## KARPAGAM ACADEMY OF HIGHER EDUCATION

*(Deemed to be University Established Under Section 3 of UGC Act 1956)*

**Coimbatore – 641 021.**

# LECTURE PLAN
# DEPARTMENT OF BIOCHEMISTRY

STAFF NAME : Dr.K.POORNIMA
SUBJECT NAME: Biostatistics and Research Methodology SUB.CODE:18BCP305A
SEMESTER : III CLASS: II M.Sc (BC)

| Sl. No | Duration of Period | Topics to be Covered | Support material |
|---|---|---|---|
| **UNIT I** | | | |
| 1 | 1 | Definitions, Scope of Biostatistics- | T1: 3-9 |
| 2 | 1 | Statistical survey-Organizing, planning and executing the survey | R1: 26-36 |
| 3 | 1 | Sources of data-primary and secondary data; Collection of data – Methods of data collection | T1:10-12<br>T2:14-38 |
| 4 | 1 | Classification of data – Characteristics and types | T1:15-18<br>T2:39-42 |
| 5 | 1 | Tabulation of data – General rules and types of tables | T1:18-23<br>T2:42-63 |
| 6 | 1 | Graphical and diagrammatic representation of data | T1:23-48<br>T2:71-100 |
| 7 | 1 | Continuation of graphical and diagrammatic representation of data | T1:23-48<br>T2:71-100 |
| 8 | 1 | Measures of central tendency – Arithmetic mean, Median and mode | T1:50-58<br>T1:61-72 |
| 9 | 1 | Quartiles, deciles and percentiles | R1: 205-210 |
| 10 | 1 | Measures of dispersion- Range, quartile deviation and mean deviation | T1:85-87<br>R1: 271-282 |
| 11 | 1 | Standard deviation, Coefficient of variation | T1:97-109<br>R1: 282-299 |
| 12 | 1 | Revision and possible questions discussion of unit I | |
| | | **Total no of hours planned for UNIT I = 12** | |
| **Unit II** | | | |
| 1 | 1 | Correlation: Meaning and definition | T1:139-141 |
| 2 | 1 | Scatter diagram | T1:141-144 |
| 3 | 1 | Karl Pearson’s correlation coefficient | T1:146-156 |
| 4 | 1 | Rank correlation | T1:156-159 |

| | | | |
|---|---|---|---|
| 5 | 1 | Correlation problems | T1:172-176 |
| 6 | 1 | Regression: Regression in two variables | T1:177-186 |
| 7 | 1 | Regression coefficient problems | T1:191-200 |
| 8 | 1 | Uses of regression | T1:106 |
| 9 | 1 | Revision and possible questions discussion of Unit II | |
| | | **Total no of hours planned for UNIT II = 09** | |
| **Unit III** | | | |
| 1 | 1 | Probability-Definition, concepts, theorems | R1: 752-765 |
| 2 | 1 | Calculations of probability-simple problems | R1: 774-780 |
| 3 | 1 | Theoretical distributions-Binomial distribution-simple problems | R1: 809-826 |
| 4 | 1 | Theoretical distributions-Poisson distribution-simple problems | R1: 826-836 |
| 5 | 1 | Theoretical distributions-Normal distribution-simple problems | R1: 836-852 |
| 6 | 1 | Practicing problems in theoretical distribution | R1: 858-870 |
| 7 | 1 | Revision and possible question discussion of unit 3 | |
| | | **Total no of hours planned for UNIT III = 07** | |
| **Unit- IV** | | | |
| 1 | 1 | Sampling distribution and test of significance-concepts of sampling, testing of hypothesis, errors in hypothesis testing | R1:882-889 |
| 2 | 1 | Standard errors and sampling distribution | R1:889-892 |
| 3 | 1 | Student's t test | R1: 910-919 |
| 4 | 1 | F test and Chi square test – goodness of fit | R1: 901-910<br>R1: 954-978 |
| 5 | 1 | Analysis of variance – one way classification | R1: 1011-1018 |
| 6 | 1 | Analysis of variance – two way classification | R1: 1018-1038 |
| 7 | 1 | CRD, RBD designs | R1:1040-1049 |
| 8 | 1 | Duncan's multiple range test | W1 |
| 9 | 1 | Revision and possible questions discussion of Unit IV | |
| | | **Total no of hours planned for UNIT IV = 09** | |
| **UNIT V** | | | |
| 1 | 1 | Research: Scope and significance-Types of research | T4:1-7 |
| 2 | 1 | Research process-Characteristics of good research | T4:10-22 |
| 3 | 1 | Problems in research-Identifying research problems | T4:24-30 |
| 4 | 1 | Research designs-Features of good designs | T4:31-32 |
| 5 | 1 | Sources of information: Journals, eJournals, books, biological abstracts | W2 |
| 6 | 1 | Preparation of index cards, review writing | W3 |
| 7 | 1 | Article writing-structure of article, selection of journals for publication-impact factor-citation index and H index | W4 |

| | | | |
|---|---|---|---|
| 8 | 1 | Proposal writing for funding, IPR and patenting, concepts and types | W5 |
| 9 | 1 | Revision and possible questions discussion of Unit V | |
| | | **Total no of hours planned for UNIT V = 09** | |
| **Discussion on previous year end semester examination question paper discussion** | | | |
| 1 | 1 | Discussion on previous year ESE question paper - 1 | |
| 2 | 1 | Discussion on previous year ESE question paper - 2 | |
| **Total number of hours planned to complete this course: 48** | | | |

## Support Materials

**T1:** S.Palanichamy and M. Manoharan (2008). Statistical methods for biologists; 3rd edition; Palani Paramount Publications.

**T2:** P. Ramakrishnan (1995). Biostatistics; 1st edition; Saras Publication.

**T3**: C.R. Kothari (2009). Research Methodology-Methods and Techniques, 3rd edition; New Age International Pvt Ltd, New Delhi.

**R1**: Gupta. S.P. (2007). Statistical Methods; Sultan and Chand Company, New Delhi.
W1: https://www.statisticshowto.datasciencecentral.com/duncans-multiple-range-test/
W2: https://guides.lib.monash.edu/subject-databases/biological-sciences
W3: https://www.enago.com/academy/how-to-write-a-good-scientific-literature-review/
W4:https://btsav.edu.va/sites/default/files/scopes/synep%20%20writing-an-academic journal.article.pdf
W5: https:// www.wipo.int/edocs/pubdocs/en/intproperty/450/wipo-pub-450.pdf

## UNIT-I
## SYLLABUS

Definitions-Scope of Biostatistics- Variables in biology, collection, classification and tabulation of data- Graphical and diagrammatic representation.
Measures of central tendency – Arithmetic mean, median and mode. Measures of dispersion-Range, standard deviation, Coefficient of variation.

### Introduction

Statistical tools are found useful in progressively increasing of disciplines. In ancient times the statistics or the data regarding the human force and wealth available in their land had been collected by the rulers. Now-a-days the fundamental concepts of statistics are considered by many to be essential part of their knowledge.

### Origin and Growth

The origin of the word 'statistics' has been traced to the Latin word 'status', the Italian word 'statista' , the French word 'statistique' and the German word 'statistik'. All these words mean political state.

### Meaning

The word 'statistics' is used in two different meanings. As a plural word it means data or numerical statements. As a singular word it means the science of statistics and statistical methods. The word 'statistics' is also used currently as singular to mean data.

### Definitions

Statistics is " the science of collection, organization, presentation, analysis and interpretation of numerical data". – Dr S.P.Gupta.

"Statistics are numerical statement of facts in any department of enquiry, placed in relation to each other". – Dr.A.L.Bowley.

### Functions

The following are the important functions of statistics.

* Collection
* Numerical Presentation
* Diagrammatic Presentation
* Condensation
* Comparison
* Forecasting
* Policy Making

* Effect Measuring
* Estimation
* Tests of significance.

**Characteristics**

* Statistics is a Quantitative Science.
* It never considers a single item.
* The values should be different.
* Inductive logic is applied.
* Statistical results are true on the average.
* Statistics is liable to be misused.

**Collection of data**

Data constitutes the base. The findings of an investigation depend on correctness and completeness of the relevant data. Sources of data are of two kinds- primary source and secondary source. The term source means origin or place from which data comes or got. A primary source is one that itself collects the data; a secondary source is one that makes available data which were collected by some other agency. Based on source, data are classified under two categories- Primary data and secondary data.

**Primary data**

The data which is collected by actual observation or measurement or count is called primary data.

**Secondary Data**

The data which are compiled from the records of others is called secondary data.

**Methods of collection of primary Data**

Primary Data is collected in any one of the following methods:

* Direct personal interviews
* Indirect oral interviews
* Information from correspondence
* Mailed questionnaire method.
* Schedules sent through enumerators.

**Sources of secondary data**

Secondary data can be compiled either from published sources or from unpublished sources.

**Classification**

Classification is the process of arranging data into groups or classes according to the common characteristics possessed by the individual items.

**Basis**

Data can be classified on the basis of one or more of the following:

**i) Geographical Classification or Spatial Classification**

Some data can be classified area-wise such as states, towns etc.

**ii)Chronological or Temporal or Historical Classification**

Some data can be classified on the basis of time and arranged chronologically or historically.

**iii) Qualitative Classification**

Some data can be classified on the basis of attributes or characteristics.

**iv)Quantitative Classification**

Some data can be classified in terms of magnitudes.

**Tabulation**

Tabulation is the process of arranging data systematically in rows and columns of a table.

There are two methods or modes in which data can be presented. They are

i) Statistical Tables

ii) Diagrams or Graphs

**Parts of a table**

A good table has the following parts or components:

* Identification number
* Title
* Prefatory Note or Head note
* Stubs
* Captions
* Body of the table
* Foot note
* Source

**Frequency Distribution**

The easiest method of organizing data is a frequency distribution, which converts raw data into a meaningful pattern for statistical analysis.

The following are the *steps* of constructing a frequency distribution:
**1.** Specify the number of class intervals. A class is a group (category) of interest. No totally accepted rule tells us how many intervals are to be used. Between 5 and 15 class intervals are generally recommended. Note that the classes must be both *mutually exclusive and all-inclusive.* Mutually exclusive means that classes must be selected such that an item can't fall into two classes, and all-inclusive classes are classes that together contain all the data.

**2.** When all intervals are to be the same width, the following rule may be used to find the required class interval width:
**W = (L - S) / K**
where:
**W=** class width, **L=** the largest data, **S=** the smallest data, **K=** number of classes

**Example**
Suppose the age of a sample of 10 students are:
20.9, 18.1, 18.5, 21.3, 19.4, 25.3, 22.0, 23.1, 23.9, and 22.5
We select K=4 and W=(25.3 - 18.1)/4 = 1.8 which is rounded-up to 2. The frequency table is as follows:

| Class Interval | Class Frequency | Relative Frequency |
|---|---|---|
| 18-20 | 3 | 30 % |
| 20-22 | 2 | 20 % |
| 22-24 | 4 | 40 % |
| 24- 26 | 1 | 10 % |

**Cumulative Frequency Distribution**
When the observations are numerical, cumulative frequency is used. It shows the total number of observations which lie above or below certain key values. Cumulative Frequency for a population = frequency of each class interval + frequencies of preceding intervals. For example, the cumulative frequency for the above problem is: 3, 5, 9, and 10.

**Diagrams and Graphs**
**Diagrams**

Diagrams are various geometrical shapes such as bars, circles etc . Diagrams are based on scales but are not confirmed to points or lines. They are more attractive and easier to understand than graphs and are widely used in advertisement and publicity.

**Rules for construction**

* Title
* Proportion between width and height
* Size
* scale
* Index
* Suitable Diagram
* Simplicity
* Neatness
* Foot-Note and source
* Identification numbers.

**Types of Diagram**

The frequently used diagrams are divided into the following four heads:

1. One Dimensional diagram- Bar Diagram
2. Two Dimensional diagram – Pie Diagram, Rectangle, squares and circles
3. Three Dimensional diagram – Cubes
4. Pictograms and Cartograms.

***Histograms*** are used to graph absolute, relative, and cumulative frequencies.

***Ogive*** is also used to graph cumulative frequency. An ogive is constructed by placing a point corresponding to the *upper end of each class* at a height equal to the cumulative frequency of the class. These points then are connected. An ogive also shows the relative cumulative frequency distribution on the right side axis.

***A less-than ogive*** shows how many items in the distribution have a value less than the upper limit of each class.

***A more-than ogive*** shows how many items in the distribution have a value greater than or equal to the lower limit of each class.

***A less-than cumulative frequency polygon*** is constructed by using the upper true limits and the cumulative frequencies.

***A more-than cumulative frequency polygon*** is constructed by using the lower true limits and the cumulative frequencies.

***Pie chart*** is often used in newspapers and magazines to depict budgets and other economic information. A complete circle (the pie) represents the total number of measurements. The size of a slice is proportional to the relative frequency of a particular category. For example, since a complete circle is equal to 360 degrees, if the relative frequency for a category is 0.40, the slice assigned to that category is 40% of 360 or (0.40)(360)= 144 degrees.

## Measures of Central Tendency and Dispersion

### INTRODUCTION

In this chapter we are going to deal with Measures of central tendency and about the measures of dispersion. The measures of central tendency concentrate about the values in the central part of the distribution. Plainly speaking an average of a statistical series is the value of the variable which is the representative of the entire distribution. If we know the average alone we cannot form a complete idea about the distribution so for the completeness of the idea we use Measures of dispersion.

### Measures of Central Tendency

According to Professor Bowley the measures of central tendency are “statistical constants which enable us to comprehend in a single effort the significance of the whole “

The following are the three measures of central tendency in this chapter we deal with

- Arithmetic Mean or simply Mean
- Median
- Mode

### Arithmetic Mean or simply Mean

Arithmetic Mean or simply Mean is the total values of the item divided by their number of the items. It is usually denoted by $\bar{X}$.

### Individual series

$$\bar{X} = \Sigma X / N$$

Example:

The expenditure of ten families are given below .Calculate arithmetic mean.

30 70 10 75 500 8 42 250 40 36

### Solution

Here N=10

Σ X = 30 +70 +10 +75 +500 +8 +42 +250 +40 +36 = 1061

$\bar{X}$ = 1061 / 10 =106.1

**Discrete series**

$\bar{X}$ = Σ f X / Σ f

**Example**

Calculate the mean number of person per house.

| | | | | | |
|---|---|---|---|---|---|
| No. of person : | 2 | 3 | 4 | 5 | 6 |
| No. of house : | 10 | 25 | 30 | 25 | 10 |

**Solution**

| X | f | f X |
|---|---|---|
| 2 | 10 | 20 |
| 3 | 25 | 75 |
| 4 | 30 | 120 |
| 5 | 25 | 125 |
| 6 | 10 60 | |

Σ f =100 Σ f X= 400

$\bar{X}$ =400 / 100 = 4 .

**Continuous series**

$\bar{X}$ = Σ f m / Σ f where m represents the mid value .

Mid-value = (upper boundary + lower boundary) / 2.

**Example**

Calculate the mean for the following.

| | | | | | | |
|---|---|---|---|---|---|---|
| Marks : | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
| No. of student : | 5 | 8 | 12 | 15 | 6 | 4 |

Solution:

| **C.I** | **f** | **m** | **f m** |
|---|---|---|---|
| 20-30 | 5 | 25 | 125 |
| 30-40 | 8 | 35 | 280 |
| 40-50 | 12 | 45 | 540 |

| | | | |
|---|---|---|---|
| 50-60 | 15 | 55 | 825 |
| 60-70 | 6 | 65 | 390 |
| 70-80 | 4 | 75 | 300 |

Σ f = 50 Σ f m= 2460

$\bar{X}$ = 2460 / 50 = 49.2.

**Median**

The median is the value for the middle most items when all the items are in the order of magnitude. It is denoted by M or Me.

