

# Biostatistics

# **Chapter # 1**

# **Introduction**

## INTRODUCTION TO BIOSTATISTICS

### Statistics: ★

It refers to the subject of scientific activity which deals with the theories and methods of collection, analysis and interpretation of such data.

### Biostatistics: ★

This term is used when tools of statistics are applied to the data that is derived from biological organisms.

### Characteristics of Statistics: ✓ ★

1. Statistics are the aggregate of facts.
2. Statistics are numerically expressed.
3. Statistics are affected by multiplicity of causes and not by single cause.
4. Statistics must be related to some field of inquiry.
5. Statistics should be capable of being related to each other, so that some cause & effect relationship can be established.
6. The reasonable standard of accuracy should be maintained in statistics.

### Importance and Usefulness of Statistics: ★

1. Statistics help in presenting large quantity of data in a simple and classified form.
2. It gives the methods of comparison of data.
3. It enlarges individual mind.
4. It helps in finding the conditions of relationship between the variables.
5. It tries to give material for the business man as well as the administrators so as to serve as a guide in planning and shaping future policies and programmes.
6. It proves useful in number of fields viz. railways, Banks, Army, etc.

### Limitation of Statistics: ★

1. Statistics laws are held to be true on the average and in the long run.
2. Statistics can be used to analyse only collective matters not individual events.
3. It is applicable only to quantitative data.
4. Statistical results are ascertained by samples. If the selection of samples is biased, errors will accumulate and results will not be reliable.

5. The greatest limitation of statistics is that only one who has an expert knowledge of statistical methods can efficiently handle statistical data.

### Application and Uses of Biostatistics: ★

#### 1. In Physiology and Anatomy

- (i) To define what is normal or healthy in a population and to find limits of normality in variables.
- (ii) To find the difference between the means and proportions of normal at two places or in different periods.
- (iii) To find out correlation between two variables  $X$  and  $Y$  such as height and weight.

#### 2. In Pharmacology

- (i) To find out the action of drug—a drug is given to animals & humans to observe the changes produced are due to the drug or by chance.
- (ii) To compare the action of different drugs or two successive dosages of the same drug.
- (iii) To find out the relative potency of a new drug with respect to a standard drug.

#### 3. In Medicine

- (i) To compare the efficacy of a particular drug. For this, the percentage of cured & died in the experiment & control groups.
- (ii) To find out an association between two attributes such as cancer and smoking.
- (iii) To identify signs and symptoms of a disease or syndrome. Cough & typhoid is found by chance and fever is found in almost every case.

#### 4. In Community Medicine and Public Health

- (i) To test usefulness of sera and Vaccines in the field—the percentage of attacks or deaths among the vaccinated subject is compared with that among the unvaccinated ones to find whether the difference observed is statistically significant.
- (ii) In epidemiological studies—the role of causative factors is statistically tested.
- (iii) In public health, the measures adopted are evaluated.

### DATA: ★

Data is a collection of observations expressed in numerical figures. The collection may be done in two ways.

- (a) by complete enumeration and
- (b) simple survey method.

Data is always in collective sense and never be used singular.

### Types of Data:

The statistical data can be divided into two broad categories:

- (a) Qualitative
- (b) Quantitative.

#### Qualitative Data :

In this type of data, there is no numerical relation with one another.

Example: Skin colour—brown, black, white  
 Eye colour—blue, brown  
 Sex—Male, Female.

#### Quantitative Data:

1. In this type of data, there is numerical relation with one another.
2. It may be continuous or discrete.



**Q1.1 Write short answers to the following questions.**

**Q.1.** Define statistics as a discipline of science. (B.I.S.E., Gujranwala 20)

**Ans.** Statistics refers to the science comprising methods which are used in the collection, presentation, analysis and interpretation of numerical data.

**Q.2.** Define population and sample.

**Ans.** A set of individuals or objects having some common measurable characteristics is called a *population*. A subset or part of the population selected for study is called a *sample*.

**Q.3.** Given an example of the population and the sample.

**Ans.** Suppose we want to find out the average age of under-graduate students of a local college where there are 2000 undergraduate students. We select 100 students and record their ages. Here the population consists of 2000 students and 100 students selected from this population is a sample.

**Q.4.** Differentiate between descriptive and inferential statistics.

**Ans.** The phase of statistics that is concerned with the description and analysis of sample or population data is called *descriptive statistics*.

The phase of statistics that is concerned with the procedures and methodology for obtaining valid conclusions is called *inferential statistics*.

**Q.5.** What is the aim of collecting numerical data for a statistical study?

**Ans.** Numerical data are collected from the universe under study. Inferential statistics is concerned with drawing conclusions about the population or universe obtained from the sample.

**Q.6.** Define complete enumeration (or census) and sampling.

**Ans.** If the information is collected from every individual of the population, the inquiry is known as *complete enumeration* or *census*.

The process of selecting a sample of selected individuals of the population is known as *sampling*.

**Q.7.** Distinguish between primary and secondary data.

(B.I.S.E., Lahore 2009)

**Ans.** The data published or used by an organization which originally collected them are called *primary data*.

The data published or used by an organization other than the one who originally collected them are known as *secondary data*.

**01.1 Write short answers to the following questions.**

**Q.1.** Define statistics as a discipline of science. (B.I.S.E., Gujranwala 2009)

**Ans.** Statistics refers to the science comprising methods which are used in the collection, presentation, analysis and interpretation of numerical data.

**Q.2.** Define population and sample.

**Ans.** A set of individuals or objects having some common measurable characteristics is called a *population*. A subset or part of the population selected for study is called a *sample*.

**Q.3.** Given an example of the population and the sample.

**Ans.** Suppose we want to find out the average age of under-graduate students of a local college where there are 2000 undergraduate students. We select 100 students and record their ages. Here the population consists of 2000 students and 100 students selected from this population is a sample.

**Q.4.** Differentiate between descriptive and inferential statistics.

**Ans.** The phase of statistics that is concerned with the description and analysis of sample or population data is called *descriptive statistics*. The phase of statistics that is concerned with the procedures and methodology for obtaining valid conclusions is called *inferential statistics*.

**Q.5.** What is the aim of collecting numerical data for a statistical study?

**Ans.** Numerical data are collected from the universe under study. Inferential statistics is concerned with drawing conclusions about the population or universe obtained from the sample.

**Q.6.** Define complete enumeration (or census) and sampling.

**Ans.** If the information is collected from every individual of the population, the inquiry is known as *complete enumeration* or *census*.

The process of selecting a sample of selected individuals of the population is known as *sampling*.

**Q.7.** Distinguish between primary and secondary data.

(B.I.S.E., Lahore 2009)

**Ans.** The data published or used by an organization which originally collected them are called *primary data*.

The data published or used by an organization other than the one who originally collected them are known as *secondary data*.



Q.8. Give an example each of primary and secondary data.

Ans. The data in the *Population Census Reports* are primary because these are collected and published by the *Population Census Organization*.

The data in the *Economic Survey of Pakistan* are secondary because these are originally collected by other agencies like *Federal Bureau of Statistics*.

Q.9. Define variable and constant.

(B.I.S.E., Rawalpindi 2007; Multan 2009)

Ans. A measurable quantity which can vary from one individual or object to another is called a *variable*.

A quantity which can assume only one value is called a *constant*.

Q.10. Distinguish between variable and attribute.

Ans. A characteristic which can be measured numerically is called a *variable*, e.g. income, weight, age, etc.

A characteristic which cannot be measured numerically but only its presence or absence can be described is called an *attribute*, e.g. marital status, sex, religion, etc.

Q.11. Differentiate between discrete and continuous variables.

Ans. A variable which can assume only some specific or limited number of values within a given range is called a *discrete variable*. A variable which can assume an infinite number of values within a given range is called a *continuous variable*.

Q.12. Give an example each of the discrete and continuous variables.

Ans. The number of children in a family is a discrete variable. The age in years of a child is a continuous variable.

Q.13. Differentiate between discrete and continuous data.

Ans. Data which can be described by a discrete variable are called *discrete data*, e.g. the number of children in 100 families.

Data which can be described by a continuous variable are called *continuous data*, e.g. heights of 100 students of a college.

Q.14. Differentiate between quantitative and qualitative data.

(B.I.S.E., Rawalpindi 2007)

Ans. Data which can be described by a quantitative variable such as height, age, weight, etc. are called *quantitative data*.

Data which can be described by a qualitative variable such as marital status, sex, religion, etc. are called *qualitative data*.

Q.15. What is the difference between primary and secondary data?

**Q.16.** Enumerate four uses of statistics.

**Ans.** (i) Statistics presents facts in a definite form. (ii) Statistics studies relationships among different facts. (iii) Statistics aids forecasting. (iv) Statistics guides the formulation of policies.

**Q.17.** Name the sources of primary data. (B.I.S.E., Gujranwala 2008)

**Ans.** (i) Direct personal observation (ii) Registration (iii) Estimates through local correspondents (iv) Investigation through enumerators (v) Information through mailed questionnaire.

**Q.18.** Name the sources of secondary data.

**Ans.** (i) Official sources, e.g. various government publications, specially publications of Federal Bureau of Statistics. (ii) Semi-official sources, e.g. publications of State Bank of Pakistan. (iii) Private sources, e.g. publications of Chamber of Commerce and Industry. (iv) Publications of Research Organizations.



Q2.1 Write short answers to the following questions.

Q.1. What is classification?

(B.I.S.E., Lahore 2009; Multan 2009)

Ans. The process of arranging data into classes or categories according to some common characteristics present in the data. For example, in classifying the population of a country by religion, we may arrange Muslims, Christians, Hindus, etc. into different groups.

Q.2. Describe the four bases of classification of data.

Ans. (i) *Qualitative* when data are classified by attributes, e.g. sex, religion etc. (ii) *Quantitative* when data are classified by quantitative characteristics, e.g. height, weight, etc. (iii) *Geographical* when data are classified by geographical regions or locations, e.g. the population of a country may be classified by provinces or districts. (iv) *Chronological or temporal* when data are arranged by their time of occurrence. Such arrangement is called a *time series*.

Q.3. What is a table?

Ans. A table is a systematic arrangement of data into vertical columns and horizontal rows.

Q.4. Define tabulation.

(B.I.S.E., Lahore 2007, 2009)

Ans. The process of arranging data into rows and columns is called tabulation.

Q.5. What is a frequency distribution?

Ans. A frequency distribution is a tabular arrangement of data in which various items in each class (called class frequencies) are stated.

Q.6. Define grouped data.

Ans. Data presented in the form of a frequency distribution are called grouped data.

Q.7. Differentiate between grouped and ungrouped data.

Ans. Data collected from the field and which have not been arranged in a systematic order are called raw data or ungrouped data. Data arranged in a systematic order as in a frequency distribution are called grouped data.

