of significance for 8 and 6 degrees of freedom is $F_{0.05} = 4.15$

(vi) Decision: Since calculated F is less than tabulated F, the null hypothesis is accepted at 5% level of significance, i.e., the two populations have the same variance.

EXERCISE 17.18

1. If two independent samples of sizes $n_1 = 13$ and $n_2 = 7$ are taken from a normal population. What is the probability that the variance of the first sample will be at least four times as large as that of the second sample?

[**Ans.:** 0.05]

2. The standard deviations calculated from two random samples of size 9 and 13 are 2 and 1.9 respectively. Can the samples be regarded as drawn from the normal populations with the same standard deviation?

[**Ans.:** The samples can be regarded as drawn from the normal populations with the same standard deviation]

3. Two samples are drawn from two normal populations. From the following data test whether the two samples have the same variance at 5% level?

Sample I	60	65	71	74	76	82	85	87		
Sample II	61	66	67	85	78	63	85	86	88	91

[**Ans.:** Two samples have the same variance]

4. The time taken by workers in performing a job by method I and method II is given below.

Method I	20	16	26	27	22		
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution in a population from which these samples are drawn do not differ significantly?

[**Ans.:** The variances of time distribution in a population from which the samples are drawn do not differ significantly]

5. Following results were obtained from two samples, each drawn from two different population A and B:

Population	A	B
Sample	I	II
Sample size	25	17
Sample SD	3	2

Test the hypothesis that the variance of brand A is more than that of B.

[**Ans.:** Variance of brand A is not more than the variance of brand B]

6. In a laboratory experiment two samples gave the following results:

Sample	*Size*	*Sample mean*	*Sum of squares of deviation from the mean*
1	10	15	90
2	12	14	108

Test the equality of sample variances at 5% level of significance.

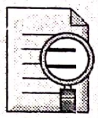
[**Ans.:** The two population have the same variance]

For interactive quiz

Scan the given QR code here

or Visit **http://qrcode.flipick.com/index.php/513**

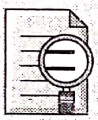
For important formulae



Scan the given QR code here

or Visit **http://qrcode.flipick.com/index.php/515**

For additional solved examples



Scan the given QR code here

or Visit **http://qrcode.flipick.com/index.php/514**