**Individual series**

For odd number of item

Position of the median = (N+1) / 2

For even number of item

Position of the median = [ (N / 2)+((N/2)+1)] / 2

**Example**

Calculate median for the following.

22 10 6 7 12 8 5

**Solution**

Here N =7

Arrange in ascending order or descending order.

5 6 7 8 10 12 22

(N+1) / 2= (7+1) /2

= 4 th item = 8

**Discrete series**

Position of the median = (N+1) / 2 th item.

**Example**

Find the median for the following.

| | | | | | |
|---|---|---|---|---|---|
| X : | 10 | 15 | 17 | 18 | 21 |
| F: | 4 | 16 | 12 | 5 | 3 |

**Solution**

| X | f | c.f |
|---|---|---|

| | | |
|---|---|---|
| 10 | 4 | 4 |
| 15 | 16 | 20 |
| 17 | 12 | 32 |
| 18 | 5 | 37 |
| 21 | 3 | 40 |
| | N= 40 | |

$(N+1)/2 = (40+1)/2 = 20.5^{th}$ item

$= (20^{th}\text{ item} + 21^{st}\text{ item})/2 = (15+17)/2$

$= 16.$

**Continuous series**

$$M = L+\left[\frac{((N/2) - c.f) \times i}{f}\right]$$

Where L- lower boundary, f-frequency, i-size of class interval, c.f- cumulative frequency.

**Example**

Calculate the median height given below.

| Height : | 145-150 | 150-155 | 155-160 | 160-165 | 165-170 | 170-175 |
|---|---|---|---|---|---|---|
| No. of student: | 2 | 5 | 10 | 8 | 4 | 1 |

**Solution :**

| Height | No. of student | c.f |
|---|---|---|
| 145-150 | 2 | 2 |
| 150-155 | 5 | 7 |
| 155-160 | 10 | 17 |
| 160-165 | 8 | 25 |
| 165-170 | 4 | 29 |
| 170-175 | 1 | 30 |

$\Sigma f = 30$

Position of the median = $N/2^{th}$ item = 30 / 2 =15.

$$M = L+\left[\frac{((N/2) - c.f) \times i}{f}\right]$$

$$= 155+\frac{[(15-7)\times 5]}{10} = 155+(40/10) = 159.$$

**Mode :**

Mode is the value which has the greatest frequency density. Mode is usually denoted by Z.

**Individual series**

The value which occur more times are identified as mode.

**Example**

Determine the mode

32, 35,42, 32, 42,32.

**Solution:**

Unimode = 32.

**Discrete series**

Determine the mode

| **Size of dress** | **No. of set** |
|---|---|
| 18 | 55 |
| 20 | 120 |
| 22 | 108 |
| 24 | 45 |

here mode represents highest frequency .

Mode = 20

**Continuous series**

$Z = L +[ i( f_1\text{-}f_0) /(2f_1 - f_0 - f_2)]$

Where L- lower boundary , $f_1$-frequency of the modal class, $f_0$ – frequency of the preceding modal class, $f_2$- frequency of the succeeding modal class, i-size of class interval , c.f- cumulative frequency.

**Example**

Determine the mode

| Marks : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 |
|---|---|---|---|---|---|
| No.of student : | 5 | 20 | 35 | 20 | 12 |

**Solution**

| **Marks** | **No. of student** |
|---|---|
| 0-10 | 5 |
| 10-20 | 20 |
| 20-30 | 35 |
| 30-40 | 20 |
| 40-50 | 12 |

$Z = L + [\ i(\ f_1\text{-}f_0)\ /(2f_1 – f_0 – f_2)]$

$= 20+[10(35-20)/(2(35)-20-20)] = 20+5$

$= 25.$

**Empirical relation**

- Mode = 3 median -2 mean.

## Measures of Dispersion

Measure of dispersion deals mainly with the following three measures

- Range
- Standard deviation
- Coefficient of variation

**Range**

Range is the difference between the greatest and the smallest value.

- Range = L – S , where L-largest value & S-Smallest value
- Coefficient of range = ( L-S) /(L+S)

**Individual series**

**Example**

Find the value of range and its coefficient of range for the following data.

8 ,10, 5, 9,12,11

**Solution**

Range = L – S

= 12- 5 =7

Coefficient of range = ( L-S) /(L+S)

= (12-5) / (12+5)

= 7 /17 = 0.4118

**Continuous series**

Range = L – S, where L-Mid-value of largest boundary & S-Mid-value of smallest boundary

Calculate the range.

| Marks | : | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|
| No.of student | : | 5 | 8 | 12 | 15 | 6 | 4 |

**Solution**

| **C.I** | **f** | **m** |
|---|---|---|
| 20-30 | 5 | 25 |
| 30-40 | 8 | 35 |
| 40-50 | 12 | 45 |
| 50-60 | 15 | 55 |
| 60-70 | 6 | 65 |
| 70-80 | 4 | 75 |

Here L=75 & S=25

Range = L – S = 75-25 = 50

**Standard deviation**

The standard deviation is the root mean square deviation of the values from the arithmetic mean .It is a positive square root of variants. It is also called root mean square deviation. This is usually denoted by σ .

**Individual series**

$$\sigma = \sqrt{(\Sigma x^2 / N) - (\Sigma x / N)^2}$$

**Example**

Calculate standard deviation for the following data.

40,41,45,49,50,51,55,59,60,60.

**Solution**

| X | $X^2$ |
|---|---|
| 40 | 1600 |
| 41 | 1681 |
| 45 | 2025 |
| 49 | 2401 |
| 50 | 2500 |
| 51 | 2601 |
| 55 | 3025 |

| | |
|---|---|
| 59 | 3481 |
| 60 | 3600 |
| 603600 | |
| 510 | $\Sigma x^2 = 26504$ |

$$\sigma = \sqrt{(\Sigma x^2 / N) - (\Sigma x / N)^2}$$

$$= \sqrt{(26514/10) - (510/10)^2}$$

$= 7.09$

**Discrete series**

$$\sigma = \sqrt{(\Sigma fx^2 / \Sigma f) - (\Sigma fx / \Sigma f)^2}$$

**Example**

Calculate standard deviation for the following data.

| | | | | | | |
|---|---|---|---|---|---|---|
| X : | 0 | 1 | 2 | 3 | 4 | 5 |
| F : | 1 | 2 | 4 | 3 | 0 | 2 |

**Solution**

| X | f | fx | $x^2$ | $fx^2$ |
|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 |
| 1 | 2 | 2 | 1 | 2 |
| 2 | 4 | 8 | 4 | 16 |
| 3 | 3 | 9 | 9 | 27 |
| 4 | 0 | 0 | 16 | 0 |
| 5 | 2 10 | 25 | 50 | |

$\Sigma f = 12$ $\Sigma fx = 29$ $\Sigma fx^2 = 95$

$$\sigma = \sqrt{(\Sigma fx^2 / \Sigma f) - (\Sigma fx / \Sigma f)^2}$$

$$= \sqrt{(95/12) - (29/12)^2}$$

$= 1.44$

**Continuous series**

$$\sigma = \sqrt{(\Sigma fm^2 / \Sigma f) - (\Sigma fm / \Sigma f)^2}$$

**Example**

| C.I | : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 |
|---|---|---|---|---|---|---|
| F | : | 2 | 5 | 9 | 3 | 1 |

**Solution**

| **C.I** | **f** | **m** | **fm** | **$m^2$** | **$fm^2$** |
|---|---|---|---|---|---|
| 0-10 | 2 | 5 | 10 | 25 | 50 |
| 10-20 | 5 | 15 | 75 | 225 | 1125 |
| 20-30 | 9 | 25 | 225 | 625 | 5625 |
| 30-40 | 3 | 35 | 105 | 1225 | 3675 |
| 40-50 | 1 | 45 | 45 | 2025 | 2025 |
| | 20 | | 460 | | 12500 |

$$\sigma = \sqrt{(\Sigma fm^2 / \Sigma f) - (\Sigma fm / \Sigma f)^2}$$

$$= \sqrt{(12500/20) - (460/20)^2}$$

$$= 9.79$$

**Coefficient of variation**

Coefficient of variation = [standard deviation / arithmetic mean ] x100

**Example**

Calculate the coefficient of variation.

Mean= 51, standard deviation = 7.09

**Solution**

Coefficient of variation = [standard deviation / arithmetic mean ] x100
= (7.09 /51) x100
= 13.9

## POSSIBLE QUESTIONS
## UNIT I

### PART A (20 x 1 = 20 Marks)
### Question number 1 – 20 online examinations

### PART B (5 x 2= 20Marks)

1. Define Statistics.
2. Write the formula to calculate the angle in Pie Diagram.
3. Define Classification.
4. Define Mid-Value and also find the Mid-Value of 100 – 110.
5. Define Frequency Distribution.
6. What do you mean by size of the Class Interval?
7. Write a formula to calculate Percentage Bar Diagram.
8. Define Histogram.
9. What are the types of classification?
10. Write about Geographical Classification with example.
11. Define Frequency Distribution.
12. Write any two functions of Statistics.
13. What is Bi-variate data?
14. Define Frequency Curve.
15. Calculate the Mean for the following.

| | | | | | |
|---|---|---|---|---|---|
| X | 20 | 30 | 35 | 15 | 10 |
| f | 2 | 3 | 4 | 3 | 2 |

16. Define Median and give Example.
17. Calculate the Range and its Coefficient for the following data.

| | | | | | | |
|---|---|---|---|---|---|---|
| X | : | 12 | 14 | 16 | 18 | 20 |
| f | : | 1 | 3 | 5 | 3 | 1 |

18. What is mean by Bimodal?
19. Calculate the Median for the following data.

    80 100 50 90 120 110

20. Write the relation between Standard Deviation and Variance.
21. Calculate the Average number of students per class for the following data.

    26 46 33 25 36 27 34 29

22. Find Median and Mode for the following data.

    13 16 17 15 18 14 19 15 12

23. Define Range.
24. Find the Arithmetic Mean for the following data.
 70 60 75 50 42 95 46
25. Calculate the Range and its Coefficient for the following data.
 17 10 56 19 12 11 18 14
26. Find the median for 57, 58, 61, 42, 38, 65, 72, and 66
27. Write the empirical relation for Mode.

**PART C (5 X 6 = 30 Marks)**

1. Explain about the Classification of data.
2. Draw a suitable Pie Diagram to represent the following submitted as a part of the budget proposal of the govt. of India for the year 1995 – 96.

| **Item of Expenditure** | **Percentage** |
|---|---|
| i) Interest | 25 |
| ii) Defense | 15 |
| iii) Other non plan expenditure | 20 |
| iv) States share of taxes and duties | 15 |
| v) State and UT plan assistance | 10 |
| vi) Central plan | 15 |
| Total | 100 |

3. You are given the average expenditure of a family for a month.

| **Item** | **Average Expenditure per month (Rs)** |
|---|---|
| Food | 2,400 |
| Clothing | 200 |
| Rent | 800 |
| Medicare | 150 |
| Entertainment | 450 |

Draw a Pie Diagram for the above data.

4. The following table shows the total sale in July 2005 of major brands of cars in India. Represent the following data by using Simple Bar Diagram.

| **Brand** | **July 2005**<br>**Sales (in Rs. '000)** |
|---|---|
| Maruti | 81 |
| Hyundai Santro | 39 |
| Honda city | 14 |
| Matiz | 20 |

| | |
|---|---|
| Opel Astra | 10 |
| Fait Uno | 12 |
| Ford | 37 |
| Mitsubishi Lancer | 11 |
| Mercedes | 3 |

5. Draw the less than Ogive and hence find the Median of the following data.

| **Marks** | 20 - 29 | 30 - 39 | 40 - 49 | 50 - 59 | 60 - 69 | 70 - 79 | 80 - 89 | 90 - 99 |
|---|---|---|---|---|---|---|---|---|
| **No. of students** | 7 | 11 | 24 | 32 | 9 | 14 | 2 | 1 |

6. Draw Percentage Bar Diagram for the following data.

| | |
|---|---|
| Food | Rs.200 |
| Clothing | Rs.60 |
| Education | Rs.70 |
| Rent | Rs.130 |
| Miscellaneous | Rs.40 |

7. Draw a Multiple Bar Diagram for the following data.

| **Year** | **Sales (000 Rs)** | **Gross Profit (000 Rs.)** | **Net Profits (000 Rs)** |
|---|---|---|---|
| 1974 | 100 | 30 | 10 |
| 1975 | 120 | 40 | 15 |
| 1976 | 130 | 45 | 25 |
| 1977 | 150 | 50 | 25 |

8. Nixon Corporation manufactures computers. The following data are the numbers of computers produced at the company for sample of 25 days.

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

Construct frequency distribution using classes 21 - 23, 24 - 26, 27 - 29, 30 - 32 and 33 - 35. And draw a Histogram to the frequency distribution.

9. The frequency distribution representing the number of days annually the employees at the Voltas Ltd. who were absent due to illness is

| **Number of days absent** | 0-2 | 3-5 | 6-8 | 9-11 | 12-14 | Total |
|---|---|---|---|---|---|---|
| **Frequency** | 5 | 12 | 20 | 10 | 3 | 50 |

Draw a Frequency Polygon to the above Frequency Distribution.

10. Calculate the Mode for the following Continuous Frequency Distribution.

| **Salary (in Rs. 1000s) :** | 0 – 19 | 20 - 39 | 40 - 59 | 60 – 79 | 80 - 99 |
|---|---|---|---|---|---|
| **No. of Employees:** | 5 | 20 | 35 | 20 | 12 |

11. Find the Mean and the Standard Deviation for the given below data set.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 14 | 20 | 12 | 21 | 16 | 19 | 17 | 14 | 25 |

12. Calculate the Standard Deviation and Coefficient of Variance (CV) for the following data.

| X | 0 – 10 | 10 - 20 | 20 - 30 | 30 – 40 | 40 - 50 |
|---|---|---|---|---|---|
| f | 2 | 5 | 9 | 3 | 1 |

13. Calculate the Median for the following Continuous Frequency Distribution.

| **Wages (in Rs.) :** | 0 - 19 | 20 - 39 | 40 - 59 | 60 – 79 | 80 - 99 |
|---|---|---|---|---|---|
| **No. of Workers:** | 5 | 20 | 35 | 20 | 12 |

14. Calculate the Coefficient of Variation for the following data.

| **X** | 6 | 9 | 12 | 15 | 18 |
|---|---|---|---|---|---|
| **f** | 7 | 12 | 13 | 10 | 8 |

15. Calculate the Median for the following.

| **Hourly Wages (in Rs.)** | 40 - 50 | 50 – 60 | 60 - 70 | 70 - 80 | 80 - 90 | 90 - 100 |
|---|---|---|---|---|---|---|
| **Number of Employees** | 10 | 20 | 15 | 30 | 15 | 10 |

16. The following data give the details about salaries (in thousands of rupees) of seven employees randomly selected from a Pharmaceutical Company.

| **Serial No.** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **Salary per Annum ( ‘000)** | 89 | 57 | 104 | 73 | 26 | 121 | 81 |

Calculate the Standard Deviation and Coefficient of variance of the given data.