Q.8. What is an array?

Ans. An arrangement of raw numerical data in ascending or descending order is called an array.

## **Chapter # 2**

# **Measures of Location**



**3.1 Introduction** In the previous chapter, we have seen that it is difficult to learn anything by looking at the data which have not been properly arranged. When the data have been arranged into a frequency distribution, the information contained in the data is easily understood. We have also seen that important features of the data become clear at a glance when the frequency distribution is represented by means of a graph. We can still go further and find a *single value* which will represent all the values of the distribution in some definite way. A value which is used in this way to represent the distribution is called an *average*. Since the averages tend to lie in the centre of a distribution they are called *measures of central tendency*. They are also called *measures of location* because they locate the centre of a distribution.

**3.2 Types of Averages** The most commonly used averages are (i) the arithmetic mean, (ii) the geometric mean, (iii) the harmonic mean, (iv) the median, and (v) the mode.

**3.3 The Arithmetic Mean** The arithmetic mean is the most commonly used average. In view of its common use, it is usually referred to as *the average* or simply *the mean*.

The arithmetic mean or simply the mean is defined as the value obtained by dividing the sum of the values by their number. Thus the mean of the values  $X_1, X_2, \dots, X_n$  denoted by  $\bar{X}$  (read as X-bar) is

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} = \frac{\sum X}{n} \quad (3.1)$$

**Example 3.1 (a)** The arithmetic mean of the values 5, 8, 10, 12 and 7 is

$$\bar{X} = \frac{5 + 8 + 10 + 12 + 7}{5} = \frac{42}{5} = 8.4.$$

**Example 3.1 (b)** Total annual incomes of eight families are Rs.3200, Rs.4000, Rs.3500, Rs.4500, Rs.3800, Rs.4200, Rs.3600 and Rs.53200. Their arithmetic mean is obtained as

$$\begin{aligned} \bar{X} &= \frac{3200 + 4000 + 3500 + 4500 + 3800 + 4200 + 3600 + 53200}{8} \\ &= \frac{80000}{8} = \text{Rs.10000.} \end{aligned}$$



$$\bar{X} = \frac{\sum_{i=1}^k f_i X_i}{\sum_{i=1}^k f_i} = \frac{\sum fX}{\sum f}$$

(3.2)

**Example 3.2 (a)** Find the mean weight of 120 students at the Punjab University from the frequency distribution in Table 2.2.

**Solution** The computation of mean is outlined in Table 3.1.

**Table 3.1**

| Weight<br>(pounds)     | Class Mark<br>(X) | Frequency<br>(f) | fX                |
|------------------------|-------------------|------------------|-------------------|
| 110 - 119              | 114.5             | 1                | 114.5             |
| 120 - 129              | 124.5             | 4                | 498.0             |
| 130 - 139              | 134.5             | 17               | 2286.5            |
| 140 - 149              | 144.5             | 28               | 4046.0            |
| 150 - 159              | 154.5             | 25               | 3862.5            |
| 160 - 169              | 164.5             | 18               | 2961.0            |
| 170 - 179              | 174.5             | 13               | 2268.5            |
| 180 - 189              | 184.5             | 6                | 1107.0            |
| 190 - 199              | 194.5             | 5                | 972.5             |
| 200 - 209              | 204.5             | 2                | 409.0             |
| 210 - 219              | 214.5             | 1                | 214.5             |
| $n = \sum f = n = 120$ |                   |                  | $\sum fX = 18740$ |

Here  $n = \Sigma f = 120$  and  $\Sigma fX = 18740$ . Using Formula (3.2) the mean weight is given by

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{18740}{120} = 156.17 \text{ pounds.}$$

**Example 3.2 (b)** Find the arithmetic mean for the following distribution showing marks obtained by 50 students in English at a certain examination.

| Marks     | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 |
|-----------|-------|-------|-------|-------|-------|-------|-------|
| Frequency | 1     | 4     | 8     | 11    | 15    | 9     | 2     |

**Solution** Computation of the arithmetic mean is outlined in Table 3.2

**Table 3.2**

| Marks   | Frequency ( $f$ )   | Class Mark ( $X$ ) | $fX$               |
|---------|---------------------|--------------------|--------------------|
| 20 - 24 | 1                   | 22                 | 22                 |
| 25 - 29 | 4                   | 27                 | 108                |
| 30 - 34 | 8                   | 32                 | 256                |
| 35 - 39 | 11                  | 37                 | 407                |
| 40 - 44 | 15                  | 42                 | 630                |
| 45 - 49 | 9                   | 47                 | 423                |
| 50 - 54 | 2                   | 52                 | 104                |
|         | $n = \Sigma f = 50$ |                    | $\Sigma fX = 1950$ |

Here  $n = 50$  and  $\Sigma fX = 1950$ . Using Formula (3.2), we get

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{1950}{50} = 39 \text{ marks.}$$

$\bar{X} = 29.84$ ,  $G = 20.57$  and  $A = 25.00$

**3.6 The Median** The median of a set of values arranged in ascending or descending order of magnitude is defined as the middle value if the number of values is odd and the mean of the two middle values if the number of values is even. The median divides a distribution into two halves and the number of values greater than the median is equal to the number of values smaller than the median.

**Example 3.20 (a)** The median of the values 4, 5, 6, 8, 10, 11 and 12 is 8.

(b) The median of the values, 4, 6, 7, 9, 11 and 13 is  $\frac{7+9}{2} = 8$ .

When the number of values is *odd*, the median is the middle value and when the number of values is *even*, the median is the mean of the two middle values. In both cases, the median is the value of  $\left(\frac{n+1}{2}\right)$ th item from either end in the array. In Example 3.20(a),  $n = 7$  and the  $\left(\frac{7+1}{2}\right)$ th



**3.6.1 The Median for Grouped Data** We have seen that the median is the value of  $\left(\frac{n+1}{2}\right)$ th item. In case of a frequency distribution, the median is the value of  $\left(\frac{n}{2}\right)$ th item from either end. Thus if we have 100 items in a frequency distribution, the median will be the value of the 50th item. To find the median from a frequency distribution, we form a cumulative frequency distribution. The median lies in the class which corresponds to the cumulative frequency in which  $\left(\frac{n}{2}\right)$  lies. It is given by the formula (obtained by interpolation)

$$\text{Median} = l + \frac{h}{f} \left( \frac{n}{2} - F \right) \quad (3.17)$$

$l$  = lower class boundary of the median class, i.e. the class corresponding to the cumulative frequency in which  $(n/2)$  lies

$h$  = class interval size of the median class

$f$  = frequency of the median class

$n$  = number of values or the total frequency

$F$  = cumulative frequency of the class preceding the median class

**3.7 The Mode** The *mode* is defined as that value in the data which occurs the-greatest number of times provided such a value exists.

**Example 3.27** (a) (i) The mode of the values 2, 5, 7, 8, 9, 9, 9 and 10 is 9.  
(ii) The mode of the values 11, 12, 12, 14, 15, 15, 15, 17, 17 and 19 is 15.

If each value occurs the same number of times, then there is no mode. If two or more values occur the same number of times but more frequently than any of the other values, then there is more than one mode. In this respect the mode differs from the mean and the median because there is only one mean and only one median.

**Example 3.27** (b) (i) The set of values 2, 3, 4, 4, 4, 6, 8, 9, 9 and 9 has two modes 4 and 9.

(ii) The set of values 1, 2, 5, 6, 12, 13 and 14 has no mode.

A distribution having only one mode is called *uni-modal distribution*, a distribution having two modes is called *bi-modal distribution* and a distribution having more than two modes is called a *multi-modal distribution*.

**3.7.1 The Mode from Grouped Data** In case of a unimodal frequency distribution where a frequency curve has been constructed for the data, the mode is defined as that value of  $X$  which corresponds to the highest point on the curve. This value of  $X$  is sometimes denoted by  $\hat{X}$  (read as  $X$ -caret).

In a frequency distribution with equal class interval sizes, the class with the highest frequency is called the modal class. The mode is given by the formula.

$$\text{Mode} = l + \frac{(f_m - f_{m-1})}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times h \quad (3.24)$$

where  $l$  = lower class boundary of the modal class, i.e. the class with the highest frequency

$f_m$  = frequency of the modal class

$f_{m-1}$  = frequency of the class preceding the modal class

$f_{m+1}$  = frequency of the class following the modal class

$h$  = class interval size of the modal class

**Example 3.28** Find the mode of the following data.



**3.5 The Harmonic Mean** The harmonic mean,  $H$ , of a set of  $n$  values  $X_1, X_2, \dots, X_n$  is the reciprocal of the arithmetic mean of the reciprocals of the values. The mean of the reciprocals

$$\frac{1}{X_1}, \frac{1}{X_2}, \dots, \frac{1}{X_n} \text{ is } \left( \frac{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}}{n} \right).$$

Hence the harmonic mean,  $H$ , is given by

$$H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^n \left( \frac{1}{X_i} \right)} = \frac{n}{\sum \left( \frac{1}{X} \right)} \quad (3.14)$$

In practice it is easier to remember

$$\frac{1}{H} = \frac{\sum_{i=1}^n \left( \frac{1}{X_i} \right)}{n} = \frac{1}{n} \sum \left( \frac{1}{X} \right) \quad (3.15)$$

**Example 3.16.** (a) The harmonic mean of the values 2, 4 and 8 is

$$H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{7/8} = 3 \times \frac{8}{7} = \frac{24}{7} = 3.43.$$

The geometric mean and the arithmetic mean of these values are 4 and 4.67 respectively. This shows that the harmonic mean of values (not all equal) is less than their geometric mean which in turn is less than their arithmetic mean.

**Example 3.16(b)** The harmonic mean of the values 3, 5, 6, 6, 7, 10 and 12 is

$$H = \frac{7}{\frac{1}{3} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6} + \frac{1}{7} + \frac{1}{10} + \frac{1}{12}} = \frac{7}{0.3333 + 0.2000 + 0.1667 + 0.1667 + 0.1429 + 0.1000 + 0.0833} = \frac{7}{1.1929} = 5.87.$$

**Example 3.16(c)** The reciprocals of the values of the variable  $X$  are 0.0500, 0.0400, 0.0200, 0.0285 and 0.0143. Find the A. M. and H. M. of  $X$ .

**Solution** The values of the variable  $X$  are  $\frac{1}{0.0500} = 20$ ,  $\frac{1}{0.0400} = 25$ ,  $\frac{1}{0.0200} = 50$ ,  $\frac{1}{0.0285} = 35.09$ , and  $\frac{1}{0.0143} = 69.93$ . The A. M. of  $X$  is given by



The relation is

**3.4 The Geometric Mean** The geometric mean,  $G$ , of a set of  $n$  positive values  $X_1, X_2, \dots, X_n$  is the  $n$ th root of the product of the values.