17. Calculate the Arithmetic Mean for the following data.

| **Height** (cms): | 160 | 161 | 162 | 163 | 164 | 165 | 166 |
|---|---|---|---|---|---|---|---|
| **No. of Persons** : | 27 | 36 | 43 | 78 | 65 | 48 | 28 |

18. Calculate the Coefficient of Variance for the following data.

77 73 75 70 72 76 75 72 74 76

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics and Research Methodology
UNIT I

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | The most suitable form of presentation for publicity and Propaganda is | **diagram** | numerals | graph | map | **diagram** |
| 2 | 1 | Mailed enquiry method can be adopted if the respondents are | illiterate | **literate** | blind | physically challenged | **literate** |
| 3 | 1 | Pie-chart represents the components of a factor by | percentages | angles | **sectors** | deciles | **sectors** |
| 4 | 1 | The number of questions of a questionnaire should be | 5 | 20 | 25 | **as small as possible keeping in view the purpose of the survey** | **as small as possible keeping in view the purpose of the survey** |
| 5 | 1 | Primary data are | **always more reliable compared to secondary data** | less reliable compared to secondary data | depends on the care with which data have been selected | depends on the agency collecting the data | **always more reliable compared to secondary data** |
| 6 | 1 | The census data published for state wise population in India will be known as | Quantitative classification | Two-way classification | **Geographical classification** | Quanlitative classification | **Geographical classification** |
| 7 | 1 | Statistics implies | both data and science | data only | **data, science and measure in samples** | samples | **data, science and measure in samples** |
| 8 | 1 | Data taken from the publication, 'Agricultural Situation in India' will be considered as | primary data | **secondary data** | primary and secondary data | published data | **secondary data** |
| 9 | 1 | Year wise recording of data of food production will be called as: | Geographical classification | **Chronological classification** | Quantitative classification | Quanlitative classification | **Chronological classification** |
| 10 | 1 | Who is the father of Biostatistics? | R.A.Fisher | W.Gosset | **Sir Francis Galton** | S.C.Gupta | **Sir Francis Galton** |
| 11 | 1 | Number of source of data is | **2** | 3 | 4 | 1 | **2** |
| 12 | 1 | Compared with primary data secondary data are | more reliable | **less reliable** | equally reliable | not reliable | **less reliable** |
| 13 | 1 | In quantitative classification data are classified on the basis of | attributes | time | location | **magnitudes** | **magnitudes** |
| 14 | 1 | Classification according to class-intervals would yield | raw data | discrete data | qualitative data | **grouped data** | **grouped data** |
| 15 | 1 | In qualitative classification data are classified on the basis of | **attributes** | time | location | class intervals | **attributes** |
| 16 | 1 | In geographical classification data are classified on the basis of | **area** | attributes | time | location | **area** |
| 17 | 1 | Which source is one that itself collects the data? | **primary data** | secondary data | published data | non published data | **primary data** |
| 18 | 1 | Individual observations are called | **raw data** | grouped data | ungrouped data | simple data | **raw data** |
| 19 | 1 | Which one is Geographical classification? | 1990-91 | **North** | Male | 442 | **North** |
| 20 | 1 | In discrete frequency distribution values are given as | class intervals | grouped data | **ungrouped data** | equal frequency data | **ungrouped data** |
| 21 | 1 | In continuous frequency distribution values are given as | **class intervals** | grouped data | ungrouped data | equal frequency data | **class intervals** |
| 22 | 1 | Which of the following is the one dimensional diagram? | **square diagram** | multiple bar diagram | rectangular diagram | Pie-Chart | **square diagram** |
| 23 | 1 | In a bar diagram, the base line is | Horizontal | **Vertical** | False baseline | true baseline | **Vertical** |
| 24 | 1 | Pictograms are shown by | Dots | Lines | Circles | **Pictures** | **Pictures** |
| 25 | 1 | Histogram is suitable for the data presented as | **continuous grouped frequency distribution** | discrete grouped frequency distribution | individual series | discrete series | **continuous grouped frequency distribution** |
| 26 | 1 | Numerical data presented in descriptive form are called | classified presentation | tabular presentation | graphical presentation | **textual presentation** | **textual presentation** |
| 27 | 1 | A simple table represents | **only one factor or variable** | always two factors or variables | two or more number of factors or variables | only four factors or variables | **only one factor or variable** |
| 28 | 1 | The row headings of a table are known as | sub-titles | **stubs** | reference notes | captions | **stubs** |
| 29 | 1 | The origin of the word statistics has been traced from the Latin word | statista | **status** | statistic | statistique | **status** |
| 30 | 1 | Relative error is always | Positive | Negative | **Positive or Negative** | Zero | **Positive or Negative** |
| 31 | 1 | The column headings of a table are known as | sub-titles | stubs | reference notes | **captions** | **captions** |
| 32 | 1 | Statistical data are collected for, | Collecting data without any purpose | **a given purpose** | any purpose | for all purpose | **a given purpose** |
| 33 | 1 | In grouped data, the number of classes preferred are | Minimum possible | **adequate** | Maximum possible | any arbitrarily chosen number | **adequate** |
| 34 | 1 | Class interval is measured as | The sum of the upper and lower limit | Half of the sum of the upper and lower limit | **The difference between the upper and lower limit** | Upper limit + lower limit | **The difference between the upper and lower limit** |
| 35 | 1 | The shape of the trilinear charts is that of a | Cone | Cube | **Equilateral triangle** | Pyramid | **Equilateral triangle** |
| 36 | 1 | Mean is a measure of | **central value** | dispersion | correlation | regression | **central value** |
| 37 | 1 | Mode is that value in a frequency distribution which possesses | minimum frequency | frequency one | **maximum frequency** | last frequency | **maximum frequency** |
| 38 | 1 | The most stable measure of central tendency is | **the mean** | the median | the mode | range | **the mean** |
| 39 | 1 | Sum of the deviations about mean is: | minimum | **zero** | maximum | one | **zero** |
| 40 | 1 | The formula used to calculate arithmetic mean for individual series by direct method is | **∑X/N** | ∑Fx/N | ∑Fm/N | N/ ∑Fx | **∑X/N** |
| 41 | 1 | The formula used to calculate arithmetic mean for individual series by short cut method is | **A+ ∑d/N** | A+ ∑Fd/N | A+∑Fm/N | A+ N/ ∑d | **A+ ∑d/N** |
| 42 | 1 | In continuous series the formula for A.M is | X/N | N/∑X | ∑X/N | **∑fm/N** | **∑fm/N** |
| 43 | 1 | A.M = 8, N = 12 then ∑X = | 76 | 80 | 86 | **96** | **96** |
| 44 | 1 | 12, 34, 56, 34, 45, 11 in this series the mode is | 12 | 56 | **34** | 11 | **34** |
| 45 | 1 | The data given as 5, 12, 16, 24, 35, 44 will be called as | a continuous series | a discrete series | **an individual series** | time series | **an individual series** |
| 46 | 1 | Which of the following divides the series into two equal parts | Mean | Mode | **Median** | Range | **Median** |
| 47 | 1 | Mode of the following data 3, 6, 5, 7, 8, 4, 9 | 3 | 7 | **no mode** | 5 | **no mode** |
| 48 | 1 | Which of the following is not a measure of dispersion? | Range | Quartile deviation | Standard deviation | **Median** | **Median** |
| 49 | 1 | Range of the given values is given by | **L-S** | L+S | S-L | LS | **L-S** |
| 50 | 1 | Which one of the following is relative measure of dispersion? | range | Q.D | S.D | **Coefficient of variation** | **Coefficient of variation** |
| 51 | 1 | Range of a set of values is 65 and maximum value in the series is 83. The minimum value of the series is | 74 | 9 | **18** | 65 | **18** |
| 52 | 1 | If standard deviation is 5, then the variance is | 5 | **25** | 625 | 2.23068 | **25** |
| 53 | 1 | If the value of mode and mean is 60 and 66 then the value of median is | **64** | 46 | 54 | 44 | **64** |
| 54 | 1 | Standard deviation is also called | **Root mean square deviation** | Mean square deviation | Root deviation | Root median square deviation | **Root mean square deviation** |
| 55<br>56 | 1<br>1 | X: 10-15 15-20 20-25 25-30 30-35 35-40<br>f: 12 4 10 6 11 8 , range is | 8 | **30** | 33 | 12 | **30** |
| 57 | 1 | If the S.D and the C.V of a series are 5 and 25, then the mean | **500** | 200 | 250 | 100 | **500** |
| 58 | 1 | Which one of the following is a measure of central tendency | **median** | range | variation | correlation | **median** |
| 59 | 1 | Mean of the following values is 5 15 20 10 20 | 5 | **14** | 41 | 20 | **14** |
| 60 | 1 | The position of the median for an individual series is taken as | **(n+1)/2** | (n+2)/2 | n/2 | n/4 | **(n+1)/2** |

## UNIT-II
## SYLLABUS

Correlation: Meaning and definition - Scatter diagram –Karl Pearson's correlation coefficient. Rank correlation.
Regression: Regression in two variables – Regression coefficient problems – uses of regression.

### Simple Linear Correlation

The term Correlation refers to the relationship between the variables. Simple correlation refers to the relationship between two variables. Various types of correlation are considered.

**Positive or Negative** when the values of two variables change in the same direction, their positive correlation between the two variables.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Example :** | X | 50 | 60 | 70 | 95 | 100 | 105 |
| | Y | 23 | 32 | 37 | 41 | 46 | 50 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Example :** | X | 34 | 25 | 18 | 10 | 7 |
| | Y | 51 | 49 | 42 | 33 | 19 |

### Simple or Partial or Multiple

When only two variables are considered as under positive or negative correlation above the correlation between them is called Simple correlation. When more than two variables as considered the correlation between two of them when all other variables are held constant, i.e., when the linear effects of all other variables on them are removed is called partial correlation. When more than two variables are considered the correlation between one of them and its estimate based on the group consisting of the other variables is called multiple correlation.

### Methods

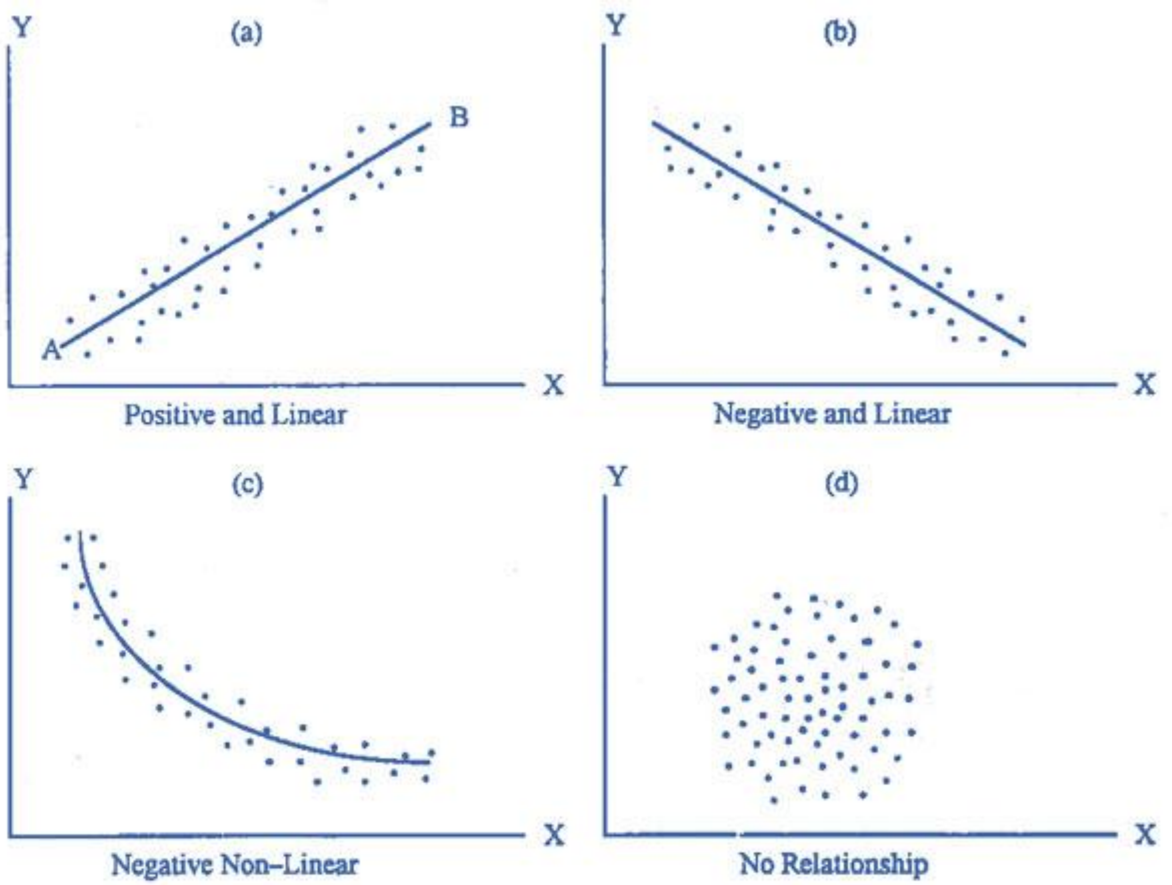
The following four methods are available under simple linear correlation and among them; product moment method is the best one.

- Scatter Diagram
- Karl Pearson's correlation coefficient or product moment correlation coefficient (r)
- Spearman's rank correlation coefficient ( ρ )

- Correlation coefficient by concurrent deviation method ( $r_c$ ).

**Scatter Diagram**

Scatter diagram is a graphic picture of the sample data. Suppose a random sample of n pairs of observations has the values $(X_1,Y_1),(X_2,Y_2),(X_3,Y_3),\ldots,(X_n,Y_n)$. These points are plotted on a rectangular co-ordinate system taking independent variable on *X*-axis and the dependent variable on *Y*-axis. Whatever be the name of the independent variable, it is to be taken on *X*-axis. Suppose the plotted points are as shown in figure (a). Such a diagram is called scatter diagram. In this figure, we see that when *X* has a small value, *Y* is also small and when *X* takes a large value, *Y* also takes a large value. This is called direct or positive relationship between *X* and *Y*. The plotted points cluster around a straight line. It appears that if a straight line is drawn passing through the points, the line will be a good approximation for representing the original data. Suppose we draw a line ***AB*** to represent the scattered points. The line ***AB*** rises from left to the right and has positive slope. This line can be used to establish an approximate relation between the random variable *Y* and the independent variable *X*. It is nonmathematical method in the sense that different persons may draw different lines. This line is called the regression line obtained by inspection or judgment.



Making a scatter diagram and drawing a line or curve is the primary investigation to assess the type of relationship between the variables. The knowledge gained from the scatter diagram can be used for further analysis of the data. In most of the cases the diagrams are not as simple as in figure (a). There are quite complicated diagrams and it is difficult to choose a proper mathematical model for representing the original data. The scatter diagram gives an indication of the appropriate model which should be used for further analysis with the help of method of least squares. Figure (b) shows that the points in the scatter diagram are falling from the top left corner to the right. This is a relation called inverse or indirect. The points are in the neighborhood of a certain line called the regression line.

As long as the scattered points show closeness to a straight line of some direction, we draw a straight line to represent the sample data. But when the points do not lie around a straight line, we do not draw the regression line. Figure (c) shows that the plotted points have a tendency to fall from left to right in the form of a curve. This is a relation called non-linear or curvilinear. Figure (d) shows the points which apparently do not follow any pattern. If *X* takes a small value, *Y* may take a small or large value. There seems to be no sympathy between *X* and *Y*. Such a diagram suggests that there is no relationship between the two variables.

**Karl Pearson's Coefficient**

Karl Pearson's Product-Moment Correlation Coefficient or simply Pearson's Correlation Coefficient for short, is one of the important methods used in Statistics to measure Correlation between two variables.

A few words about Karl Pearson. Karl Pearson was a British mathematician, statistician, lawyer and a eugenicist. He established the discipline of mathematical statistics. He founded the world's first statistics department In the University of London in the year 1911. He along with his colleagues Weldon and Galton founded the journal "Biometrika" whose object was the development of statistical theory.

The Correlation between two variables X and Y, which are measured using Pearson's Coefficient, give the values between +1 and -1. When measured in population the Pearson's Coefficient is designated the value of Greek letter rho (ρ). But, when studying a sample, it is designated the letter r. It is therefore sometimes called Pearson's r. Pearson's coefficient reflects the linear relationship between two variables. As mentioned above if the correlation coefficient is +1 then there is a perfect positive linear relationship between variables, and if it is -1 then there is a perfect negative linear relationship between the variables. And 0 denotes that there is no relationship between the two variables.

The degrees -1, +1 and 0 are theoretical results and are not generally found in normal circumstances. That means the results cannot be more than -1, +1. These are the upper and the lower limits.

Pearson's Coefficient computational formula

$$r=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{\left(\sum X^{2}-\frac{(\sum X)^{2}}{N}\right)\left(\sum Y^{2}-\frac{(\sum Y)^{2}}{N}\right)}}$$

Sample question: compute the value of the correlation coefficient from the following table:

| Subject | Age x | Weight Level y |
|---|---|---|
| 1 | 43 | 99 |
| 2 | 21 | 65 |
| 3 | 25 | 79 |
| 4 | 42 | 75 |
| 5 | 57 | 87 |
| 6 | 59 | 81 |

**Step 1:** Make a chart. Use the given data, and add three more columns: xy, x2, and y2.

.

| Subject | Age x | Weight Level y | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| 1 | 43 | 99 | | | |
| 2 | 21 | 65 | | | |
| 3 | 25 | 79 | | | |
| 4 | 42 | 75 | | | |
| 5 | 57 | 87 | | | |
| 6 | 59 | 81 | | | |

**Step 2:** Multiply x and y together to fill the xy column. For example, row 1 would be 43 × 99 = 4,257

**Step 3:** Take the square of the numbers in the x column, and put the result in the $x^2$ column.

| Subject | Age x | Weight Level y | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| 1 | 43 | 99 | 4257 | 1849 | |
| 2 | 21 | 65 | 1365 | 441 | |
| 3 | 25 | 79 | 1975 | 625 | |
| 4 | 42 | 75 | 3150 | 1764 | |
| 5 | 57 | 87 | 4959 | 3249 | |
| 6 | 59 | 81 | 4779 | 3481 | |

**Step 4:** Take the square of the numbers in the y column, and put the result in the $y^2$ column.

**Step 5:** Add up all of the numbers in the columns and put the result at the bottom.2 column. The Greek letter sigma (Σ) is a short way of saying "sum of."

| Subject | Age x | Weight Level y | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| 1 | 43 | 99 | 4257 | 1849 | 9801 |
| 2 | 21 | 65 | 1365 | 441 | 4225 |
| 3 | 25 | 79 | 1975 | 625 | 6241 |
| 4 | 42 | 75 | 3150 | 1764 | 5625 |
| 5 | 57 | 87 | 4959 | 3249 | 7569 |
| 6 | 59 | 81 | 4779 | 3481 | 6561 |
| Σ | 247 | 486 | 20485 | 11409 | 40022 |

**Step 6:** Use the following formula to work out the correlation coefficient. The answer is: 1.3787 × 10-4 the range of the correlation coefficient is from -1 to 1. Since our result is 1.3787 × 10-4, a tiny

positive amount, we can’t draw any conclusions one way or another.

**Spearman’s Rank Correlation Coefficient**

The Spearman correlation coefficient is often thought of as being the Pearson correlation coefficient between the ranked variables. In practice, however, a simpler procedure is normally used to calculate ρ. The $n$ raw scores $X_i$, $Y_i$ are converted to ranks $x_i$, $y_i$, and the differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated.

**If there are no tied ranks, then ρ is given by**

$$\rho = 1 - \left(\frac{6\sum d^2}{N(N^2 -1)}\right)$$

If tied ranks exist, Pearson's correlation coefficients between ranks should be used for the calculation:

One has to assign the same rank to each of the equal values. It is an average of their positions in the ascending order of the values.

**Example**

X : 21 36 42 37 25

Y : 47 40 37 42 43. For the data given above , calculate the rank correlation coefficient.

**Solution**

| | | RANK | | | |
|---|---|---|---|---|---|
| X | Y | X | Y | d | $D^2$ |
| 21 | 47 | 5 | 1 | 4 | 16 |
| 36 | 40 | 3 | 4 | -1 | 1 |
| 42 | 37 | 1 | 5 | -4 | 16 |
| 37 | 42 | 2 | 3 | -1 | 1 |
| 25 | 43 | 4 | 2 | 2 | 4 |
| Total | | | | $\sum d = 0$ | $\sum d^2 = 38$ |

$$\rho = 1 - \left(\frac{6\sum d^2}{N(N^2 -1)}\right)$$

$$= 1- \left(\frac{6 \times 38}{5\,(5^2 - 1)}\right)$$

= 1- 1.9 = - 0.9

**Tied Ranks**

When one or more values are repeated the two aspects- ranks of the repeated values and changes in the formula are to be considered.

**Example**

Find the rank correlation coefficient for the percentage of marks secured by a group of 8 students in Economics and Statistics.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Marks in Economics: | 50 | 60 | 65 | 70 | 75 | 40 | 70 | 80 |
| Marks in Statistics: | 80 | 71 | 60 | 75 | 90 | 82 | 70 | 50 |

**Solution**

Let X be Marks in Economics and Y be Marks in Statistics

| | | RANK | | | |
|---|---|---|---|---|---|
| X | Y | X | Y | d | $D^2$ |
| 50 | 80 | 7 | 3 | 4 | 16 |
| 60 | 71 | 6 | 5 | 1 | 1 |
| 65 | 60 | 5 | 7 | -2 | 4 |
| 70 | 75 | 3.5 | 4 | -0.5 | 0.25 |
| 75 | 90 | 2 | 1 | 1 | 1 |
| 40 | 82 | 8 | 2 | 6 | 36 |
| 70 | 70 | 3.5 | 6 | -2.5 | 6.25 |
| 80 | 50 | 1 | 8 | -7 | 49 |
| | | Total | | $\sum d = 0$ | $\sum d^2 = 113.5$ |

$$\rho = 1 - \left( \frac{6\{\sum d^2 + m(m^2-1)/12\}}{N(N^2-1)} \right)$$

When m=2 , $m(m^2-1)/12 = 0.5$

Therefore $\rho = 1 - \left( 6\{113.5+0.5\}/8(8^2-1)\} \right)$

= 1- 1.3571 = -0.3571

**Simple Linear Regression**

The line which gives the average relationship between the two variables is known as the regression equation. The regression equation is also called estimating equation.

**Uses**

1. Regression analysis is used in statistics and other displines.
2. Regression analysis is of practical use in determining demand curve, supply curve, consumption function, etc from market survey.
3. In Economics and Business, there are many groups of interrelated variables.
4. In social resarch, the relation between variables may not known; the relation may differ from place to place.
5. The value of dependent variable is estimated corresponding to any value of the independent variable using the appropriate regression equation.

**Method of Least Squares**

from a scatter diagram, there is virtually no limit as to the number of lines that can be drawn to make a linear relationship between the 2 variables

- the objective is to create a BEST FIT line to the data concerned
- the criterion is the called the method of least squares
- i.e. the *sum of squares* of the *vertical deviations* from the points to the line be a minimum (based on the fact that the dependent variable is drawn on the vertical axis)
- the linear relationship between the dependent variable (Y) and the independent variable(x) can be written as Y = a + bX , where a and b are parameters describing the vertical intercept and the slope of the regression.
- Similarly the linear relationship between the dependent variable (XY) and the independent variable(Y) can be written as X = a’ + b’Y , where a and b are parameters describing the vertical intercept and the slope of the regression.
-

**Calculating a and b:**

The values of a and b for the given pairs of values of (xi,yi) i=1,2,3…..are determined,
Using the normal equations as ,

$\sum y = Na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

Similarly, the values of a’ and b’ for the given pairs of values of (xi,yi) i=1,2,3…..are determined,
Using the normal equations as ,

$\sum x = Na' + b'\sum y$

$\sum xy = a'\sum y + b'\sum y^2$

**Methods of forming the regression equations**

- Regression equations on the basis of normal equations.
- Regression equations on the basis of X and Y and $b_{YX}$ and $b_{XY}$.

**Problem**

From the following data, obtain the two regression equations.

| | | | | | |
|---|---|---|---|---|---|
| X | 6 | 2 | 10 | 4 | 8 |
| Y | 9 | 11 | 5 | 8 | 7 use normal equations. |

**Solution**

| **X** | **Y** | **XY** | **$X^2$** | **$Y^2$** |
|---|---|---|---|---|
| 6 | 9 | 54 | 36 | 81 |
| 2 | 11 | 22 | 4 | 121 |
| 10 | 5 | 50 | 100 | 25 |
| 4 | 8 | 32 | 16 | 64 |
| 8 | 7 | 56 | 64 | 49 |
| $\sum x=0$ | $\sum y=0$ | $\sum xy=214$ | $\sum x^2=220$ | $\sum y^2=340$ |

Let the regression equation Y on X is Y = a + bX

The normal equations are ,

$\sum y = Na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

By substituting the values from the table, we get

5a+30b = 40 -------1

30a + 180b = 214 --------2

Solving these two equations we get,

a=11.90 and b= -0.65

Therefore the regression Y on X is Y = 11.90-0.65X.

Let the regression equation X on Y is X = a' + b'Y

The normal equations are,

$\sum x = Na + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

By substituting the values from the table, we get

5a’+40’b = 30 -------3

40a’ + 340b’ = 214 --------4

Solving these two equations we get,

a’ = 16.40 and b= -1.30

Therefore the regression equation X on Y is X = 16.40-1.30Y

**Example** From the data given below, find

(i) the two regression equations

(ii) The correlation coefficient between the variables X and y

(iii) The value of Y when X= 30

| X : | 25 | 28 | 35 | 32 | 31 | 36 | 29 | 38 | 34 | 32 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y : | 43 | 46 | 49 | 41 | 36 | 32 | 31 | 30 | 33 | 39 |

**Solution**

| X | Y | x= X- X` | Y= Y-Y` | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|---|
| 25 | 43 | -7 | 5 | -35 | 49 | 25 |
| 28 | 46 | -4 | 8 | -32 | 16 | 64 |
| 35 | 49 | 3 | 11 | 33 | 9 | 121 |
| 32 | 41 | 0 | 3 | 0 | 0 | 9 |
| 31 | 36 | -1 | -2 | 2 | 1 | 4 |
| 36 | 32 | 4 | -6 | -24 | 16 | 36 |
| 29 | 31 | -3 | -7 | 21 | 9 | 49 |
| 38 | 30 | 6 | -8 | -48 | 36 | 64 |
| 34 | 33 | 2 | -5 | -10 | 4 | 25 |
| 32 | 39 | 0 | 1 | 0 | 0 | 1 |
| 320 | 380 | 0 | 0 | -93 | 140 | 398 |

X` = 32, Y`= 38, $b_{xy} = \sum xy / \sum y^2 = -0.2337$, $b_{yx} = \sum xy / x^2 = -0.6643$

iv) Regression equation of Y on X ,(Y - Y` )= $b_{yx}$ (X-X`)

( Y – 38 ) = -0.6643(X-32) ⇒ Y = 59.26-0.6643X

(ii) Regression equation of X on Y , (X - X` )= $b_{xy}$ (Y-Y`)

( X – 32) = -0.2337Y +8.88 ⇒ X = 40.88 - 0.233 Y

(iii) $r = + \sqrt{b_{yx} b_{xy}} = -0.3940$

(iv) Y = 59.26-0.6643x30= 39

**Properties of Regression coefficients**

1. The two regression equations are generally different and are not to be interchanged in their usage. ` `
2. The two regression lines intersect at (X, Y).
3. Correlation coefficient is the geometric mean of two regression coefficients.
4. The two regression coefficients and the correlation coefficient have the same sign.
5. Both the regression coefficients and the correlation coefficient cannot be greater than one numerically and simultaneously.
6. Regression coefficients are independent of change of origin but are affected by the change of scale.
7. Each regression coefficient is in the unit of the measurement of the dependent variable.
8. Each regression coefficient indicates the quantum of change in the dependent variable corresponding to unit increase in the independent variable.

## POSSIBLE QUESTIONS
## UNIT II

**PART A (20 x 1 = 20 Marks)**
**Question number 1 – 20 online examinations**

**PART B (5 x 2= 20Marks)**

1) What are the types of Correlation?
2) Write any two properties of Correlation.
3) What is the range of Correlation Coefficient?
4) Define Positive Correlation.
5) What is meant by Regression?
6) What are the formulae for Regression co-efficients?
7) Distinguish between Correlation and Regression.
8) Write the formula for Rank Correlation, when more than one rank is repeated.
9) If $b_{xy}$ = -0.2337 and $b_{yx}$ = - 0.6643 then find the Correlation Coefficient.

10) What is Negative Correlation? Give an example?
11) Write down the formula for Karl Pearson's Coefficient of Correlation.
12) Define Scatter Diagram.
13) What is Simple Correlation?
14) Define Regression Equation.
15) When X = 40, Y = 60, $\sigma_x = 10$, $\sigma_Y = 15$ and r = 0.7 find the Regression Equation of Y on X.

**PART C (5 X 6 = 30 Marks)**

1) Calculate the Correlation Coefficient from the following variables.

| **Sales in ('0000)** | 57 | 58 | 59 | 59 | 60 | 61 | 62 | 64 |
|---|---|---|---|---|---|---|---|---|
| **Advertisement Expenditure ('000)** | 17 | 16 | 15 | 18 | 12 | 14 | 19 | 11 |

2) Marks obtained by 8 students in Accountancy (X) and Statistics (Y) are given below. Compute Rank Correlation Coefficient.

| X | 25 | 20 | 28 | 22 | 40 | 60 | 20 |
|---|---|---|---|---|---|---|---|
| Y | 40 | 30 | 50 | 30 | 20 | 10 | 30 |

3) Calculate the two Regression Equations from the following data.

| X | 10 | 12 | 13 | 12 | 16 | 15 |
|---|---|---|---|---|---|---|
| Y | 40 | 38 | 43 | 45 | 37 | 43 |

4) Calculate Karl Pearson's Coefficient of Correlation from the following data.

| **Wages** | 100 | 101 | 102 | 102 | 100 | 99 | 97 | 98 |
|---|---|---|---|---|---|---|---|---|
| **Cost of Living** | 98 | 99 | 99 | 97 | 95 | 92 | 95 | 94 |

5) From the data given below find the two Regression Equations.

| **X** | 10 | 12 | 13 | 12 | 16 | 15 |
|---|---|---|---|---|---|---|
| **Y** | 20 | 28 | 23 | 25 | 27 | 30. |

i) Estimate Y when X = 20. ii) Estimate X when Y = 35.

6) A comparison of the undergraduate Grade Point Averages of 10 corporate employees with their scores in a managerial trainee examination produced the results shown *in* the following table.

| **Exam Score** | 89 | 83 | 79 | 91 | 95 | 82 | 69 | 66 | 75 | 80 |
|---|---|---|---|---|---|---|---|---|---|---|
| **GPA** | 2.4 | 3.1 | 2.5 | 3.5 | 3.6 | 2.5 | 2.0 | 2.2 | 2.6 | 2.7 |

Measure the Correlation Coefficient between Exam scores and GPA by using Rank Method and also interpret the data given with the help of Scatter Diagram.

7) Develop the Regression Equation that best fit the data given below using annual income as an independent variable and amount of life insurance as dependent variable.

| **Annual Income (Rs. in 000's)** | 62 | 78 | 41 | 53 | 85 | 34 |
|---|---|---|---|---|---|---|

| **Amount of Life Insurance (Rs. in 00's)** | 25 | 30 | 10 | 15 | 50 | 7 |
|---|---|---|---|---|---|---|

8) The ranks of ten students in Economics and Statistics subjects are as follows.

| **Economics** | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Statistics** | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Calculate Spearman's Rank Correlation Coefficient.

9) You are given the following data:

| | X | Y |
|---|---|---|
| Arithmetic Mean | 36 | 85 |
| Standard Deviation | 11 | 8 |

Correlation coefficient between X and Y = 0.66

Find the two Regression Equations. And also find Correlation Coefficient.