Thus

$$G = \sqrt[n]{X_1 \cdot X_2 \cdot \dots \cdot X_n} = (X_1 \cdot X_2 \cdot \dots \cdot X_n)^{1/n} \quad (3.10)$$

**Example 3.11** The geometric mean of the values 2, 4 and 8 is

$$G = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = 4.$$

In practice, it is difficult to extract higher roots. The geometric mean is, therefore, computed using logarithms. It is given by

$$\begin{aligned} \log G &= \frac{[\log X_1 + \log X_2 + \dots + \log X_n]}{n} = \frac{1}{n} \left( \sum_{i=1}^n \log X_i \right) \\ &= \frac{\sum \log X}{n} \end{aligned} \quad (3.11)$$

Here we assume that all the values are positive, otherwise the logarithms are not defined.

**Example 3.12** Find the geometric mean of the values

- (i) 3, 5, 6, 6, 7, 10, 12.
- (ii) 7.96, 13.82, 22.95, 35.34.

**Solution**

$$\begin{aligned} \text{(i) } \log G &= \frac{(\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 10 + \log 12)}{7} \\ &= \frac{0.47712 + 0.69897 + 0.77815 + 0.77815 + 0.84510 + 1.00000 + 1.07918}{7} \\ &= 5.65667/7 = 0.8081 \end{aligned}$$

$$G = \text{antilog}(0.8081) = 6.42836.$$

The arithmetic mean of these values is

$$\bar{X} = \frac{3 + 5 + 6 + 6 + 7 + 10 + 12}{7} = \frac{49}{7} = 7.$$

This illustrates that the geometric mean of a set of values (not all equal) is less than their arithmetic mean. Moreover, if any one of the original values is zero, their geometric mean is zero.

$$\begin{aligned} \text{(ii) } \log G &= \frac{\log 7.96 + \log 13.82 + \log 22.95 + \log 35.34}{4} \\ &= \frac{0.90091 + 1.14051 + 1.36078 + 1.54827}{4} \\ &= 4.95047/4 = 1.23762 \end{aligned}$$

$$G = \text{antilog}(1.23762)$$



# Chapter # 3

## Measures of dispersion

**1.1 Introduction** We have discussed the measures of central tendency or location in the preceding chapter. In that chapter, we attempted to find a single value (e.g. mean, median, mode) which would help us to find the centre of a distribution. A measure of location or an average alone is, however, not sufficient to describe all the characteristics of a distribution. Two or more distributions may have the same average and yet may differ from each other in other respects. Consider the following distributions:

|     |    |    |    |    |    |    |    |    |                |
|-----|----|----|----|----|----|----|----|----|----------------|
| I   | 67 | 67 | 67 | 67 | 67 | 67 | 67 | 67 | $\bar{X} = 67$ |
| II  | 66 | 66 | 66 | 67 | 67 | 68 | 68 | 68 | $\bar{X} = 67$ |
| III | 52 | 53 | 61 | 67 | 71 | 72 | 78 | 82 | $\bar{X} = 67$ |
| IV  | 43 | 44 | 50 | 55 | 66 | 90 | 91 | 97 | $\bar{X} = 67$ |

All the above distributions have the same mean, viz. 67. But these distributions differ greatly in their dispersion. (By *dispersion*, we mean the extent to which the values are spread out from the average) In the first distribution, all the values are equal to the mean. Thus there is no dispersion. In the second distribution, all the values lie within one unit of the mean on either side. In the third distribution, only one value (viz. 67) is equal to the mean. Some of the values in this distribution are as many as 15 units away from the mean. In the fourth distribution, only one value (viz. 66) is close to the mean and some of the values are as far away from the mean as 30 units. Although there is a great difference in the dispersion of the values of the distributions, yet each of these distributions is described by the same mean, 67. We, therefore, need a measure which would tell us how dispersed the data are. The measures used for this purpose are called *measures of dispersion* or *measures of variation*.

The most common measures of dispersion are

- (i) The Range
- (ii) The Semi-Interquartile Range or The Quartile Deviation
- (iii) The Mean Deviation or The Average Deviation
- (iv) The Standard Deviation

**1.2 The Range** The *range* is the simplest measure of dispersion. It is defined as the difference between the largest value and the smallest value in the data. If  $X_0$  is the smallest value and  $X_m$  is the largest value, then the range, denoted by  $R$ , is defined as

$$R = X_m - X_0$$



**Example 4.1** For the set of values 13, 23, 11, 17, 25, 18, 14 and 24, largest value is 25 and the smallest value is 11. Thus the range is  $25 - 11 = 14$ .

Sometimes the range is indicated simply by writing the smallest and the largest values. For instance, in Example 4.1, the range could be indicated as 11 to 25 or  $11 - 25$ .

The range is specially easy to find if the data have been arranged in order of magnitude; one merely notes the largest and the smallest values and finds the difference between them.

**Example 4.2** Find the range for each of the following sets of data:

- (a) 13 7 3 6 16 5 18 12    (b) 7 3 8 7 8 7 8 7 18

**Solution** Arranging the data in ascending order, we have

- (a) 3 5 6 7 12 13 16 18    (b) 3 7 7 7 7 8 8 8 18

In both cases, the largest value is 18 and the smallest value is 3, the range being  $18 - 3 = 15$ .

**4.3 The Range for Grouped Data** When the data have been grouped into a frequency table, the smallest possible value is the lower class boundary of the lowest class and the greatest possible value is the upper class boundary of the highest class. The range for grouped data may therefore, be defined as the difference between the upper class boundary of the highest class and the lower class boundary of the lowest class. We may also define the range for grouped data as the difference between the class mark of the highest class and the class mark of the lowest class. Both values of the range are approximate and either approximation is good enough. The latter definition, however, tends to eliminate the extreme cases to some extent.

**Example 4.3** Find the range for the frequency distribution of weights of 120 students given in Table 3.1 (Chapter 3).

**Solution** Range = upper class boundary of the highest class  
- lower class boundary of the lowest class  
=  $219.5 - 109.5 = 110$  pounds

Range = class

$$2/|X - \bar{X}| = 286$$

86  
05  
2  
1  
0

**4.6 The Standard Deviation** We have seen that the range is not stable on account of its dependence on extreme values whose size is largely a matter of chance. The semi-interquartile range excludes half of the items from consideration. The mean deviation neglects the fact that some deviations are negative and some are positive; it treats all of them as positive. Consequently we need some measure of variation which is free from these demerits. The standard deviation is one such measure of dispersion which is, to a considerable extent, free from these demerits.

The *standard deviation* is defined as the positive square root of the mean of the squared deviations of the values from their mean. The standard deviation of a set of  $n$  values,  $X_1, X_2, \dots, X_n$ , denoted by  $S$ , is

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} \quad (4.7)$$



**Example 4.8(b)** Find the standard deviation for the values 2, 3, 6, 8, 11.

**Solution** Here  $\bar{X} = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$ .

$$\begin{aligned} S &= \sqrt{\frac{\sum (X - \bar{X})^2}{n}} \\ &= \sqrt{\frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5}} \\ &= \sqrt{\frac{16+9+0+4+25}{5}} = \sqrt{\frac{54}{5}} = \sqrt{10.8} = 3.286. \end{aligned}$$

**4.6.1 The Standard Deviation for Grouped Data** In case of a frequency distribution with  $X_1, X_2, \dots, X_k$  as class marks and  $f_1, f_2, \dots, f_k$  as the corresponding class frequencies, the standard deviation is given by

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^k f_i (X_i - \bar{X})^2} = \sqrt{\frac{\sum f(X - \bar{X})^2}{n}} \quad (4.8)$$

where  $n = f_1 + f_2 + \dots + f_k = \sum f$ .

**Example 4.9** Find standard deviation for the frequency distribution of marks in Example 3.2(b) (Chapter 3).

**Solution** From Example 3.2(b),  $\bar{X} = 39$ . Computation of the standard deviation is outlined in the following table. Using Formula (4.8), we get

$$S = \sqrt{\frac{\sum f(X - \bar{X})^2}{n}} = \sqrt{\frac{2350}{50}} = \sqrt{47} = 6.86 \text{ or } 7 \text{ marks.}$$

|                   | $f$ | $X$ | $(X - \bar{X})$ | $f(X - \bar{X})$ | $f(X - \bar{X})^2$             |
|-------------------|-----|-----|-----------------|------------------|--------------------------------|
| 20 - 24           | 1   | 22  | -17             | -17              | 289                            |
| 25 - 29           | 4   | 27  | -12             | -48              | 576                            |
| 30 - 34           | 8   | 32  | -7              | -56              | 392                            |
| 35 - 39           | 11  | 37  | -2              | -22              | 44                             |
| 40 - 44           | 15  | 42  | 3               | 45               | 135                            |
| 45 - 49           | 9   | 47  | 8               | 72               | 576                            |
| 50 - 54           | 2   | 52  | 13              | 26               | 338                            |
| $n = \sum f = 50$ |     |     |                 |                  | $\sum f(X - \bar{X})^2 = 2350$ |

**4.7 The Variance** The variance is defined as the square of the standard deviation, i.e. the mean of the squared deviations from the mean. It is given by

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\sum (X - \bar{X})^2}{n} \quad (\text{for ungrouped data})$$

$$S^2 = \frac{1}{n} \sum_{i=1}^k f_i (X_i - \bar{X})^2 = \frac{\sum f(X - \bar{X})^2}{n} \quad (\text{for grouped data})$$



Corresponding to Formulae (4.9) to (4.13) for computing standard deviation, short formulae for computing variance for ungrouped and grouped data are

$$S^2 = \frac{\sum X^2}{n} - \left( \frac{\sum X}{n} \right)^2 = \frac{\sum D^2}{n} - \left( \frac{\sum D}{n} \right)^2 \quad (\text{for ungrouped data})$$

$$\left. \begin{aligned} S^2 &= \frac{\sum X^2}{n} - \left( \frac{\sum X}{n} \right)^2 \\ &= \frac{\sum D^2}{n} - \left( \frac{\sum D}{n} \right)^2 \\ &= h^2 \left( \frac{\sum fu^2}{\sum f} - \left( \frac{\sum fu}{\sum f} \right)^2 \right) \end{aligned} \right\} \quad (\text{for grouped data})$$

While standard deviation is the most useful single measure of dispersion, the variance will be used as a measure of dispersion in some later chapters.

#### 4.8 Properties of the Standard Deviation and Variance

(i) The standard deviation or variance of a constant is zero. Symbolically, if  $X = a$  (a constant), then  $S.D. (a) = 0$  and  $\text{Var}(a) = 0$ .