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics and Research Methodology
UNIT II

**UNIT=II**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | The study of regression is considerably used by | **Economists and Businessmen** | Bioscience researcher | Mathematician | Astronaut | **Economists and Businessmen** |
| 2 | 2 | The average relationship existing between X and Y variables is described by | Regression graph | Regression diagram | **Regression line** | Regression bar | **Regression line** |
| 3 | 2 | The regression equation of Y on X is expressed as | X=a+bY | X=b+aY | **Y=a+bX** | Y=b +aX | **Y=a+bX** |
| 4 | 2 | The regression coefficient of X on Y is | **r * $(\sigma_x/\sigma_{y)}$** | r * $(\sigma_y/\sigma_{x)}$ | $\sigma_x/\sigma_y$ | $\sigma_y/\sigma_x$ | **r * $(\sigma_x/\sigma_{y)}$** |
| 5 | 2 | The under root of the product of two regression coefficients is equal to | coefficient of variation | coefficient of regression | **correlation coefficient** | correlation zero | **correlation coefficient** |
| 6 | 2 | Scatter diagram method is a | **Graphic** | Mathematical | Numerical | Algebraical | **Graphic** |
| 7 | 2 | If every change in X produces a corresponding decrease in Y then X and Y are said to be | uncorrelated | independent | positively correlated | **negatively correlated** | **negatively correlated** |
| 8 | 2 | Rank correlation was discovered by | R.A.Fisher | Sir Francis Galton | Karl Pearson | **Spearman** | **Spearman** |
| 9 | 2 | Correlation is used to measure | **closeness of relationship between variables** | one variable from another | nature of the distribution | central value | **closeness of relationship between variables** |
| 10 | 2 | Coefficient of correlation lies between | **1 and –1** | 0 and 1 | 0 and ∞ | 0 and –1 | **1 and –1** |
| 11 | 2 | While drawing a scatter diagram if all points appear to form a straight linegetting downward from left to right, then it is inferred that there is | a perfect positive correlation | simple positive correlation | **a perfect negative correlation** | no correlation | **a perfect negative correlation** |
| 12 | 2 | The range of the rank correlation coefficient is | 0 to 1 | **-1 to 1** | 0 to ∞ | -∞ to +∞ | **-1 to 1** |
| 13 | 2 | The technique used in measuring the closeness of the relationship between theVariables is referred to as | Range | standard deviation | median | **correlation analysis** | **correlation analysis** |
| 14 | 2 | If both the variables are varying in the same direction is known as | **positive correlation** | negative correlation | linear correlation | non linear correlation | **positive correlation** |
| 15 | 2 | X : 10 12 15 18 20 Y : 15 20 22 25 37 These two variables are | **Positively correlated** | negatively correlated | linearly correlated | non linearly correlated | **Positively correlated** |
| 16 | 2 | X : 20 30 40 60 80 Y : 40 30 22 15 10 These two variables are | Positively correlated | **negatively correlated** | linearly correlated | non linearly correlated | **negatively correlated** |
| 17 | 2 | If the two variables are varying in opposite direction the correlation is said to be | positive | **negative** | simple | partial | **negative** |
| 18 | 2 | When we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is the problem of | **multiple correlation** | Simple correlation | non linear correlation | linear correlation | **multiple correlation** |
| 19 | 2 | Study of three or more variables simultaneously is known as | Simple correlation | **multiple correlation** | linear correlation | non linear correlation | **multiple correlation** |
| 20 | 2 | If the amount of change in one variable tends to bear constant ratio to the amount ofchange in the other variable then the correlation is said to be | non linear | **linear** | multiple | simple | **linear** |
| 21 | 2 | X : 10 20 30 40 50 Y : 70 140 210 280 350 The above said two variables are | non linear | **linear** | multiple | simple | **linear** |
| 22 | 2 | If the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable is known as | **non linear** | linear | multiple | simple | **non linear** |
| 23 | 2 | In scatter diagram if the points are lying in a straight line from lower left hand corner to the upper right hand corner the correlation is said to be | perfectly negative | **perfectly positive** | positive | negative | **perfectly positive** |
| 24 | 2 | In scatter diagram if the points are lying in a straight line from upper left hand corner to the lower right hand corner the correlation is said to be | **perfectly negative** | perfectly positive | positive | negative | **perfectly negative** |
| 25<br>26 | 2<br>2 | A simple and non mathematical method of studying correlation between the variables is | Karl pearson's coefficient of correlation | **scatter diagram method** | concurrent deviation method | rank correlation | **scatter diagram method** |
| 27 | 2 | Correlation is the | Relationship between two values | **Relationship between two variables** | Significance between two variables | Significant level between two values | **Relationship between two variables** |
| 28 | 2 | If r =+1, the given two variables are | **perfectly positive** | perfectly negative | no correlation | Negative | **perfectly positive** |
| 29 | 2 | If r =-1, the given two variables are | perfectly positive | **perfectly negative** | positive | negative | **perfectly negative** |
| 30 | 2 | If r =0, the given two variables are having | perfect positive correlation | perfect negative correlation | **no correlation** | slight correlation | **no correlation** |
| 31 | 2 | Coefficient of correlation lies between | **1 and –1** | 0 and 1 | 0 and ∞ | 0 and –1 | **1 and –1** |
| 32 | 2 | The range of the rank correlation coefficient is | 0 to 1 | **-1 to 1** | 0 to ∞ | -∞ to +∞ | **-1 to 1** |
| 33 | 2 | Formula for rank correlation is | | | | | |
| 34 | 2 | The formula for computing Pearsonian r is | **$\Sigma xy / N\sigma_x \sigma_y$** | $\Sigma x / N\sigma_x \sigma_y$ | $\Sigma y / N\sigma_x \sigma_y$ | $\Sigma xy / \sigma_x \sigma_y$ | **$\Sigma xy / N\sigma_x \sigma_y$** |
| 35 | 2 | In rank correlation the sum of the differences of ranks between two variables shall be | **zero** | more than 1 | less than 1 | 10 | **zero** |
| 36 | 2 | Estimation of the value of one variable from the given value of another variable is done by | correlation analysis | **regression analysis** | chi square test | student's t test | **regression analysis** |
| 37 | 2 | If advertising and sales are correlated the expected amount of sales for a give Advertising expenditure is calculated by | correlation analysis | **regression analysis** | chi square test | student's t test | **regression analysis** |
| 38 | 2 | If advertising and sales are correlated the required amount of expenditure for attaininga given amount of sales is calculated by | correlation analysis | **regression analysis** | chi square test | student's t test | **regression analysis** |
| 39 | 2 | If yield of rice and rainfall are correlated the amount of rain required to achieve a certain production figure is calculated by | correlation analysis | **regression analysis** | chi square test | student's t test | **regression analysis** |
| 40 | 2 | The dictionary meaning of the term 'regression' is | **act of returning** | removing | reacting | relating | **act of returning** |
| 41 | 2 | The term 'Regreeion' was first used by | R.A.Fisher | **Sir Francis Galton** | Karl Pearson | Spearman | **Sir Francis Galton** |
| 42 | 2 | In 1877, the relationship between the height of the fathers and sons was studied by | R.A.Fisher | **Sir Francis Galton** | Karl Pearson | Spearman | **Sir Francis Galton** |
| 43 | 2 | The regression equation of X on Y is expressed as | **X=a+bY** | X=b+aY | Y=a+bX | Y=b +Ax | **X=a+bY** |
| 44 | 2 | The measure of the average relationship between two or more variables in terms ofthe original units of the data is referred to as | correlation | **regression** | standard deviation | correlation coefficient | **regression** |
| 45 | 2 | One of the most frequently used techniques in Economics and Business research is | correlation | **regression** | standard deviation | correlation coefficient | **regression** |
| 46 | 2 | The variable which is used to predict the variable of interest is called the | dependent variable | **independent variable** | common variable | multidependent variable | **independent variable** |
| 47 | 2 | The variable we are trying to predict is called | **dependent variable** | independent variable | common variable | multidependent variable | **dependent variable** |
| 48 | 2 | In correlation analysis $r_{xy}$ and $r_{yx}$ are | **symmetric** | not symmetric | multiples of two | multiples of five | **symmetric** |
| 49 | 2 | In regression analysis the regression coefficients $b_{xy}$ and $b_{yx}$ are | symmetric | **not symmetric** | multiples of two | multiples of five | **not symmetric** |
| 50 | 2 | In the regression equation Y = a+bX, Y is a | **dependent variable** | independent variable | common variable | multidependent variable | **dependent variable** |
| 51 | 2 | In the regression equation Y = a+bX, X is a | dependent variable | **independent variable** | common variable | multidependent variable | **independent variable** |
| 52 | 2 | The regression coefficient of Y on X is | r * $(\sigma_x/\sigma_{y)}$ | **r * $(\sigma_y/\sigma_{x)}$** | $\sigma_x/\sigma_y$ | $\sigma_y/\sigma_x$ | **r * $(\sigma_y/\sigma_{x)}$** |
| 53 | 2 | In the regression equation Y = a+bX, b is the | X intercept | Y intercept | **slope** | dependent variable | **slope** |
| 54 | 2 | In the regression equation Y = a+bX, a is the | X intercept | **Y intercept** | slope | dependent variable | **Y intercept** |
| 55 | 2 | In rank correlation the value of D is | $R_1+R_2$ | **$R_1$-$R_2$** | $R_1xR_2$ | $R_1/R_2$ | **$R_1$-$R_2$** |
| 56 | 2 | If three ranks are equal at 5th place, the rank given to them is | 5 | **6** | 7 | 5.5 | **6** |
| 57 | 2 | When equal ranks are assigned the adjustments made by adding | **1/12(m3-m)** | 1/6(m2-m) | 1/5(m2+m) | 1/12(m2-m) | **1/12(m3-m)** |
| 58 | 2 | The only method used with ranks not the actual value is | Karl pearson's coefficient of correlation | **rank correlation** | scatter diagram method | concurrent deviation method | **rank correlation** |
| 59 | 2 | If the data are of a qualitative nature like honesty,efficiency and intelligence etc the method used is | Karl pearson's coefficient of correlation | **rank correlation** | scatter diagram method | concurrent deviation method | **rank correlation** |
| 60 | 2 | If the dots in scatter diagram is too scattered we can say that | r = +1 | r = -1 | **r = 0** | r = 5 | **r = 0** |

## UNIT-III
## SYLLABUS

Probability- Definition, concepts, theorems (proofs of the theorems not necessary) and calculations of probability-simple problems, theoretical distributions-Binomial, Poisson and Normal distribution – simple problems

**Introduction:**

The theory of probability has its origin in the games of chance related to gambling such as tossing in a coin, throwing of a die, drawing cards from a pack of cards etc. Jerome Cardon, an Italian mathematician wrote a book on “GAMES OF CHANCE” which was published on 1663. Starting with games of chance, probability has become one of the basic tools of statistics. The knowledge of probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples.

Probability theory is being applied in the solution of social, economic, business problems. Today the concept of probability has assumed greater importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision – making research. Probability theory, in fact, is the foundation of statistical inferences.

**Definitions and basic concepts:**

The following definitions and terms are used in studying the theory of probability.

**Random experiment:**

Random experiment is one whose results depend on chance, that is the result cannot be predicted. Tossing of coins, throwing of dice are some examples of random experiments.

**Trial:**

Performing a random experiment is called a trial.

**Outcomes:**

The results of a random experiment are called its outcomes. When two coins are tossed the possible outcomes are HH, HT, TH& TT.

**Event:**

An outcome or a combination of outcomes of a random experiment is called an event. For example tossing of a coin is a random experiment and getting a head or tail is an event.

**Sample space:**

Each conceivable outcome of an experiment is called a sample point. The totality of all sample points is called a sample space and is denoted by S.

For example:

When a coin is tossed, the sample space is S = {H, T}. H and T are the sample points of the sample space S.

**Equally likely events:**

Two or more events are said to be equally likely if each one of them has an equal chance of occurring.

For example:

In tossing of a coin, the event of getting a head and the event of getting a tail are equally likely events.

**Mutually exclusive events:**

Two or more events are said to be mutually exclusive, when the occurrence of any one event excludes the occurrence of the other event. Mutually exclusive events cannot occur simultaneously.

For example:

When a coin is tossed, either the head or the tail will come up. Therefore the occurrence of the head completely excludes the occurrence of the tail. Thus getting head or tail in tossing of a coin is a mutually exclusive event.

**Exhaustive events:**

Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment.

For example:

While throwing a die, the possible outcomes are {1, 2, 3, 4, 5, 6} and hence the number of cases is 6.

**Complementary events:**

The event ‘A occurs’ and the event ‘A does not occur’ are complementary events of each other. The event ‘A does not occur’ is denoted by A' or A or $A^c$. The event and its complements are mutually exclusive.

For example:

In throwing a die, the event of getting odd numbers is {1, 3, 5} and getting even numbers is {2, 4, 6}. These two events are mutually exclusive and complement to each other.

**Independent events:**

Events are said to be independent if the occurrence of one does not affect the others. In the experiment of tossing a fair coin, the occurrence of the event ‘head’ in the first toss is independent of the occurrence of the event ‘head’ in the second toss, third toss and subsequent tosses.

**Definitions of probability:**

There are two types of probability. They are Mathematical probability and Statistical probability.

**1) Mathematical probability or a priori probability:**

If the probability of an event can be calculated even before the actual happening of the event, that is, even before conducting the experiment, it is called ***Mathematical probability.***

If the random experiments results in µ Q¶ exhaustive, mutually exclusive and equally likely cases, out of which µ P¶ are favourable to the occurrence of an event A, then the ratio *m/n* is called the probability of occurrence of event A, denoted by P(A), is given by

$$\mathbf{P(A) = \frac{m}{n} = \frac{\text{Number of cases favourable to the event A}}{\text{Total number of exhaustive cases}}}$$

Mathematical probability is often called ***classical probability*** or a ***priori probability*** because if we keep using the examples of tossing of fair coin, dice etc., we can state the answer in advance (*prior*), without tossing of coins or without rolling the dice etc.,

The above definition of probability is widely used, but it cannot be applied under the following situations:

(1) If it is not possible to enumerate all the possible outcomes for an experiment.
(2) If the sample points (outcomes) are not mutually independent.
(3) If the total number of outcomes is infinite.
(4) If each and every outcome is not equally likely.

Some of the drawbacks of classical probability are removed in another definition given below:

**2) Statistical probability or a posteriori probability:**

If the probability of an event can be determined only after the actual happening of the event, it is called ***statistical probability.***

If an event occurs ***m*** times out of ***n,*** its relative frequency is ***m/n.***

In the limiting case, when ***n*** becomes sufficiently large it corresponds to a number which is called the probability of that event.

In symbol, P(A) = Limit (*m/n*)
→**n**∞

The above definition of probability involves a concept which has a long term consequence. This approach was initiated by the mathematician Von Mises.

If a coin is tossed 10 times we may get 6 heads and 4 tails or 4 heads ad 6 tails or any other result. In these cases the probability of getting a head is **not 0.5** as we consider in Mathematical probability.

However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails and we can see that the probability of getting head approaches 0.5. The statistical probability calculated by conducting an actual experiment is also called a ***posteriori probability*** or ***empirical probability.***

**Axiomatic approach to probability:**

The modern approach to probability is purely axiomatic and it is based on the set theory. The axiomatic approach to probability was introduced by the Russian mathematician A.N. Kolmogorov in the year 1933.

**Axioms of probability:**

Let S be a sample space and A be an event in S and P(A) is the probability satisfying the following axioms:

(1) The probability of any event ranges from zero to one.

i.e. $0 \leq P(A) \leq 1$

(2) The probability of the entire space is 1.

i.e. $P(S) = 1$

(3) Is $A_1$ , $A_2$,…… is a sequence of mutually exclusive events in S, then

$P(A_1 \cup A_2 \cup \ldots..) = P_1(A) + P_2(A) + \ldots..$

**Interpretation of statistical statements in terms of set theory:**

S → Sample space.

A → A does not occur.

$A \cup A = S$

$A \cap B = \emptyset$ → A and B are mutually exclusive.

$A \cup B$ → Event A occurs or B occurs or both A and B occurs.
(at least one of the events A or B occurs)

$A \cap B$ → Both the events A and B occur.

$\overline{A} \cap \overline{B}$ → Neither A nor B occurs.

$A \cap \overline{B}$ → Event A occurs and B does not occur.

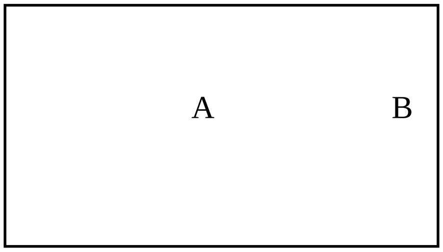
$\overline{A} \cap B$ → Event A does not occur and B occurs.

**Addition theorem on probabilities:**

We shall discuss the addition theorem on probabilities for mutually exclusive events and not mutually exclusive events.

**1) Addition theorem on probabilities for mutually exclusive events:**

If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of individual properties of A and B. i.e$P(A \cup B) = P(A) + P(B)$. This is clearly stated in axioms of probability.



**2) Addition theorem on probabilities for not – mutually exclusive events:**

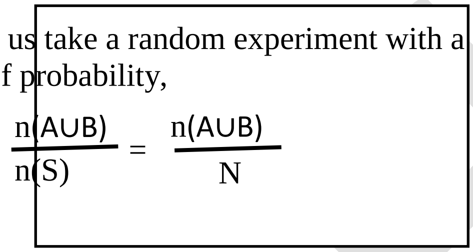
If two events A and B are not – mutually exclusive, the probability of the event that either A or B or Both occur is given as

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

**Proof:**

Let us take a random experiment with a sample space S of N sample points. Then by the definition of probability,

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N}$$



From this diagram, using the axiom for the mutually exclusive events, we write

$$P(A \cup B) = \frac{n(A) + n(\bar{A} \cap B)}{N}$$

Adding and subtracting n(A∩B) in the numerator,

$$= \frac{n(A) + n(\bar{A} \cap B) + n(A \cap B) - n(A \cap B)}{N}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N}$$

$$= \frac{n(A)}{N} + \frac{n(B)}{NN} - \frac{n(A \cap B)}{}$$

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

**Note:**

In the case of three events A,B,C is

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

**Compound events:**

The joint occurrence of two or more events is called compound events. Thus compound simply the simultaneous occurrence of two or more simple events.

For example,

In tossing of two fair coins simultaneously, the event of getting 'atleast one head' is a compound event as it consists of joint occurrence of two simple events.
Namely,
Event A = one head appears i.e A = {HT, TH} and
Event B = two heads appears i.e B = {HH}

Similarly, if a bag contains 6 white and 6 red balls and we make a draw of 2 balls at random, then the events that 'both are white' or 'one is white and one is red' are compound events.

The compound events may be further classified as

(1) Independent event
(2) Dependent event

**(1) Independent event:**

If two or more events occur in such a way that the occurrence of one does not affect the occurrence of another, they are said to be independent events.

For example,

If a coin is tossed twice, the results of the second throw would in no way be affected by the results of the first throw.

Similarly, if a bag contains 5 white and 7 red balls and then two balls are drawn one by one in such a way that the first ball is replaced before the second one is drawn. In this situation, the two events 'the first ball is white' and 'second ball is red', will be independent, since the composition of the balls in the bag remains unchanged before a second draw is made.

**Dependent events:**

If the occurrence of one event influences the occurrence of the other, then the second event is said to be dependent on the first.

In the above example, if we do not replace the first ball drawn, this will change the composition of balls in the bag while making the second draw and therefore the event of 'drawing a red ball' in the second will depend on event (first ball is red or white) occurring in first draw.

Similarly, is a person draw a card from a full pack and does not replace it, the result of the draw made afterwards will be dependent on the first draw.

**Conditional probability:**

Let A be any event with p(A) > 0. The probability that an event b occurs subject to the condition that A has already occurred is known as the conditional probability of occurrence of the event B on the assumption that the event A has already occurred and is denoted by the symbol P(B/A) or P(B│A) and is read as the probability of B given A.

The same definition can be given as follows also:

Two events A and B are said to be dependent when A can occur only when B is known to have occurred (or vice versa). The probability attached to such an event is called **conditional probability** and is denoted by P(B/A) or, in other words, probability of B is given that A has occurred.

If two events A and B are dependent, then the conditional probability of B given A is

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

Similarly, the conditional probability of A given B is given as

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

**Note:**

If the events A and B are independent, that is the probability of occurrence of any one of the P(A/B) = P(A) and P(B/A) = P(B)

**Multiplication theorem on probabilities:**

We shall discuss multiplication theorem on probabilities for both independent and dependent events.

**(1) Multiplication theorem on probabilities for independent events:**

If two events A and B are independent, the probability that both of them occur is equal to the product of their individual probabilities.

i.e. $P(A\cap B) = P(A).P(B)$

**Proof:**

Out of $n_1$ possible cases let $m_1$ cases be favourable for the occurrence of the event A.

Therefore, $P(A) = \dfrac{m_1}{n_1}$

Out of $n_2$ possible cases, let $m_2$ cases be favourable for the occurrence of the event B.

Therefore, $P(B) = \dfrac{m_2}{n_2}$

Each of $n_1$ possible cases can be associated with each of the $n_2$ possible cases.

Therefore, the total number of possible cases for the occurrence of the event 'A' and 'B' is $n_1$ x $n_2$. Similarly, each of the $m_1$ favourable cases can be associated with each of the $m_2$ favourable cases. So the total number of favourable cases for the event 'A' and 'B' is $m_1$ x $m_2$

$$\text{Therefore, } P(A\cap B) = \frac{m_1\ m_2}{n_1\ n_2}$$

$$= \frac{m_1}{n_1}\ \frac{m_2}{n_2}$$

$$= P(A).P(B)$$

**Note:**

The theorem can be extended to three or more independent events. If A, B, C……. be independent events, then

$$P(A\cap B\cap C\ldots\ldots) = P(A).P(B).P(C)\ldots\ldots$$

**Note:**

If A and B are independent then the complements of A and B are also independent.

$$\text{i.e. } P(\overline{A}\cap\overline{B}) = P(\overline{A}).P(\overline{B})$$

**(2) Multiplication theorem for dependent events:**

If A and B be two dependent events, i.e. the occurrence of one event is affected by the occurrence of the other event, then the probability that both A and B will occur is

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

**Proof:**

Suppose an experiment results in n exhaustive, mutually exclusive and equally likely outcomes, m of them being favourable to the occurrence of the event A.

Out of these n outcomes let $m_1$ be favourable to the occurrence of another event B.

Then the outcomes favourable to the happening of the events 'A and B' are $m_1$ .

$$\text{Therefore, } P(A\cap B) = \frac{m_1}{n}$$

$$\frac{m_1}{n} \qquad \frac{m}{m} \qquad \frac{mm_1}{nm}$$

$$= \frac{\quad}{\quad} \text{ x } \frac{\quad}{\quad} = \frac{\quad}{\quad}$$

$$= \frac{m}{{}_1\, n} \text{ x } \frac{m}{{}_1\, m}$$

Therefore, P(A∩B) = P(A) . P(B/A)

**Note:**

In the case of three events A, B, C, P(A∩B∩C) = P(A) . P(B/A) . P(C/A∩B). i.e. the probability of A, B and C is equal to the probability of A times the probability of B given that A has occurred, times the probability of C given that both A and B have occurred.

**BAYE'S Theorem:**

The concept of conditional probability discussed earlier takes into amount information about the occurrence of one event to predict the probability of another event. This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to specific cause. The procedure for revising these probabilities is known as Bayes theorem.

The principle was given by Thomas Bayes in 1763. By this principle, assuming certain prior probabilities, the posterior probabilities are obtained. That is why Bayes' probabilities are also called posteriori probabilities.

**Bayes' Theorem or Rule (Statement only):**

Let $A_1$ , $A_2$, $A_3$,…………A□ be a set of n mutually exclusive and collectively exhaustive events and P(A), P($A_1$ ),………P(A□ ) are their corresponding probabilities. If B is another event such that P(B) is not zero and the priori probabilities P(BA)│i , i = 1, 2 …….. n are also known. Then

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

**Basic principles of Permutation and Combination Factorial:**

The consecutive product of first *n* natural numbers is known as ***factorial* n** and is denoted as **n!** *n*
or ∟

That is n! = 1x2x3x4x5x……..n
3! = 3x2x1
4! = 4x3x2x1
5! = 5x4x3x2x1
Also, 5! = 5x(4x3x2x1) = 5x(4!)

Therefore, this can be algebraically written as n! = nx(n – 1)!, note that 1! = 1 and 0! = 1.

**Permutations:**

Permutation means arrangement of things in different ways. Out of three things A, B, C taking two at a time, we can arrange them in the following manner.

| | |
|---|---|
| A B | B A |
| A C | C A |
| B C | C B |

Here we find 6 arrangements. In the arrangements order of arrangement is considered. The arrangement AB and the other arrangement BA are different.

The number of arrangements of the above is given as the number of permutations of 3 things taken 2 at a time which gives the value 6. This is written symbolically, ${}_3P_2 = 6$.

Thus the number of arrangements that can be made out of *n* things taken *r* at a time is known as the number of permutation of *n* things taken *r* at a time and is denoted as nPr.

The expression of nPr is given below:

nPr = n (n-1) (n-2) …………….[n – (r-1)]

The same can be written in factorial notation as follows:

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

For example,

To find ${}_{10}P_3$ we write this as follows:

$${}_{10}P_3 = 10\ (10-1)\ (10-2)$$
$$= 10x9x8$$
$$= 720.$$

[To find ${}_{10}P_3$, start with 10, write the product of 3 consecutive natural numbers in the descending order]

Simplifying ${}_{10}P_3$ using factorial notation:

$${}_{10}P_3 = \frac{\_10!}{(10-3)!} \quad \frac{10x9x8x7x6x5x4x3x2x1}{7x6x5x4x3x2x1}$$

$$= 10x9x8$$
$$= 720.$$

Note that,

$\square P_0 = 1$, $\square\ P_1 = n$, $\square\ P\square\ = n!$

**Combinations:**

A combination is a selection of objects without considering the order of arrangements.

For example,

Out of three things A, B, C we have to select two things at a time.

This can be selected in three different ways as follows:

A B A C B C

Here, the selection of the object A B and B A re one and the same. Hence the order of arrangement is not considered in combination. Here the number of combinations from 3 different things taken 2 at a time is 3.

This is written symbolically, $_3C_2 = 3$

Thus, the number of combination of n different things, taken r at a time is given by

$$□C_r = \frac{□\,P_r}{r!} \quad \text{(or)} \quad □\,C_r = \frac{n!}{(n-r)!\,r!}$$

Note that $□\,C_0 = 1$, $□\,C_1 = n$, $□\,C□ = 1$

Find $_{10}C_3$, $\quad _{10}C_3 = \frac{_{10}P_3}{3!} = \frac{10x9x8}{1x2x3} = 120$

Find $_8C_4$, $\quad _8C_4 = \frac{8x7x6x5}{1x2x3x4} = 70$

[To find $_8C_4$: In the numerator, first write the product of 4 natural numbers starting with 8 in descending order and in the denominator write the factorial 4 and then simplify.]

Compare $_{10}C_8$ and $_{10}C_2$,

$$_{10}C_8 = \frac{10x9x8x7x6x5x4x3}{1x2x3x4x5x6x7x8} = \frac{10x9}{1x2} = 45$$

$$_{10}C_2 = \frac{10x9}{1x2} = 45$$

From the above, we find $_{10}C_8 = {}_{10}C_2$

This can be got by the following method also:

$_{10}C_8 = {}_{10}C_{(10-8)} = {}_{10}C_2$

This method is very useful, when the difference between *n* and *r* is very high in $□C_r$.

This property of the combination is written as $□C_r = □\,C_{(□-r)}$.

To find $_{200}C_{198}$ we can use the above formula as follows:

$$_{200}C_{198} = {}_{200}C_{(200-198)} = {}_{200}C_2 = \frac{200x199}{1x2} = 19900.$$

**Example:**

Out of 13 players, 11 players are to be selected for a cricket team. In how many ways can this be done?

Out of 13 players, 11 players are selected in $_{13}C_{11}$ ways

i.e. $_{13}C_{11} = {}_{13}C_2 = \frac{13x12}{1x2} = 78.$

**Example 1:**

Three coins are tossed simultaneously. Find the probability that (i) no head (ii) one head

(iii) two heads (iv) atleast two heads (v) atmost two heads appear.

**Solution:**

The sample space for the 3 coins is
S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}; n(S) = 8

(i) No head appear A = {TTT}; n(A) = 1
∴P(A) = 1/8

(ii) One head appear B = {HTT, THT, TTH}; n(B) = 3
∴P(B) = 3/8

(iii) Two heads appear C = {HHT, HTH, THH}; n(C) = 3
∴ P(C) = 3/8

(iv) Atleast two heads appear D = {HHT, HTH, THH, HHH}; n(D) = 4
∴P(D) = 4/8 = 1/2

(v) Atmost two heads appear E = {TTT, HTT, THT, TTH, HHT, HTH, THH}; n(E) = 7
∴P(E) = 7/8

**Example 2:**

When two dice are thrown, find the probability of getting doublets (same number on both dice)

**Solution:**

When two dice are thrown, the number of points in the sample space is n(S) = 36
Getting doublets A = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}
∴P(A) = 6/36 = 1/6.

**Example 3:**

A card is drawn at random from a well shuffled pack of 52 cards. What is the probability that it is
(i) an ace (ii) a diamond card.

**Solution:**

We know that the pack contains 52 cards ∴ n(S) = 52

(i) There are 4 aces in a pack ∴n(A) = 4
∴P(A) = 4/52 = 1/13

(ii) There are 13 diamonds in a pack ∴n(B) = 13
∴P(B) = 13/52 =1/4

**Example 4:**

A ball is drawn at random from a box containing 5 green, 6 red and 4 yellow balls. Determine the probability that the ball drawn is (i) green (ii) red (iii) yellow (iv) green or red (v) not yellow.

**Solution:**

Total number of balls in the box = 5+6+4 =15 balls

(i) Probability of drawing a green ball = 5/15 = 1/3

(ii) Probability of drawing a red ball = 6/15 = 2/5

(iii) Probability of drawing a yellow ball = 4/15

(iv) Probability of drawing a green or a red ball = 5/15 + 6/15 = 11/15

(v) Probability of getting not an yellow ball = 1 – P(yellow)
= 1 – 4/15
= 11/15

**Example 5:**

Two dice are thrown, what is the probability of getting the sum being 8 or the sum being 10?

**Solution:**

Number of sample points in throwing two dice at a time is n(S) = 36

Let A = {the sum being 8}

$\therefore$ A = {(6,2), (5,3), (4,4), (3,5), (2,6)}; P(A) = 5/36

B = {the sum being 10}

$\therefore$ B = {(6,4), (5,5), (4,6)}; P(B) = 3/36

A∩B = {0}; n(A∩B) = 0

$\therefore$the two events are mutually exclusive

$\therefore$P(A∪B) = P(A) + P(B)
= 5/36 + 3/36
= 8/36 = 2/9.

**Example 6:**

Two dice are thrown simultaneously. Find the probability that the sum being 6 or same number on both dice.

**Solution:**

n(S) = 36

The total is 6

∴ A = {(5,1), (4,2), (3,3), (2,4), (1,5)}; P(A) = 5/36

Same number on both dice

∴ B = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}; P(B) = 6/36

∴ A∩B = {(3,3)}; P(A∩B) = 1/36

Here, the events are not mutually exclusive.