(ii) The standard deviation and the variance are independent of origin, i.e. they remain unchanged when the values are increased or decreased by a constant. Symbolically

$$S.D. (X + a) = S.D. (X) \text{ and } \text{Var} (X + a) = \text{Var} (X)$$

$$S.D. (X - a) = S.D. (X) \text{ and } \text{Var} (X - a) = \text{Var} (X)$$

where  $a$  is a constant.

We have seen that the standard deviation for the frequency distribution of weights of 120 students is the same (19.03) as calculated in Examples 4.10 and 4.16.

(iii) When all the values are multiplied or divided by a constant, the standard deviation of these values is multiplied or divided by the constant and the variance is multiplied or divided by the square of the constant. Symbolically

$$S.D. (aX) = a S.D. (X) \text{ and } \text{Var} (aX) = a^2 \text{Var} (X)$$

$$S.D. \left( \frac{X}{a} \right) = \left( \frac{1}{a} \right) S.D. (X) \text{ and } \text{Var} \left( \frac{X}{a} \right) = \left( \frac{1}{a^2} \right) \text{Var} (X)$$

S.D.  $(Y) = a$  S.D.  $(X)$  and  $\text{Var}(Y) = a^2 \text{Var}(X)$ . Similarly, if  $Y = \frac{X}{a}$ ,  
 S.D.  $(Y) = \frac{1}{a}$  S.D.  $(X)$  and  $\text{Var}(Y) = \frac{1}{a^2} \text{Var}(X)$ .

(iv) If two sets of data consisting of  $n_1$  and  $n_2$  values have variances  $S_1^2$  and  $S_2^2$  respectively, the combined variance of both sets of data is given by

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} + \frac{n_1 + n_2}{(n_1 + n_2)^2} (\bar{X}_1 - \bar{X}_2)^2 \quad (4.14)$$

and the combined standard deviation is given by

$$S = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} + \frac{n_1 + n_2}{(n_1 + n_2)^2} (\bar{X}_1 - \bar{X}_2)^2}$$

Formula (4.14) is a weighted mean of two variances. This result can be generalized to more than two, i.e.  $k$  sets of data by the formula

$$S^2 = \frac{\sum_{i=1}^k n_i [S_i^2 + (\bar{X}_i - \bar{X})^2]}{\sum_{i=1}^k n_i} \quad (4.15)$$

(v) The variance of the sum or difference of two independent random variables is the sum of their respective variances. Thus if  $X$  and  $Y$  are independent random variables

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{and } \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

(vi) The variance has the minimal property. This means that the variance or the standard deviation is a minimum if and only if the deviations are taken from the mean. In other words,

$$\frac{1}{n} \sum_{i=1}^n (X_i - a)^2 \text{ is a minimum if and only if } a = \bar{X}.$$

This property provides the

# Chapter # 4

## Regression & Correlation



What is meant by regression?

Q.1

Ans.

Regression is concerned with the study of a relationship between two variables in such a way that one variable is dependent and the other variable is independent. For instance one can predict the height of a person when his weight is given with the help of a regression equation when some observed values of heights and weights are known. Similarly weight of person can be predicted if his height is given.

Q.2

Ans.

Define simple linear regression.

The relationship between one dependent variable and another independent variable such that the relationship is approximated by a straight line is called simple linear regression.

Q.3

Ans.

Explain the terms regression and linear regression.

Regression is concerned with the study of a relationship between two variables so that one variable is dependent and the other variable is independent. If such a relationship between two variables is approximately by a straight line, the relationship is called linear regression.

Q.4

Ans.

Explain the terms regressand and regressor.

In the study of simple regression, we have two variables. One variable is independent and the other is dependent. One can predict the value of dependent variable with the help of independent variable. The dependent variable is called regressand and the independent variable is called regressor.

Q.5 Explain the difference between fixed variable and random variable.

Ans.

The independent variable is also called a fixed variable because its value is decided by the experimenter and is fixed before the prediction. It is also called regressor or predictor. The variable which is influenced by the independent variable is called dependent variable. It is also called regressand or predictand. This variable is of random nature and cannot be determined exactly for a given value of X. It is also called random variable.

**Q.6** Differentiate between linear regression and curvilinear regression.

**Ans.** Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line is known as linear regression. When the relationship between the variables is not approximated by a straight line but approximated by a curve, the regression is called curvilinear regression.

**Q.7** What is meant by residual?

**Ans.** In regression analysis residual means the difference between the observed value of a variable and the corresponding estimated value of a variable

$$Y - \hat{Y} = e$$

**Q.8** Discuss the method of least squares for fitting the regression lines of Y on X and X on Y.

**Ans.** (1) The regression line of Y and X is  $\hat{Y} = a + bX$  and the regression line of X on Y is  $\hat{X} = a + bY$

(2) Since we have to find the unknown values of a and b, we must have two equations in both the cases. i.e.

$$\Sigma Y = na + b \Sigma X$$

$$\Sigma X = na + b \Sigma Y$$

$$\Sigma XY = a \Sigma X + b \Sigma X^2$$

$$\Sigma XY = a \Sigma Y + b \Sigma Y^2$$

(3) We can find the values of a and b by solving the two equations simultaneously in both the cases.

**Q.9** What is meant by regression coefficient?

**Ans.** The average change in the dependent variable for a unit change in the independent variable is called regression coefficient. The regression coefficient may be positive or negative, depending on the relationship between the two variables.

**Q.10** Write a short note on slope of regression line.

**Ans.** The average change in the dependent variable for a unit change in the independent variable is called slope of regression line. The slope of regression line may be positive or negative, depending on the relationship between the two variables.

**Q.11** Differentiate between linear regression and non-linear regression.

**Ans.** Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line is known as linear regression. When the relationship between the two variables is not approximated by a straight line, the regression is called non-linear regression.



Q.12 Write down the properties of regression coefficients.

Ans. Properties of regression coefficients are:

- (1) The range of regression coefficient ( $b_{yx}$  or  $b_{xy}$ ) is  $-\infty$  to  $+\infty$ .
- (2) If one of the regression coefficient is greater than one, the other is less than one.
- (3) If X and Y are independent random variables, then regression coefficients are zero.
- (4) The correlation coefficient  $r$  is the geometric mean (square root) of the two regression coefficients.
- (5) The signs of regression coefficients and correlation coefficient are always the same.
- (6) The mean of the regression coefficients is greater than or equal to the correlation coefficient.

Q.13 Write a short note on scatter diagram.

Ans. Scatter diagram is a graphic picture of the observed sample data. Suppose a random sample of  $n$  pairs of observations such as  $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$  are plotted on a rectangular co-ordinate system taking independent variable on X-axis and the dependent variable on Y-axis. The diagram so obtained is called a scatter diagram.

14 Write down the properties of the least squares regression line.

s. The regression line  $\hat{Y} = a + bX$  has the following properties:

- (1) The regression line always passes through the means of the two variables  $\bar{X}$  and  $\bar{Y}$ .
- (2) The sum of deviations of observed values and estimated values is always zero i.e.  $\sum(Y - \hat{Y}) = 0$ .

Write down any three formulas of the regression coefficients of Y on X and X on Y.

The Regression Coefficient  
of Y on X is:

$$(1) \quad b_{yx} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$(2) \quad b_{yx} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$(3) \quad b_{yx} = \frac{n\sum D_x D_y - (\sum D_x)(\sum D_y)}{n\sum D_x^2 - (\sum D_x)^2}$$

The Regression Coefficient  
of X on Y is:

$$(1) \quad b_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}$$

$$(2) \quad b_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$$

$$(3) \quad b_{xy} = \frac{n\sum D_x D_y - (\sum D_x)(\sum D_y)}{n\sum D_y^2 - (\sum D_y)^2}$$



Q.12 Write down the properties of regression coefficients.

Ans. Properties of regression coefficients are:

- (1) The range of regression coefficient ( $b_{yx}$  or  $b_{xy}$ ) is  $-\infty$  to  $+\infty$ .
- (2) If one of the regression coefficient is greater than one, the other is less than one.
- (3) If X and Y are independent random variables, then regression coefficients are zero.
- (4) The correlation coefficient  $r$  is the geometric mean (square root) of the two regression coefficients.
- (5) The signs of regression coefficients and correlation coefficient are always the same.
- (6) The mean of the regression coefficients is greater than or equal to the correlation coefficient.

Q.13 Write a short note on scatter diagram.

Ans. Scatter diagram is a graphic picture of the observed sample data. Suppose a random sample of  $n$  pairs of observations such as  $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$  are plotted on a rectangular co-ordinate system taking independent variable on X-axis and the dependent variable on Y-axis. The diagram so obtained is called a scatter diagram.

Q.14 Write down the properties of the least squares regression line.

Ans. The regression line  $\hat{Y} = a + bX$  has the following properties:

- (1) The regression line always passes through the means of the two variables  $\bar{X}$  and  $\bar{Y}$ .
- (2) The sum of deviations of observed values and estimated values is always zero i.e.  $\sum(Y - \hat{Y}) = 0$ .

Write down any three formulas of the regression coefficients of Y on X and X on Y.

The Regression Coefficient of Y on X is:

$$(1) \quad b_{yx} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$(2) \quad b_{yx} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$(3) \quad b_{yx} = \frac{n\sum D_x D_y - (\sum D_x)(\sum D_y)}{n\sum D_x^2 - (\sum D_x)^2}$$

The Regression Coefficient of X on Y is:

$$(1) \quad b_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}$$

$$(2) \quad b_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$$

$$(3) \quad b_{xy} = \frac{n\sum D_x D_y - (\sum D_x)(\sum D_y)}{n\sum D_y^2 - (\sum D_y)^2}$$

Write down the normal equations for  $Y = a + bX$  and  $X = c + dY$ .

16. (1) **Normal Equations for  $Y = a + bX$**   
 $\Sigma Y = na + b\Sigma X$  and  $\Sigma XY = a\Sigma X + b\Sigma X^2$
- (2) **Normal Equations for  $X = c + dY$**   
 $\Sigma X = nc + d\Sigma Y$  and  $\Sigma XY = c\Sigma Y + d\Sigma Y^2$

17 Define correlation.

18. Correlation is the interdependence between two variables which means that how the increase or decrease in one variable brings the increase or decrease in other variable. For instance the increase of age is associated with the increase of weight in children.

19 Differentiate between linear and non-linear correlation.

20. When the ratio of variations between the variables is constant, the correlation is said to be linear correlation. When the ratio of variations between the variables is fluctuating, the correlation is said to be non-linear correlation.

21 Distinguish between positive and negative correlation.

22. If the increase in one variable is associated with increase in the other variable or decrease in one variable is associated with decrease in the other variable, the correlation between the two variables is positive.

On the other hand, if the increase in one variable is associated with the decrease in the other variable, the correlation is said to be negative or inverse.

23 Define the terms no correlation and curvilinear correlation.

24. When two variables result in the answer of the correlation coefficient  $r = 0$ , there will be no correlation between the variables. When points on a scatter diagram are not in a straight line but in a curvilinear position, the relationship is said to be curvilinear

Write down the properties of the correlation coefficient.