∴ P(A∪B) = P(A) + P(B) – P(A∩B)

= 5/36 + 6/36 – 1/36

= 10/36 = 5/18.

**Example 7:**

Two persons A and B appeared for an interview for a job. The probability of selection of A is 1/3 and that of B is 1/2. Find the probability that (i) both of them will be selected

(ii) only one of them will be selected (iii) none of them will be selected

**Solution:**

P(A) = 1/3; P(B) = 1/2; P(A) = 2/3; P(B) = 1/2

Selection or non – selection of any one of the candidate is not affecting the selection of the other. Therefore, A and B are independent events.

(i) Probability of selecting both A and B

P(A∩B) = P(A) . P(B)

= 1/3 x ½

= 1/6

(ii) Probability of selecting any one of them = P (selecting A and not selecting B)

+

P (not selecting A and selecting B)

∴ P(A∩B) + P(A∩B) = P(A) . P(B) + P(A) . P(B)

= 1/3 x 1/2 + 2/3 x 1/2

= 1/6 + 2/6

= 3/6

= 1/2

(iii) Probability of not selecting both A and B

i.e. P(A∩B) = P(A) . P(B)

= 2/3 . 1/2

= 1/3

**Example 8:**

There are three T.V programmes A, B and C which can be received in the city of 2000

families. The following information is available on the basis of surgery.
1200 families listen to programme A; 1100 families listen to programme B; 800 families listen to programme C; 765 families listen to programme A and B; 450 families listen to programme A and C; 400 families listen to programme B and C; 100 families listen to programme A, B and C. Find the probability that a family selected at random listens atleast one or more T.V programmes.

**Solution:**

Total number of families n(S) = 2000

Let,

$n(A) = 1200$
$n(B) = 1100$
$n(C) = 800$
$n(A\cap B) = 765$
$n(A\cap C) = 450$
$n(B\cap C) = 400$
$n(A\cap B\cap C) = 100$

Let us first find $n(A\cup B\cup C) = ?$

$$n(A\cup B\cup C) = n(A) + n(B) + n(C) - n(A\cap B) - n(A\cap C) - n(B\cap C) + n(A\cap B\cap C)$$
$$= 1200 + 1100 + 800 – 765 – 450 – 400 +100$$
$$n(A\cup B\cup C) = 1585$$

$$\text{Now, } P(A\cup B\cup C) = \frac{n(A\cup B\cup C)}{n(S)}$$
$$= \frac{1585}{2000}$$
$$= 0.792$$

Therefore, about 79% chance that a family selected at random listens to one or more T.V programmes.

**Example 9:**

A stockist has 20 items in a lot. Out of which 12 are non – defective and 8 are defective. A customer selects 3 items from the lot. What is the probability that out of these three items
(i) three items are non – defective (ii) two are non - defective and one is defective

**Solution:**

(i) Let the event, that all the three items are non – defective, be denoted by $E_1$. There are 12 non – defective items and out of them 3 can be selected in $12C_3$ ways i.en$(E_1) = 12C_3$
Total number of ways in which 3 items can be selected are $20C_3$i.e $n(S) = 20C_3$

$n(E_1)$ $12C_3$
$n(S)$ $20C_3$

$\therefore P(E_1) = \frac{\quad}{\quad} = \frac{\quad}{\quad}$

$$= \frac{12x11x10}{20x19x18}$$

$= 0.193$

(ii) Let the event, that two items are non – defective and one is defective be denoted by $E_2$. Two non – defective items out of 12 can be selected in $12C_2$ ways. One item out of 8 defective can be selected in $8C_1$ ways. Thus, $n(E_2) = 12C_2 . 8C_1$

Then the probability $P(E_2) = \frac{n(E_2)}{n(S)}$

$$= \frac{12C_2 . 8C_1}{20C_3}$$

$$= \frac{12x11x8x3}{20x19x18} = 0.463$$

**Example 10:**

A test paper containing 10 problems is given to three students A, B, C. It is considered that student A can solve 60% problems, student B can solve 40% problems and student C can solve 30% problems. Find the probability that the problem chosen from the test paper will be solved by all the three students.

**Solution:**

Probability of solving the problem by A = 60%
Probability of solving the problem by B = 40%
Probability of solving the problem by C = 30%

Solving the problem by a student is independent of solving the problem by the other students.

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

$=(60/100)\times(40/100)\times(30/100)$

$=0.6\times0.4\times0.3$

0.072.

**Example 11:**

From a pack of 52 cards, 2cards are drawn at random. Find the probability that one is king and the other is queen.

**Solution:**

From a pack of 52 cards 2 cards are drawn n(S)=52C2

Selection of one king is in 4C1ways

Selection of one queen is in 4C1 ways

Selection of one king and one queen is in 4C1 .4C ways
ie n(E) = 4C1 .4C1

P(E)= $\frac{n(E)}{n(S)}$

= $\frac{C4.C4}{C52}$

=4×4÷$\frac{52\times51}{1\times2}$

=$\frac{4\times4\times2}{52\times51}$

= $\frac{8}{663}$.

**Example 12:**
An urn contains 4 black balls and 6 white balls. If 3 balls are drawn at random, find the probability that (i) all are black
(ii) all are white
**Solution:**
Total number of balls = 10
Total number ways of selecting 3 balls = 10C3

(i) Number of ways of drawing 3 black balls = 4C3

Probability of drawing 3 black balls = $\frac{C4}{C10}$

=$\frac{4\times3\times2}{10\times9\times8}$

=$\frac{1}{30}$

(ii) Number of ways of drawing 3 white balls = 6C3

Probability of drawing 3 white balls $=\frac{C6}{C10}$

$=\frac{6\times5\times4}{10\times9\times8}$

$=\frac{1}{6}$.

**Example 13:**
A box containing 5 green balls and 3 red colour balls. Find the probability of selecting 3 green colour balls one by one
(i) without replacement (ii) with replacement
**Solution:**
(i) Selection without replacement
Selecting 3 balls out of 8 balls = 8C3 ways
i.e n(S) = 8C3
Selecting 3 green balls in 5C3 ways

P(3 green balls) $=\frac{C3(5)}{C3(8)}$

$=\frac{5\times4\times3}{8\times7\times6}$

$=\frac{5}{28}$.

(ii) Selection with replacement
When a ball is drawn and replaced before the next draw, the number of balls in the box remains the same. Also the 3 events
of drawing a green ball in each case is independent.
Therefore, Probability of drawing a green ball in each case is $\frac{5}{8}$

The event of selecting a green ball in the first, second and third event are same,

Probability of drawing 3 green balls = $\frac{5}{8}\times\frac{5}{8}\times\frac{5}{8} = \frac{125}{512.}$

**Example 14:**
A box contains 5 red and 4 white marbles. Two marbles are drawn successively from the box without replacement and it is noted that the second one is white. What is the probability that the first is also white?
**Solution:**
If w1 , w2 are the events ‘ white on the first draw’ , ‘ white on the second draw’ respectively.
Now we are looking for P(w1 /w2 )

$$P(w1/w2) = \frac{P(w1\cap w2)}{P(w2)}$$

$$=\frac{(4/9)(3/8)}{(3/8)}$$

$$=\frac{4}{9.}$$

**Example 15:**
A bag contains 6 red and 8 black balls. Another bag contains 7 red and 10 black balls. A bag is selected and a ball is
drawn. Find the probability that it is a red ball.
**Solution:**
There are two bags
Therefore, probability of selecting a bag = $\frac{1}{2.}$

Let A denote the first bag and B denote the second bag.
Then P(A) = P(B)= $\frac{1}{2}$

Bag ‘ A’ contains 6 red and 8 black balls.
Then, Probability of drawing a red ball is 6

$$\frac{\;}{14}$$

Probability of selecting bag A and drawing a red ball from that bag is P(A). P(R/A) =$\frac{1}{2}\times\frac{6}{14}=\frac{3}{14}$.

Similarly probability of selecting bag B and drawing a red ball from that bag is P(B). P(R/B) = $\frac{1}{2}\times\frac{7}{17}=\frac{7}{34}$

All these are mutually exclusive events
Then,
Probability of drawing a red ball either from the bag A or B is
P(R) = P(A) P(R/A) + P(B) P(R/B)

$$= \frac{3}{14}+\frac{7}{34}$$

$$= \frac{17\times3+7\times7}{238}$$

$$=\frac{51+49}{238}$$

$$=\frac{100}{238}$$

$$=\frac{50}{119}.$$

**Example 16:**
If P(A ⋂ B) = 0.3, P(A) = 0.6, P(B) = 0.7 Find the value of P(B/A) and P(A/B)

Solution:

P(B/A)=P(A∩B)

$$=\frac{0.3}{\;}$$

0.6

$= \frac{1}{2.}$

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

$P(B)$

$= \frac{0.3}{0.7}$

$= \frac{3}{7.}$

**Example 17:**
In a certain town, males and females form 50 percent of the population. It is known that 20 percent of the males and 5 percent
of the females are unemployed. A research student studying the employment situation selects unemployed persons at random.
What is the probability that the person selected is (i) a male (ii) a female?
**Solution:**
Out of 50% of the population 20% of the males are unemployed.

i.e., $\frac{50}{100} \times \frac{20}{100} = \frac{10}{100} = 0.10$

Out of 50% the population 5% of the females are unemployed.

i.e. , $\frac{50}{100} \times \frac{5}{100} = \frac{25}{1000} = 0.025$

Based on the above data we can form the table as follows:

| | Employed | Unemployed | Total |
|---|---|---|---|
| Males | 0.400 | 0.100 | 0.50 |
| Females | 0.475 | 0.025 | 0.50 |
| Total | 0.875 | 0.125 | 1.00 |

Let a male chosen be denoted by M and a female chosen be denoted by F.

1)$P(M/U) = P(M\cap U)$ 0.10

——— = ——— = 0.80
P(U) 0.125
2)P(F/U) = P(F∩U) 0.025
——— = ——— =0.20
P(U) 0.125

**Example 18:**

Two sets of candidates are competing for the positions on the Board of directors of a company. The probabilities that the first and second sets will win are 0.6 and 0.4 respectively. If the first set
wins, the probability of introducing a new product is 0.8, and the corresponding probability if the second set wins is 0.3. What is the probability that the new product will be introduced?

**Solution:**

Let the probabilities of the possible events be:
P(A1 ) = Probability that the first set wins = 0.6
P(A2 )= Probability that the second set wins = 0.4
P(B) =Probability that a new product is introduced
P(B/A1 ) =Probability that a new product is introduced given that the first set wins = 0.8
P(B/A2 ) = Probability that a new product is introduced given that the second set wins = 0.3

Then the rule of addition gives:
P(new product) = P(first set and new product) + P(second set and new product)
i.e P(B) = P(A1B) + P(A2B)
= P(A1) P(B/A1 ) + P(A2 ).P(B/A2 )

= 0.6×0.8+0.4×0.3
= 0.60

**Example 19:**

Three persons A, B and C are being considered for the appointment as the chairman for a company whose chance of being selected for the post are in the proportion 4:2:3 respectively. The probability that A, if selected will introduce democratization in the company structure is 0.3 the corresponding probabilities for B and C doing the same are respectively 0.5 and 0.8. What is the probability that democratization would be introduced in the company?

**Solution:**

Let A1 and A2 and A3 denote the events that the persons A, B and C respectively are selected as chairman and let E be the event of introducing democratization in the company structure.
Then we are given,

$P(A1)=\frac{4}{9}$ $P(A2)=\frac{2}{9}$ $P(A3)=\frac{3}{9}$

P(E/A1) = 0.3
P(E/A2 ) = 0.5
P(E/A ) = 0.8

The event E can materialize in the following mutually exclusive ways:
(i) Person A is selected and democratization is introduced
ie A⋂ E happens

(ii) Person B is selected and democratization is introduced
i.e., A2⋂E happens

(iii) Person C is selected and democratization is introduced
ie A ⋂ E happens

Thus E = (A1 ⋂ E) U (A2⋂E) U(A3⋂ E) , where these sets are disjoint

Hence by addition rule of probability we have
P(E) = P(A1⋂ E) + P(A2⋂ E) + P(A3⋂ E)
= P(A1) P(E/A1 ) + P(A2) P(E/A2 ) + P(A3 ) P(E/A3 )

$$=\frac{4}{9}\times 0.3+\frac{2}{9}\times 0.5+\frac{3}{9}\times 0.8$$

$$=\frac{46}{90}$$

$$=\frac{23}{45}$$

**Example 21:**
A company has two plants to manufacture motorbikes. Plant I manufactures 80 percent of motor bikes, and plant II manufactures 20 percent. At Plant I 85 out of 100 motorbikes are rated standard quality or better. At plant II only 65 out of 100 motorbikes are rated standard quality or better.

(i) What is the probability that the motorbike, selected at random came from plant I. if it is known that the motorbike is of standard quality?

(ii) What is the probability that the motorbike came from plant II if it is known that the motor bike is of standard quality?

**Solution:**

Let A1 be the event of drawing a motorbike produced by plant I.

A2be the event of drawing a motorbike produced by plant II. B be the event of drawing a standard quality motorbike produced by plant I or plant II.

Then from the first information, P(A1 ) = 0.80, P(A2 ) = 0.20

From the additional information
P(B/A1) = 0.85

P(B/A2 ) = 0.65
The required values are computed in the following table.

The final answer is shown in last column of the table.

| Event | Prior Probability | Conditional probability of event B given A1 P(B/A1) | Joint probability P(A1∩B)= P(A1)P(B/A1) | Posterior (revised) Probability P(A1/B)=$\frac{P(A\cap B)}{P(B)}$ |
|---|---|---|---|---|
| A1 | 0.80 | 0.85 | 0.68 | 0.68/0.81=68/81 |
| A2 | 0.20 | 0.65 | 0.13 | 0.13/0.81=13/81 |
| Total | 1.00 | | P(B)=0.81 | 1 |

Without the additional information, we may be inclined to say that the standard motor bike is drawn from plant I output, since
P(A ) = 80% is larger than P(A ) =20%

**SOME IMPORTANT THEORETICAL DISTRIBUTIONS**

## 3.1 BINOMIAL DISTRIBUTION

### 3.1.0 Introduction:

In this chapter we will discuss the theoretical discrete distributions in which variables are distributed according to some definite probability law, which can be expressed mathematically. The Binomial distribution is a discrete distribution expressing the probability of a set of dichotomous alternative i.e., success or failure. This distribution has been used to describe a wide variety of
process in business and social sciences as well as other areas.

### 3.1.
### 1 Bernoulli Distribution:

A random variable X which takes two values 0 and 1 with probabilities q and p i.e., P(x=1) = p and P(x=0) = q, q = 1 p, is
called a Bernoulli variate and is said to be a Bernoulli Distribution, where p and q takes the probabilities for success and failure respectively. It is discovered by Swiss Mathematician James Bernoulli (1654-1705).

Examples of Bernoulli' s Trails are:

1) Toss of a coin (head or tail)
2) Throw of a die (even or odd number)
3) Performance of a student in an examination (pass or fail)

### 3.1.
### 2 Binomial Distribution:

A random variable X is said to follow binomial distribution,
if its probability mass function is given by
P (X = x) = P(x) = nC x pqx n-x ; x = 0, 1,2, …,n
0 ; otherwise

Here, the two independent constants n and p are known as the ' parameters' of the distribution. The distribution is completely
determined if n and p are known. x refers the number of successes.

If we consider N sets of n independent trials, then the number of
times we get x success is N(nCpq x *x*n-x ).
It follows that the terms in the expansion of N (q + p) gives the frequencies of the occurrences n of 0,1,2,...,x,...,n success in the N sets of independent trials.

### 3.1.

**3 Condition for Binomial Distribution:**

We get the Binomial distribution under the following experimental conditions.

1) The number of trials ‘ n’ is finite.
2) The trials are independent of each other.
3) The probability of success ‘ p’ is constant for each trial.
4) Each trial must result in a success or a failure.