The important properties of the correlation coefficient are given below:

- (1) The correlation coefficient  $r$  is symmetrical with respect to the variable  $X$  and  $Y$  i.e.  $r_{xy} = r_{yx}$ .
- (2) The correlation coefficient  $r$  is free from units of measurements.
- (3) The correlation coefficient  $r$  lies between  $-1$  and  $+1$  i.e.  $-1 \leq r \leq +1$
- (4) The correlation coefficient  $r$  is independent of origin and scale i.e.  $r_{xy} = r_{uv}$ .
- (5) The correlation coefficient  $r$  is the geometric mean of two regression coefficients i.e.  $r = \pm \sqrt{b_{xy} \cdot b_{yx}}$
- (6) If  $X$  and  $Y$  are independent variates, then they are uncorrelated i.e.  $r_{xy} = 0$



Q.27 Write down the formula of correlation coefficient:

(a) When  $b_{yx}$  and  $b_{xy}$  are negative.

(b) When  $b_{yx}$  and  $b_{xy}$  are positive

Ans. (a)  $r_{xy} = -\sqrt{b_{yx} \cdot b_{xy}}$ , the correlation coefficient will be negative when  $b_{yx}$  and  $b_{xy}$  are negative.

(b)  $r_{xy} = +\sqrt{b_{yx} \cdot b_{xy}}$ , the correlation coefficient will be positive when  $b_{yx}$  and  $b_{xy}$  are positive.

Q.28 Differentiate between regression and correlation.

Ans. Correlation is the interdependence between two variables. Whereas regression is the prediction of the value of dependent variable with the help of an independent variable.

Q.29 Define the terms regression analysis and correlation analysis.

Ans. The regression analysis is the study of a relationship between one dependent variable and another independent variable. The correlation analysis is helpful in finding that how strong or weak is the relationship between two variables and in which direction.

Q.30 Write down the aims of regression and correlation analysis.

Ans. The aims of regression and correlation analysis are:

- (1) Regression analysis provides estimates of the dependent variable for given values of the independent variable.
- (2) Regression analysis provides measures of errors that are likely to be involved in using the regression line to estimate the dependent variable.
- (3) Regression analysis provides an estimate of the effect on the mean value of Y of a unit change in X.
- (4) Correlation analysis helps us to find that how strong or weak is the relationship between two variables.

Q.31 Differentiate between perfect positive and perfect negative correlation.

Ans. When the movement in two variables is in the same direction and increase or decrease is in a proportionate manner, there is perfect positive correlation between the variables. The answer of the correlation coefficient will be equal to + 1.

When the movement in two variables is in opposite direction and the movement is proportionate, there is perfect negative correlation. The answer of the correlation coefficient will be equal to - 1.



**Q.1** Compute the regression equation of Y on X from the following data using normal equations.

|   |    |    |    |    |    |
|---|----|----|----|----|----|
| X | 25 | 30 | 40 | 50 | 65 |
| Y | 6  | 5  | 4  | 8  | 7  |

**Solution:**

The regression equation of Y on X is  $Y = a + bX$

The normal equations are  $\Sigma Y = na + b \Sigma X$  and  $\Sigma XY = a \Sigma X + b \Sigma X^2$

The necessary calculations are given below:

| X                | Y               | XY                 | $X^2$               |
|------------------|-----------------|--------------------|---------------------|
| 25               | 6               | 150                | 625                 |
| 30               | 5               | 150                | 900                 |
| 40               | 4               | 160                | 1600                |
| 50               | 8               | 400                | 2500                |
| 65               | 7               | 455                | 4225                |
| $\Sigma X = 210$ | $\Sigma Y = 30$ | $\Sigma XY = 1315$ | $\Sigma X^2 = 9850$ |

Substituting the values from the table in the normal equations, we have

$$30 = 5a + 210b \quad \dots\dots (1) \quad 1315 = 210a + 9850b \quad \dots\dots (2)$$

The values of a and b are given by solving (1) and (2). We multiply equation (1) by 42 and subtract from equation (2), we get

$$1315 = 210a + 9850b$$

$$1260 = 210a + 8820b$$

---


$$55 = 1030b \quad \text{or} \quad b = \frac{55}{1030} = 0.053$$

Substituting  $b = 0.053$  in equation (1), we get

$$30 = 5a + 210(0.053) \quad \text{or} \quad 5a = 30 - 11.13 = 18.87$$

$$\text{or } a = \frac{18.87}{5} = 3.774$$

Hence the regression equation of Y on X is  $\hat{Y} = 3.774 + 0.053 X$

**Q.2** The following sample observations were randomly selected:

|   |   |   |   |   |    |
|---|---|---|---|---|----|
| X | 4 | 5 | 3 | 6 | 12 |
| Y | 4 | 6 | 5 | 7 | 8  |

Determine the value of  $\hat{Y}$  when X is 7.

Solution

The necessary calculations are given below:

| X  | Y  | XY  | X <sup>2</sup> |
|----|----|-----|----------------|
| 4  | 4  | 16  | 16             |
| 5  | 6  | 30  | 25             |
| 3  | 5  | 15  | 9              |
| 6  | 7  | 42  | 36             |
| 12 | 8  | 96  | 144            |
| 30 | 30 | 199 | 230            |

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6 \text{ and } \bar{Y} = \frac{\Sigma Y}{n} = \frac{30}{5} = 6$$

The regression coefficient of Y on X is

$$b_{yx} = \frac{n \Sigma XY - (\Sigma X)(\Sigma Y)}{n \Sigma X^2 - (\Sigma X)^2} = \frac{5(199) - 30(30)}{5(230) - (30)^2} = \frac{95}{250} = 0.38$$

The regression equation of Y on X is

$$\hat{Y} - \bar{Y} = b_{yx} (X - \bar{X})$$

$$\hat{Y} - 6 = 0.38 (X - 6)$$

$$\hat{Y} = 6 + 0.38X - 2.28 = 3.72 + 0.38X$$

When  $X = 7$ , then  $\hat{Y} = 3.72 + 0.38(7) = 6.38$

Q.42 The marks of 8 students in Economics (X) and Statistics (Y) are given:

|   |    |    |    |    |    |    |    |    |
|---|----|----|----|----|----|----|----|----|
| X | 50 | 43 | 52 | 40 | 80 | 65 | 85 | 81 |
| Y | 68 | 57 | 70 | 55 | 90 | 75 | 95 | 66 |

Show that the correlation coefficient is the square root of regression coefficients.

**Solution:**

The necessary calculations are given below:

| X   | Y   | XY    | X <sup>2</sup> | Y <sup>2</sup> |
|-----|-----|-------|----------------|----------------|
| 50  | 68  | 3400  | 2500           | 4624           |
| 43  | 57  | 2451  | 1849           | 3249           |
| 52  | 70  | 3640  | 2704           | 4900           |
| 40  | 55  | 2200  | 1600           | 3025           |
| 80  | 90  | 7200  | 6400           | 8100           |
| 65  | 75  | 4875  | 4225           | 5625           |
| 85  | 95  | 8075  | 7225           | 9025           |
| 81  | 66  | 5346  | 6561           | 4356           |
| 496 | 576 | 37187 | 33064          | 42904          |

The correlation coefficient is

$$\begin{aligned}
 r_{xy} &= \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2] [n \sum Y^2 - (\sum Y)^2]}} \\
 &= \frac{8(37187) - 496(576)}{\sqrt{[8(33064) - (496)^2] [8(42904) - (576)^2]}} \\
 &= \frac{11800}{\sqrt{(18496)(11456)}} = \frac{11800}{14556.44792} = 0.81
 \end{aligned}$$

The regression coefficient of Y on X is

$$b_{yx} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{8(37187) - (496)(576)}{8(33064) - (496)^2} = \frac{11800}{18496} = 0.64$$

The regression coefficient of X on Y is

$$b_{xy} = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum Y^2 - (\sum Y)^2} = \frac{8(37187) - 496(576)}{8(42904) - (576)^2} = \frac{11800}{11456} = 1.03$$

$$\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{(0.64)(1.03)} = 0.81 = r_{xy}$$

Hence the correlation coefficient is the square root of regression coefficients.



# Chapter # 5

## Probability

07.1 Write short answers to the following questions.

Q.1. What is a random experiment? (B.I.S.E., Rawalpindi 2007)

Ans. A random experiment (or an experiment) is a process which generates raw data. For example, tossing a coin, rolling a die and drawing a ball from a bag containing balls of different colours are random experiments.

Q.2. State the properties of a random experiment.

Ans. A random experiment has two properties in common. Firstly, each experiment has several possible outcomes which can be described in advance. For example, in tossing a coin, the possible outcomes are head and tail.

Secondly, we are not certain about the outcome or result of the experiment. In tossing a coin, although the possible outcomes are a head and a tail, but we are not certain whether it will be a head or a tail.

Q.3. What is the difference between an outcome and an event?

Ans. An outcome is a particular result of an experiment, whereas an event is the collection of one or more outcomes of an experiment.

Q.4. Define the sample space.

Ans. The set or collection of all possible outcomes of an experiment is called the sample space. It is denoted by  $S$ .

Q.5. What is a sample point? (B.I.S.E., Multan 2009)

Ans. Each element of a sample space is called a *sample point*. For example, when we toss a coin, the sample space is  $S = \{H, T\}$  and each of the elements  $H$ (head) and  $T$ (tail) is a sample point.

Q.6. Distinguish between simple and compound events.

(B.I.S.E., Rawalpindi 2007)

Ans. When an event consists of only one sample point or outcome, it is called a simple or elementary event.

If the event consists of more than one sample points, it is called a compound event.

Q.7. Give an example each of simple and compound events.

Ans. When two coins are tossed, the event  $A = \{HH\}$  that two heads appear is a simple event but the event  $B = \{HT, TH\}$  that one head appears is a compound event.

Counting sample points: For experiments with large number of outcomes, it may be very difficult to list all the sample points. For this purpose, the following rules are used.

Law of multiplication, Suppose an experiment has 'n' possible outcomes and for each of these another experiment has m possible outcomes, then the total number of outcomes in which these experiments result when performed together is  $m \times n$ .

If a coin and a die are tossed together, then the total number of outcomes is  $2 \times 6 = 12$ .

Rule of combination, Suppose "r" objects are to be selected out of a total of "n" objects, such that their order of selection is not important, then these are known as combinations. Total number of such combinations is  $nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$  (Order of selection is not important)

Rule of permutation: Suppose "r" objects are to be selected out of a total of "n" objects such that their order of selection is important, then these are known as permutations. Total number of such permutations are  $nPr = \frac{n!}{(n-r)!}$  (Order of selection is important)

... of n objects



(iii) Let  $C$  denote the event that at most one head appears.

$$C = \{T, TT, HTT, THT, T, TH\}$$

$$P(C) = n(C)/n(S) = 4/8$$

✓ Question: If two fair dice are thrown once, what is the prob that the sum of two dots is  
(i) less than 4 (ii) more than 9 (iii) divisible by 8 (iv) between 7 and 10.

Solution:  $S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$

(i) Let  $A$  = sum of dots is less than 4.

$$A = \{(1,1), (1,2), (2,1)\} \Rightarrow P(A) = n(A)/n(S) = 3/36$$

(ii) Let  $B$  = sum of dots is more than 9

$$B = \{(4,6), (5,5), (5,6), (6,4), (6,5), (6,6)\}$$

$$P(B) = n(B)/n(S) = 6/36$$

(iii) Let  $C$  = sum of dots is divisible by 8

$$C = \{(2,6), (4,4), (3,5), (5,3), (6,2)\}$$

$$P(C) = n(C)/n(S) = 5/36$$

(iv) Let  $D$  = sum of dots is between 7 and 10.

$$D = \{(2,6), (3,5), (3,6), (4,4), (4,5), (5,3), (5,4), (6,2), (6,3)\} \Rightarrow P(D) = n(D)/n(S) = 9/36$$

(i) it is an even number (ii) it is an even number and divisible by 3?

Solution:  $S = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

(i) Let  $A =$  chosen number is even

$$A = \{4, 6, 8, 10, 12\} \Rightarrow P(A) = n(A)/n(S) = 5/10 = 1/2$$

(ii) Let  $B =$  the chosen number is even and divisible by 3

$$B = \{6, 12\} \Rightarrow P(B) = n(B)/n(S) = 2/10 = 1/5$$

✓ Question. In a single throw of two fair dice, find the prob. that the product of the dots is (i) between 8 and 16 (both inclusive) (ii) divisible by 4

Solution:  $S = \{(1,1), (1,2), \dots, (6,6)\}$

(i) Let  $A =$  the product of dots is between 8 and 16

$$A = \{(2,4), (2,5), (2,6), (3,3), (3,4), (3,5), (4,2), (4,3), (4,4), (5,2), (5,3), (6,2)\} \Rightarrow P(A) = n(A)/n(S) = 12/36 = 1/3$$

(ii) Let  $B =$  the product of dots is divisible by 4.

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

$$P(B) = n(B)/n(S) = 14/36 = 7/18$$

✓ Question. A card is drawn at random from a pack of 52 cards. Find the prob. that

it is (i) a king (ii) ace of spade (iii) a card between 2 and 6 (iv) face card (v) black queen

Solution,  $n(S) = \binom{52}{1} = 52$

(i) Let  $A$  = the selected card is a king

|       |        |
|-------|--------|
| Kings | others |
| 4     | 48     |

$$n(A) = \binom{4}{1} \binom{48}{0} = 4 \times 1 = 4 \Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{1}{13}$$

(ii) Let  $B$  = the selected card is ace of spade.

|              |        |
|--------------|--------|
| Ace of spade | others |
| 1            | 51     |

$$n(B) = \binom{1}{1} \binom{51}{0} = 1 \times 1 = 1 \Rightarrow P(B) = \frac{n(B)}{n(S)} = \frac{1}{52}$$

(iii) Let  $C$  = the selected card is between 2 and 6.

|                       |        |
|-----------------------|--------|
| $2 < \text{card} < 6$ | others |
| 12                    | 40     |

$$n(C) = \binom{12}{1} \binom{40}{0} = 12 \times 1 = 12 \Rightarrow P(C) = \frac{n(C)}{n(S)} = \frac{3}{13}$$

(iv) Let  $D$  = the selected card is face card

|            |        |
|------------|--------|
| face cards | others |
| 12         | 40     |

$$n(D) = \binom{12}{1} \binom{40}{0} = 12 \times 1 = 12 \Rightarrow P(D) = \frac{n(D)}{n(S)} = \frac{3}{13}$$

(v) Let  $E$  = the selected card is black queen.

|              |        |
|--------------|--------|
| Black queens | others |
| 2            | 50     |

$$n(E) = \binom{2}{1} \binom{50}{0} = 2 \times 1 = 2 \Rightarrow P(E) = \frac{n(E)}{n(S)} = \frac{1}{26}$$



### Addition law of probability for mutually exclusive events:

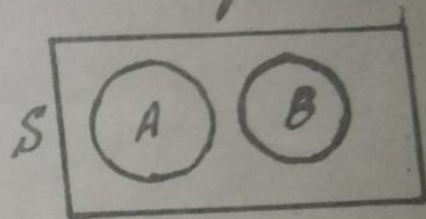
Let  $A$  and  $B$  be two mutually exclusive events, then the probability that one of these occur, denoted by  $P(A \cup B)$ , is given as

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

Proof, Suppose a sample space consists of  $n$  sample points. Further suppose two mutually exclusive events  $A$  and  $B$  having  $m_1$  and  $m_2$  sample points respectively.

$$\text{Then } P(A) = \frac{n(A)}{n(S)} = \frac{m_1}{n}$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{m_2}{n}$$



Since  $A$  and  $B$  are mutually exclusive events, therefore, the number of sample points in  $A \cup B$  is  $m_1 + m_2$  i.e.  $n(A \cup B) = m_1 + m_2$

$$\begin{aligned} \therefore P(A \cup B) &= \frac{n(A \cup B)}{n(S)} = \frac{(m_1 + m_2)}{n} = \frac{m_1}{n} + \frac{m_2}{n} \\ &= P(A) + P(B) \end{aligned}$$

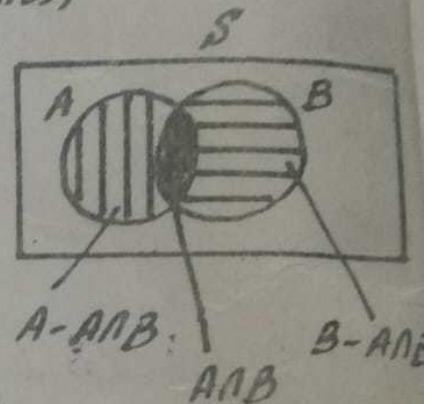
$$P(B) \geq P(A)$$

$$\text{since } P(A \cap B) \geq 0$$

\* General law of addition. If  $A$  and  $B$  are any two events, then the probability that at least one of these occur, denoted by  $P(A \cup B)$ , is given as  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof. If  $A$  and  $B$  are any two events, then the event  $A \cup B$  can be written as the union of three mutually exclusive events,  $A - A \cap B$ ,  $A \cap B$  and  $B - A \cap B$  i.e.

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



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

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

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

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

✓ Question, A marble is drawn at random from a box containing 10 red, 30 white, 20 blue and 15 orange marbles. Find the prob. that it is (i) orange or red (ii) not red or blue (iii) not blue (iv) red, white or blue.

Solution:

| Red | White | Blue | Orange | Total |
|-----|-------|------|--------|-------|
| 10  | 30    | 20   | 15     | 75    |

$$n(S) = \binom{75}{1} = 75$$

ii) Let A denote the event that orange marble is drawn and B denote the event that red marble is drawn.

$$n(A) = \binom{15}{1} \binom{60}{0} = 15 \times 1 = 15 \Rightarrow P(A) = n(A)/n(S) = 15/75$$

$$n(B) = \binom{10}{1} \binom{65}{0} = 10 \times 1 = 10 \Rightarrow P(B) = n(B)/n(S) = 10/75$$

$$P(\text{orange or red}) = P(A \cup B) = P(A) + P(B) = 15/75 + 10/75 = 1/3$$

iii) Let C denote the event that blue marble is drawn

$$n(C) = \binom{20}{1} \binom{55}{0} = 20 \times 1 = 20 \Rightarrow P(C) = n(C)/n(S) = 20/75$$

$$P(\text{red or blue}) = P(B \cup C) = P(B) + P(C) = 10/75 + 20/75 = 30/75$$

$$P(\text{not "red or blue"}) = 1 - P(\text{red or blue}) = 1 - 30/75 = 3/5$$

OR

Let D denote that event that not "red or blue"

$$P(D) = n(D)/n(S) = \binom{10+20}{0} \binom{30+15}{1} / 75 = 3/5$$

$$(iii) P(\text{not blue}) = 1 - P(\text{blue}) = 1 - P(C) = 1 - 20/75 = 11/15$$



Let  $E$  denote the event that white marble is drawn  
 $n(E) = \binom{30}{1} \binom{45}{0} = 30 \times 1 = 30 \Rightarrow P(E) = \frac{n(E)}{n(S)} = \frac{30}{75}$

$$P(\text{red, white or blue}) = P(B \cup E \cup C) = P(B) + P(E) + P(C) \\ = \frac{10}{75} + \frac{30}{75} + \frac{20}{75} = \frac{60}{75} = \frac{4}{5}$$

\* Question: One integer is chosen at random from the numbers 1, 2, 3, ..., 50. What is the prob. that the chosen number is divisible by 6 or by 8?

Solution:  $S = \{1, 2, \dots, 50\}$  or  $\binom{50}{1} = 50$

Let the event  $A$  = the number is divisible by 6  
 and the event  $B$  = " " " " " 8

$$A = \{6, 12, 18, 24, 30, 36, 42, 48\} \Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{8}{50}$$

$$B = \{8, 16, 24, 32, 40, 48\} \Rightarrow P(B) = \frac{n(B)}{n(S)} = \frac{6}{50}$$

$$A \cap B = \{24, 48\} \Rightarrow P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{50}$$

$$P(\text{number divisible by 6 or by 8}) = P(A \cup B) =$$

$$P(A) + P(B) - P(A \cap B) = \frac{8}{50} + \frac{6}{50} - \frac{2}{50} = \frac{6}{25}$$

Question: A class contains 10 men and 20 women of which half the men and half the women have brown eyes. Find the prob. that the person chosen at random is man or has brown eyes.

Solution:

|            | Men | Women |    |
|------------|-----|-------|----|
| Brown eyes | 5   | 10    | 15 |
| Other eyes | 5   | 10    | 15 |
|            | 10  | 20    | 30 |

$$n(S) = \binom{30}{1} = 30$$

Let the event  $A$  = the person chosen is man  
 and " "  $B$  = " " " " has brown eyes.

$$n(A) = \binom{10}{1} \binom{20}{0} = 10 \times 1 = 10 \Rightarrow P(A) = n(A)/n(S) = 10/30$$

$$n(B) = \binom{15}{1} \binom{15}{0} = 15 \times 1 = 15 \Rightarrow P(B) = n(B)/n(S) = 15/30$$

The event  $A \cap B$  = the person chosen is man and has brown eyes.

$$n(A \cap B) = \binom{5}{1} \binom{25}{0} = 5 \times 1 = 5 \Rightarrow P(A \cap B) = n(A \cap B)/n(S) = 5/30$$

$$P(\text{the person is man or has brown eyes}) = P(A \cup B) =$$

$$P(A) + P(B) - P(A \cap B) = 10/30 + 15/30 - 5/30 = 2/3$$

Question: The prob. that a student's passing Maths is  $2/3$  and the prob. of passing English is  $4/9$ . If the prob. of passing at least one subject is  $4/5$ , what is the prob. that he will pass both subjects?

Solution: Let the event  $A$  = the student passes Maths  
and " "  $B$  = " " " English  
Then the event  $A \cup B$  = the student passes at least one  
and " "  $A \cap B$  = " " " both

$$\text{Then } P(A) = 2/3 ; P(B) = 4/9 ; P(A \cup B) = 4/5$$

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

$$4/5 = 2/3 + 4/9 - P(A \cap B) \Rightarrow P(A \cap B) = 14/45$$

# Chapter # 6

## **Testing of Hypothesis**



## SHORT QUESTIONS

- Q.1 Describe the hypothesis.  
Ans. Hypothesis is a statement which may or may not appears be true after conclusion. or  
A statement about a population parameter developed for the purpose of testing.
- Q.2 Define the hypothesis testing.  
Ans. A procedure based on sample evidence and probability theory to determine whether the hypothesis is a reasonable statement or not is called hypothesis testing.
- Q.3 Explain the terms hypothesis and tests of hypothesis.  
Ans. Any notion that is formed about the population is called hypothesis.  
The procedures which are adopted to examine whether a certain opinion or notion about the population is true or false are called tests of hypothesis e.g. Z-test, t-test and  $\chi^2$ -test etc.
- Q.4 Define a null hypothesis.  
Ans. A null hypothesis is any hypothesis which is tested for possible rejection or acceptance under the assumption that it is true.
- 5 Write down the definition of alternative hypothesis or research hypothesis.  
Ans. A statement specifying that the population parameter is some value other than the one specified under the null hypothesis.
- 6 Define a simple hypothesis.  
Ans. A hypothesis is said to be a simple hypothesis if the hypothesis uniquely specifies the distribution from which the sample is taken.
- What is meant by composite hypothesis?  
Ans. A hypothesis which does not specify all values of parameters of a distribution is called composite hypothesis. or  
A hypothesis is said to be a composite hypothesis if it does not completely specify the probability distribution.

**Q.8** Write down the definition of tests of significance.

**Ans.** A significance test is a statistical test laying down the procedure for deciding whether to accept or reject a statistical hypothesis.

**Q.9** What is meant by a statistical hypothesis?

**Ans.** Any opinion or idea which is formed about the population under study is called a statistical hypothesis e.g. the average I.Q. scores of a university students is 118, a medicine of allergy gives relief to 80% of the patients etc. In simple words, a statistical hypothesis is a statement about the unknown value of the population parameter. The statement may be true or false.

**Q.10** Describe a two-tailed test.

**Ans.** A statistical test in which the critical region is located at both ends of sampling distribution is known as two-tailed test.

**Q.11** Explain the difference between one-sided and two-sided tests. When should each be used?

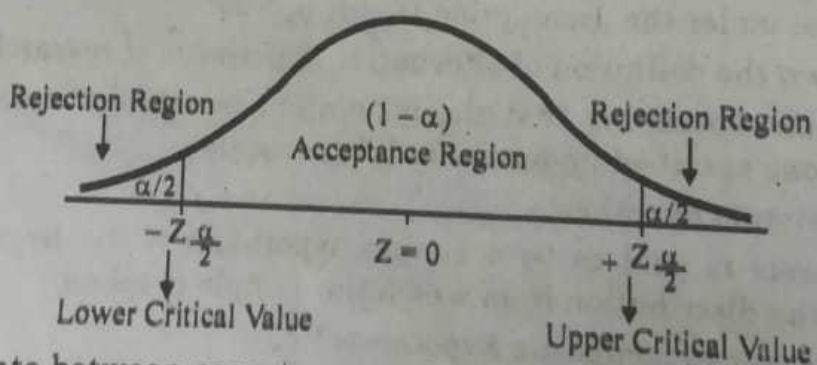
**Ans.** When critical region is located on one end of the sampling distribution, either on left end or right end, the test to be applied is called a one-sided test. Whereas in two-sided test the critical region is located in both ends of a sampling distribution. When alternative hypothesis consists of inequality sign as "less than" or "greater than", we use one-tailed test. If the alternative hypothesis contains inequality sign ( $\neq$ ), we use two-tailed test.

**Q.12** Define the acceptance region.

**Ans.** The portion of the area under a curve that includes those values of a statistic that lead to acceptance of the null hypothesis.

**Q.13** Explain with example the difference between acceptance region and rejection region.

**Ans.** Acceptance region consists of those values of the sampling distribution that leads us to accept the null hypothesis. The rejection region consists of those values of the sampling distribution that leads us to reject the null hypothesis. The rejection region is equal to  $\alpha$  and the acceptance region is denoted by  $1 - \alpha$



**Q.14** Differentiate between acceptance region and rejection region.

**Ans.** Acceptance region consists of those values of the sampling distribution that leads us to accept the null hypothesis. It means that the sample values have provided enough evidence to accept the null hypothesis. On the other hand, the rejection region consists of those values of the sampling distribution that leads us to reject the null hypothesis.



Q.15 What is meant by critical region?

Ans. A critical region consists of those values of the sampling distribution that leads us to reject the null hypothesis  $H_0$ . It means that the test has not given enough evidence that the null hypothesis could have been accepted. So we have to reject the null hypothesis. The critical region is equal to  $\alpha$ .

Q.16 What is meant by critical value?

Ans. The value which separates the rejection region from acceptance region is called critical value. In other words a critical value is a demarcation line between acceptance region and rejection region e.g.  $-\frac{Z_{\alpha}}{2}$ ,  $\frac{Z_{\alpha}}{2}$ ,  $-Z_{\alpha}$  or  $Z_{\alpha}$  are critical values.

Q.17 Explain the terms level of significance and tests of significance.

Ans. The level of significance is the probability of rejecting a true null hypothesis. It is denoted by  $\alpha$  and is usually 1% or 5%. A significance test is a statistical test laying down the procedure for deciding whether to accept or reject a statistical hypothesis.

Q.18 Define a type-I error.

Ans. If we reject a true null hypothesis, the error is called a type-I error or type-I error is the rejection of  $H_0$  when it is true.

Q.19 Differentiate between type-I error and type-II error.

Ans. If we reject a null hypothesis when it is actually true is called a type-I error. The probability of making type-I error is denoted by  $\alpha$ . If we accept a null hypothesis when it is actually false is called a type-II error. The probability of making type-II error is denoted by  $\beta$ .

Q.20 Differentiate between one-tailed test and two-tailed test.

Ans. In two-tailed test, the rejection region is located in both ends of the sampling distribution. In one-tailed test, the rejection region is located in one end of the sampling distribution, either in the right tail or in the left tail.

Q.21 Distinguish between null hypothesis and alternative hypothesis.

Ans. Null hypothesis is any hypothesis which is tested for possible rejection under the assumption that it is true. It is denoted by the symbol  $H_0$ .

Alternative hypothesis is formed versus the null hypothesis and it is accepted when null hypothesis has been rejected. It is denoted by the symbol  $H_1$  or  $H_A$ .

Q.22 Differentiate between simple hypothesis and composite hypothesis.

Ans. A hypothesis having a single value for the population parameter is called a simple hypothesis e.g.  $H_0: \mu = 10$  kg. is a simple hypothesis. Any hypothesis which specifies a range of values for the population parameter is called a composite hypothesis e.g.  $H_0: \mu \geq 10$  kg. is a composite hypothesis.

Q.23 What is meant by test-statistic?

Ans. A test-statistic is a rule or formula on which the decision whether to accept or reject a null hypothesis is based. The commonly used test statistics are "Z-statistic", "t-statistic" and " $\chi^2$ -statistic".



**Q.24** Define null hypothesis and describe the general procedure for its testing.

**Ans.** A null hypothesis denoted by the symbol  $H_0$  is any hypothesis which is to be tested for possible rejection under the assumption that it is true.

**General Procedure**

- (1) (a)  $H_0 : \theta = \theta_0$  (b)  $H_0 : \theta \geq \theta_0$  (c)  $H_0 : \theta \leq \theta_0$   
 $H_1 : \theta \neq \theta_0$   $H_1 : \theta < \theta_0$   $H_1 : \theta > \theta_0$
- (2) Level of significance  $\alpha$  is chosen.
- (3) A suitable test-statistic is selected.
- (4) We compute the value of test-statistic using the given information about sample.
- (5) Critical region is formed according to alternative hypothesis.
- (6) Conclusion whether to accept the null hypothesis or reject it.

**Q.25** Write a short note on power curve.

**Ans.** A graph of the probability of rejecting  $H_0$  for all possible values of the population parameter not satisfying the null hypothesis is known as power curve.

**Q.26** Define the terms power of a test and power curve.

**Ans.** The power of a test is the probability that the test will lead to a rejection of the null hypothesis  $H_0$  when, in fact, the alternative hypothesis  $H_1$  is true.

A graph of the probability of rejecting  $H_0$  for all possible values of the population parameter not satisfying the null hypothesis is known as power curve.

**Q.27** Describe the procedure for testing hypothesis about mean of a normal population when population standard deviation is known.

**Ans. General Procedure**

- (1) (a)  $H_0 : \mu = \mu_0$  and  $H_1 : \mu \neq \mu_0$  (Two-sided)  
 (b)  $H_0 : \mu \leq \mu_0$  and  $H_1 : \mu > \mu_0$  (One-sided)  
 (c)  $H_0 : \mu \geq \mu_0$  and  $H_1 : \mu < \mu_0$  (One-sided)
- (2) Choose the level of significance  $\alpha$
- (3) The test-statistic to be used is  $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$
- (4) The critical region is defined as:
  - (a)  $H_1 : \mu \neq \mu_0, |Z| > Z_{\frac{\alpha}{2}}$
  - (b)  $H_1 : \mu > \mu_0, Z > Z_{\alpha}$
  - (c)  $H_1 : \mu < \mu_0, Z < -Z_{\alpha}$
- (5) The calculation of the test-statistic
- (6) Conclusion: Reject  $H_0$  if  $Z$  lies in the critical region, otherwise accept it

Q.28 Explain the general procedure for testing of hypothesis regarding the population mean when population standard deviation is unknown and the sample size is large.

Ans. **General Procedure**

- (1) (a)  $H_0 : \mu = \mu_0$  and  $H_1 : \mu \neq \mu_0$  (Two-sided)
- (b)  $H_0 : \mu \leq \mu_0$  and  $H_1 : \mu > \mu_0$  (One-sided)
- (c)  $H_0 : \mu \geq \mu_0$  and  $H_1 : \mu < \mu_0$  (One-sided)
- (2) Choose the level of significance  $\alpha$
- (3) The test-statistic to be used is  $Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- (4) The critical region is defined as:
  - (a)  $H_1 : \mu \neq \mu_0, |Z| > Z_{\frac{\alpha}{2}}$
  - (b)  $H_1 : \mu > \mu_0, Z > Z_{\alpha}$
  - (c)  $H_1 : \mu < \mu_0, Z < -Z_{\alpha}$
- (5) The calculation of the test-statistic
- (6) Conclusion : Reject  $H_0$  if  $Z$  lies in the critical region, otherwise accept it

Q.29 Describe the procedure for testing hypothesis about mean of a normal population when population standard deviation is unknown and the sample size is small.

Ans. **General Procedure**

- (1) (a)  $H_0 : \mu = \mu_0$  and  $H_1 : \mu \neq \mu_0$  (Two-Sided)
- (b)  $H_0 : \mu \leq \mu_0$  and  $H_1 : \mu > \mu_0$  (One-Sided)
- (c)  $H_0 : \mu \geq \mu_0$  and  $H_1 : \mu < \mu_0$  (One-Sided)
- (2) Choose the level of significance  $\alpha$ .
- (3) The test-statistic to be used is  $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
- (4) The critical region is defined as:
  - (a)  $H_1 : \mu \neq \mu_0, |t| > t_{\frac{\alpha}{2}}(n-1)$
  - (b)  $H_1 : \mu > \mu_0, t > t_{\alpha}(n-1)$
  - (c)  $H_1 : \mu < \mu_0, t < -t_{\alpha}(n-1)$
- (5) The calculation of the test-statistic.
- (6) Conclusion: Reject  $H_0$  if  $t$  lies in the critical region, otherwise accept it.



Q.1 The heights of college male students are known to be normally distributed with a mean of 67.39 inches and  $\sigma = 1.30$  inches. A random sample of 400 students showed a mean height of 67.47 inches. Using a 0.05 level of significance, test the hypothesis  $H_0: \mu = 67.39$  against the alternative  $H_1: \mu > 67.39$ .

**Solution:**

1. Null hypothesis:  $H_0: \mu = 67.39$
2. Alternative hypothesis:  $H_1: \mu > 67.39$
3. Level of significance:  $\alpha = 0.05$

4. Test - statistic: 
$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

5. Critical region:  $Z > 1.645$

(From the area table of normal distribution, we have  $Z_\alpha = Z_{0.05} = 1.645$ )

6. Computations: Here,  $n = 400$ ,  $\bar{X} = 67.47$ ,  $\sigma = 1.30$  and hence

$$Z = \frac{67.47 - 67.39}{\frac{1.30}{\sqrt{400}}} = 1.231$$

Conclusion: Since the calculated value of  $Z = 1.231$  falls in the acceptance region, so we accept our null hypothesis  $H_0: \mu = 67.39$  at 5 % level of significance.

2 A sample of 100 observations from a population with  $\sigma = 2$  inches has  $\bar{X} = 66.5$  inches. Test the null hypothesis  $H_0: \mu = 67$  against the alternative hypothesis  $H_1: \mu \neq 67$ . Use 1% level of significance.

**Solution:**

1. Null hypothesis:  $H_0: \mu = 67$
2. Alternative hypothesis:  $H_1: \mu \neq 67$
3. Level of significance:  $\alpha = 0.01$

4. Test - statistic: 
$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

5. Critical region:  $|Z| > 2.575$  ( $Z < -2.575$  and  $Z > 2.575$ )  
(From the area table of normal distribution, we have  $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$ )

6. Computations: Here,  $n = 100$ ,  $\bar{X} = 66.5$ ,  $\sigma = 2$  and hence

$$Z = \frac{66.5 - 67}{\frac{2}{\sqrt{100}}} = -2.5$$



**Computations:**

Here,  $n = 36$ ,  $\bar{X} = 106$ ,  $\sigma^2 = 225$ ,  $\sigma = 15$ , and hence

$$Z = \frac{106 - 100}{\frac{15}{\sqrt{36}}} = \frac{6}{15}(6) = 2.4$$

**Conclusion:**

Since the calculated value of  $Z = 2.4$  falls in the critical region, thus the null hypothesis is rejected at 5 % level of significance. We may conclude that the IQS of the high school students in this city are higher than 100.

- Q.5 Suppose that scores on an aptitude test used for determining admission to graduate study in statistics are known to be normally distributed with a mean of 500 and a population standard deviation of 100. If a random sample of 64 applicants from a college has a sample mean of 537, is there any evidence that their mean score is different from the mean expected of all applicants? Use  $\alpha = 0.01$ .

**Solution:**

1. Null hypothesis:  $H_0: \mu = 500$

Alternative hypothesis:  $H_1: \mu \neq 500$

2. Level of significance:  $\alpha = 0.01$

3. Test - statistic:  $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

4. Critical region:  $|Z| > 2.575$  ( $Z < -2.575$  and  $Z > 2.575$ )  
(From the area table of normal distribution, we have  $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$ )

5. Computations: Here,  $n = 64$ ,  $\bar{X} = 537$ ,  $\sigma = 100$ , and hence

$$Z = \frac{537 - 500}{\frac{100}{\sqrt{64}}} = \frac{37}{100}(8) = 2.96$$

6. Conclusion:

Since the calculated value of  $Z = 2.96$  falls in the critical region, so we reject our null hypothesis  $H_0: \mu = 500$  at 1 % level of significance. On the basis of the evidence, we may conclude that their mean score is different from the mean expected of all applicants.

- Q.6 Let  $X \sim N(\mu, 100)$  and  $\bar{X}$  be the mean of a random sample of 64 observations of  $X$ , given that  $\bar{X} = 15$ . Test  $H_0: \mu = 12$  against the alternative  $H_1: \mu > 12$ . Use  $\alpha = 0.05$

Q.8. A random sample of 64 drinks from a soft-drink machine has an average content of 21.9 deciliters, with a standard deviation of 1.42 deciliters. Test the hypothesis that  $\mu = 22.2$  deciliters against the alternative hypothesis  $\mu < 22.2$ , at the 5 % level of significance.

**Solution:**

1. Null hypothesis:

$$H_0 : \mu = 22.2$$

Alternative hypothesis:  $H_1 : \mu < 22.2$

2. Level of significance:  $\alpha = 0.05$

3. Test - statistic:

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

4. Critical region:

$$Z < -1.645$$

(From the area table of normal distribution, we have  $-Z_\alpha = -Z_{0.05} = -1.645$ )

5. Computations:

Here,  $n = 64$ ,  $\bar{X} = 21.9$  and  $S = 1.42$ ; we get

$$Z = \frac{21.9 - 22.2}{\frac{1.42}{\sqrt{64}}} = \frac{-0.3}{1.42} (8) = -1.69$$

6. Conclusion:

Since the calculated value of  $Z = -1.69$  falls in the critical region, so we reject our null hypothesis  $H_0 : \mu = 22.2$  at 5 % level of significance.

Q.9 A random sample of 200 trucks were driven on the average 16300 miles a year with a sample standard deviation of 3100 miles. Test the null hypothesis that the average truck mileage in the population is 17000 miles a year against the alternative hypothesis that the average is less. Use the 5 % level of significance.

**Solution:**

1. Null hypothesis:

$$H_0 : \mu = 17000 \text{ miles}$$

Alternative hypothesis:  $H_1 : \mu < 17000 \text{ miles}$

2. Level of significance:  $\alpha = 0.05$

3. Test - statistic:

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

4. Critical region:

$$Z < -1.645$$

(From the area table of normal distribution, we have  $-Z_\alpha = -Z_{0.05} = -1.645$ )

5. Computations:

Here,  $n = 200$ ,  $\bar{X} = 16300$ ,  $S = 3100$ , thus

$$Z = \frac{16300 - 17000}{\frac{3100}{\sqrt{200}}} = -\frac{700}{3100} \sqrt{200} = -3.19$$

Q.12 A random sample of 10 from a population gave  $\bar{X} = 20$  and sum of square of deviations from mean is 144 test  $H_0: \mu = 19.5$  against  $H_1: \mu > 19.5$ . At  $\alpha = 0.05$ .

**Solution:**

1. Null hypothesis:  $H_0: \mu = 19.5$

Alternative hypothesis:  $H_1: \mu > 19.5$

2. Level of significance:  $\alpha = 0.05$

3. Test - statistic:  $t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$

4. Critical region:  $t > 1.833$

(From the t - table, we have  $t_{\alpha(n-1)} = t_{0.05(9)} = 1.833$ )

5. Computations: Here,  $n = 10$ ,  $\bar{X} = 20$ ,  $\Sigma(X - \bar{X})^2 = 144$ ,

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{144}{9} = 16,$$

$$s = 4, \text{ and hence}$$

$$t = \frac{20 - 19.5}{\frac{4}{\sqrt{10}}} = \frac{0.5}{4} \sqrt{10} = 0.395$$

**Conclusion:**

Since the calculated value of  $t = 0.395$  falls in the acceptance region, so we accept our null hypothesis  $H_0: \mu = 19.5$  at 5 % level of significance.

13 Ten cartons are taken at random from an automatic filling machine. The mean net weight of the ten cartons is 15.90 ounces and the sum of squared deviations from this mean is 0.276 (ounces)<sup>2</sup>. Does the sample mean differ significantly from the intended weight of 16 ounces. Use  $\alpha = 0.02$ .



**Solution:**

1. Null hypothesis:  $H_0: \mu = 16$

Alternative hypothesis:  $H_1: \mu \neq 16$

2. Level of significance:  $\alpha = 0.02$

3. Test - statistic:  $t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$

4. Critical region:  $|t| > 2.821$  ( $t < -2.821$  and  $t > 2.821$ )

(From the t-table, we have  $t_{\frac{\alpha}{2}(n-1)} = t_{0.01(9)} = 2.821$ )

5. Computations: Here,  $n = 10$ ,  $\bar{X} = 15.90$ ,  $\sum (X - \bar{X})^2 = 0.276$ ,

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1} = \frac{0.276}{10 - 1} = 0.0307, s = 0.175$$

and hence

$$t = \frac{15.90 - 16}{0.175 / \sqrt{10}} = -\frac{0.10}{0.175} \sqrt{10} = -1.81$$

6. Conclusion:

Since the calculated value of  $t = -1.81$  falls in the acceptance region, so we accept our null hypothesis  $H_0: \mu = 16$  at 2 % level of significance. On the basis of evidence we may conclude that the sample mean does not differ significantly from the intended weight of ounces.

Q.14 In a sample of