The problems relating to tossing of coins or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probability distribution.

**3.1.4 Characteristics of Binomial Distribution:**

1. Binomial distribution is a discrete distribution in which the random variable X (the number of success) assumes the values 0,1, 2, ….n, where n is finite.
2. Mean = np, variance = npq and standard deviation σ= √npq ,
   Coefficient of skewness = q-p/√*npq*,
   Coefficient of kurtosis = 1 - 6pq /npq , clearly each of the probabilities is non-negative and sum of all probabilities is
   1 ( p< 1 , q < 1 and p + q =1, q = 1-p ).
3. The mode of the binomial distribution is that value of the variable which occurs with the largest probability. It may have either one or two modes.
4. If two independent random variables X and Y follow binomial distribution with parameter (n , p1) and (n , p2) respectively, then their sum (X+Y) also follows Binomial distribution with parameter (n1+ n2 , p).

5. If n independent trials are repeated N times, N sets of n trials are obtained and the expected frequency of x success is N(nC x p *x* q n-x ). The expected frequencies of 0,1,2… n success are the successive terms of the binomial distribution of N(q + p) n .

**Example 1:**

Comment on the following: “ The mean of a binomial distribution is 5 and its variance is 9”

**Solution:**

The parameters of the binomial distribution are n and p
We have mean-np = 5
Variance -npq = 9
Therefore,

q = npq /np = 9 /5
q = 9 /5 >1
Which is not admissible since q cannot exceed unity. Hence the given statement is wrong.

**Example 2:**
Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.
**Solution:**
Here number of trials, n = 8, p denotes the probability of getting a head.
Therefore,
p = 1 /2 and q = 1 /2

If the random variable X denotes the number of heads, then the probability of a success in n trials is given by
P(X = x) = nxcpxqn-x ,
*x= 0 , 1, 2, ..., n*
= 8C x$(1/2)^x$ x(1/2)8-x= 8Cx=(1/2)8
=1/2^8 8Cx
Probability of getting atleast six heads is given by

P(x >_6)=1/2^8(28+8+1)
$$=\frac{37}{256}.$$

### 3.2 POISSON DISTRIBUTION:

**3.2.0 Introduction**:

Poisson distribution was discovered by a French Mathematician-cum-Physicist Simeon Denis Poisson in 1837. Poisson distribution is also a discrete distribution. He derived it as a limiting case of Binomial distribution. For n-trials the binomial distribution is (q + p) ; the probability of x successes is given by n P(X=x) = nC p q x x n-x . If the number of trials n is very large and the probability of success ‘ p’ is very small so that the product np = m is non – negative and finite. The probability of x success is given by
P( X = x ) = {e^-m m^x/x for x = 0,1,2, …
0 ; otherwise

Here m is known as parameter of the distribution so that m >0.
Since number of trials is very large and the probability of success p is very small, it is clear that the event is a rare event. Therefore Poisson distribution relates to rare events.
**Note:**

1) e is given by e = 1 + 1 1! + 1 2! + 1 3! +….. = 2.71828
2) P(X=0) =
m0em 0! , 0! = 1 and 1! = 1
3) P(X=1) =
m1em 1!
Some examples of Poisson variates are :
1. The number of blinds born in a town in a particular year.
2. Number of mistakes committed in a typed page.
3. The number of students scoring very high marks in all subjects
4. The number of plane accidents in a particular week.
5. The number of defective screws in a box of 100, manufactured by a reputed company.
6. Number of suicides reported in a particular day.

**3.2.1 Conditions**:

Poisson distribution is the limiting case of binomial distribution under the following conditions:
1. The number of trials n is indefinitely large i.e., n ->infinity
2. The probability of success ‘ p’ for each trial is very small; i.e., p -> 0
3. np = m (say) is finite , m > 0

**3.2. 2 Characteristics of Poisson Distribution**:

The following are the characteristics of Poisson distribution
1. Discrete distribution: Poisson distribution is a discrete distribution like Binomial distribution, where the random variable assume as a countably infinite number of values 0,1,2 ….
2. The values of p and q: It is applied in situation where the probability of success p of an event is very small and that of failure q is very high almost equal to 1 and n is very large.
3. The parameter: The parameter of the Poisson distribution is m. If the value of m is known, all the probabilities of the Poisson distribution can be ascertained.
4. Values of Constant: Mean = m = variance; so that standard deviation = m Poisson distribution may have either one or two modes.
5. Additive Property: If X and Y are two independent Poisson distribution with parameter m 1 and m 2 respectively. Then (X+Y) also follows the Poisson distribution with parameter (m1 + m2)
6. As an approximation to binomial distribution: Poisson distribution can be taken as a limiting form of Binomial distribution when n is large and p is very small in such a way that product np = m remains constant.
7. Assumptions: The Poisson distribution is based on the following assumptions.
i) The occurrence or non- occurrence of an event does not influence the occurrence or non-occurrence of any other event.
ii) The probability of success for a short time interval or a small region of space is proportional to the length of the time interval or space as the case may be.

iii) The probability of the happening of more than one event is a very small interval is negligible.

**Example 8:**

Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year? [given that e - 2 = 0.13534]

Mean, x = np , n = 2000 and p = 1/ 1000
= 2000 x1/ 1000
m = 2

The Poisson distribution is

$$P(X=x) = e^{m} m^{x} /x!$$

$$P(X =5) = (0.13534) x32 /120 = 0.036$$

(Note: The values of e - m are given in Appendix )

**Example 9**:

In a Poisson distribution 3P(X=2) = P(X=4) Find the parameter ‘ m’ .

Solution:

Poisson distribution is given by $P(X=x) = {}^{x}emx /em\ x!$

Given that 3P(x=2) = P(x= 4)

$3.\,{}^{m}e\ m_2\ /2!$

m = ± 6

Since mean is always positive m = 6

**Fitting of Poisson distribution**:

The process of fitting of Poisson distribution for the probabilities of x = 0, 1,2,... success are given below :

i) First we have to calculate the mean = x Σfx/Σf = m

ii) The value of e - m is obtained from the table (see Appendix)

iii) By using the formula P(X=x) = mx e .m /x! Substituting x = 0,

P(0) = e - m Then f(0) = N P(0)

The other expected frequencies will be obtained by using the recurrence formula f(x+1) = m /$x_+$ 1 f(x) ; x = 0,1,2,…

**NORMAL DISTRIBUTION:**

**Introduction:**

In the preceding sections we have discussed the discrete distributions, the Binomial and Poisson distribution. In this section we deal with the most important continuous distribution, known as normal probability distribution or simply normal distribution. It is important for the reason that it plays a vital role in the theoretical and applied statistics. The normal distribution was first discovered by DeMoivre (English Mathematician) in 1733 as limiting case of binomial distribution. Later it was applied in natural and social science by Laplace (French Mathematician) in 1777. The normal distribution is also known as Gaussian distribution in honour of Karl Friedrich Gauss(1809).

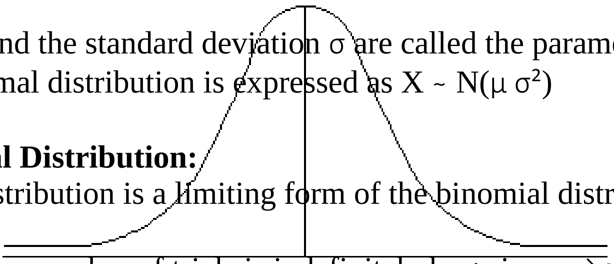
**Definition:**

A continuous random variable X is said to follow normal distribution with mean μ and standard deviation σ, if its probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2}\prod} e^{-1/2\,(x - \mu/\sigma)^2} \quad ; -\infty< x <\infty, -\infty<\mu<\infty, > 0$$

**Note:**

The mean μ and the standard deviation σ are called the parameters of Normal distribution. The normal distribution is expressed as X ~ N(μ σ²)



**Condition of Normal Distribution:**

(i) Normal distribution is a limiting form of the binomial distribution under the following conditions.

a) n, the number of trials is indefinitely large ie., n →∞ and

b) Neither p nor q is very small.

(ii) Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter m →∞

(iii) Constants of normal distribution are mean = μ, Variation = σ², Standard deviation = σ

**Normal probability curve:**

The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean (μ), bell – shaped and the two tails on the right and left sides of the mean extends to the infinity. The shape of the curve is shown in the following figure.

$-\infty$ $x = \mu$ $\infty$

**Properties of Normal Distribution:**

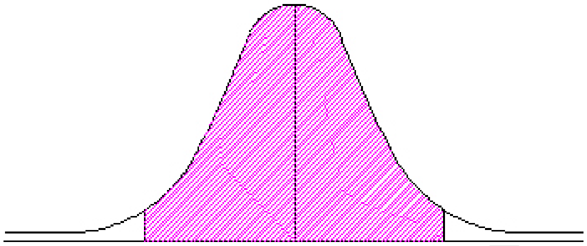
1. The normal curve is bell shaped and is symmetric at $x = \mu$.
2. Mean, median and mode of the distribution are coincide i.e., Mean = Median = Mode = $\mu$.
3. It has only one mode at $x = \mu$ (i.e., unimodal).
4. Since the curve is symmetrical, Skewness = $\beta_1 = 0$ and Kurtosis = $\beta_1 = 3$.
5. The points of inflection are at $x = \mu\pm\sigma$.
6. The maximum ordinate occurs at $x = \mu$ and its value is = $1/\sigma\sqrt{2}$.
7. The x axis is an asymptote to the curve (i.e. the curve continues to approach but never touches the x axis).
8. The first and third quartiles are equidistant from median.
9. The mean deviation about mean is $0.8\ \sigma$.
10. Quartile deviation = $0.6745\ \sigma$.
11. If X and Y are independent normal variates with mean $\mu_1$ and $\mu_2$, and variance $\sigma_1^2$ and $\sigma_2^2$ respectively then their sum (X+Y) is also a normal variate with mean ($\mu_1 + \mu_2$) and variance ($\sigma_1^2 + \sigma_2^2$).
12. Area property $P(\mu - \sigma\ x <\mu< + \sigma) = 0.6826$
    $P(\mu - 2\sigma\ x <\mu< + 2\sigma) = 0.9544$
    $P(\mu - 3\sigma\ x <\mu< + 3\sigma) = 0.9973$

**Standard Normal distribution:**

Let X be random variable which follows normal distribution with mean $\mu$ and variance $\sigma^2$. The standard normal variate is defined as $Z = X - \mu / \sigma$ which follows standard normal distribution with mean 0 and standard deviation 1 i.e., $Z \sim N(0,1)$. The standard normal distribution is given by $\phi(z) = 1/\sqrt{2\pi}\ e^{-\frac{1}{2}z^2}$; $-\infty< z <\infty$. The advantage of the above function is that it doesn't contain any parameter. This enable us to compute the area under the normal probability curve.

**Area properties of Normal curve:**

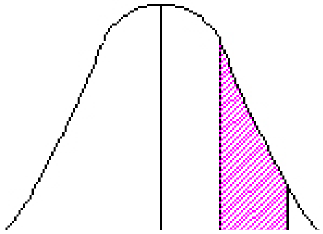
The total area under the normal probability curve is 1. The curve is also called standard probability curve. The area under the curve between the ordinates at x = a and x = b where a < b, represents the probabilities that x lies between x = a and x = b i.e., $P(a \leq X \leq b)$

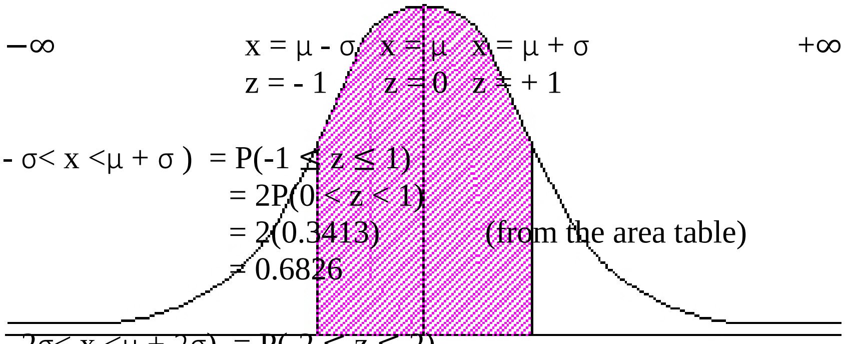


To find any probability value of x, we first standardize it by using Z = X - μ / σ, and use the area probability normal table. (given in the Appendix).

**For Example**:

The probability that the normal random variable x to lie in the interval (μ - σ ,μ + σ )is s given by





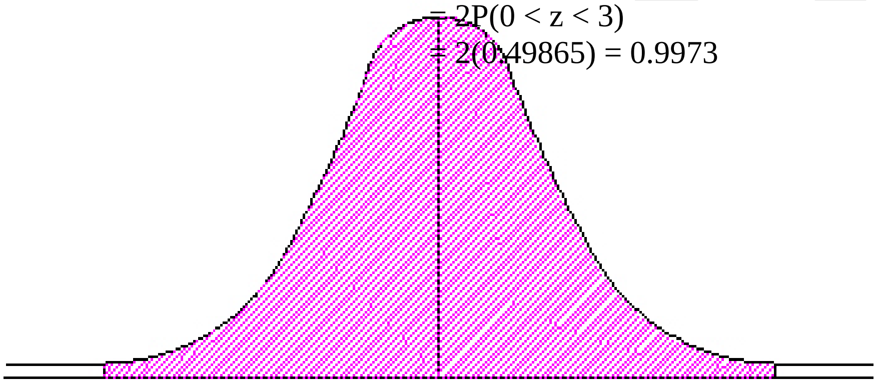
P(μ - σ< x <μ + σ ) = P(-1 ≤ z ≤ 1)
= 2P(0 < z < 1)
= 2(0.3413) (from the area table)
= 0.6826

P(μ - 2σ< x <μ + 2σ) = P(-2 ≤ z ≤ 2)
= 2P(0 < z < 2)
= 2(0.4772) = 0.9544

$-\infty$ $x = \mu - 2\sigma$ $x = \mu$ $x = \mu + 2\sigma$ $+\infty$

$z = -2$ $z = 0$ $z = +2$

$$P(\mu - 3\sigma < x < \mu + 3\sigma) = P(-3 \leq z \leq 3)$$
$$= 2P(0 < z < 3)$$
$$= 2(0.49865) = 0.9973$$



$-\infty$ $x = \mu - 3\sigma$ $x = \mu$ $x = \mu + 3\sigma$ $+\infty$

$z = - 3$ $z = 0$ $z = +3$

The probability that a normal variate x lies outside the range $\mu \pm \sigma 3$ is given by

$$P(|x - \mu| > \sigma 3\ ) = P(|z| > 3)$$
$$= 1 - P(\ -3 \leq z \leq 3)$$
$$= 1 - 0.\ 9773 = 0.0027$$

Thus, we expect that the values in a normal probability curve will lie between the range $\mu \pm 3\sigma$ though theoretically it range from - $\infty$to $+\infty$ .

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics and Research Methodology
UNIT III

**UNIT III**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 3 | The hypothesis under test is | simple hypothesis | alternative hypothesis | **null hypothesis** | complex hypothesis | **null hypothesis** |
| 2 | 3 | A test based on a test statistic is classified as | **randomized test** | non-randomized test | sequential test | Bayes test | **randomized test** |
| 3 | 3 | Student's t test is applicable in case of | **small samples** | for sample of size between 5 and 30 | large samples | sample size more than 100 | **small samples** |
| 4 | 3 | Reject $H_0$ when it is false is known as | Type I error | Type II error | **Correct decision** | wrong decision | **Correct decision** |
| 5 | 3 | Accept $H_0$ when it is true is known as | Type I error | Type II error | **Correct decision** | wrong decision | **Correct decision** |
| 6 | 3 | Accept $H_0$ when it is false is known as | Type I error | **Type II error** | Correct decision | wrong decision | **Type II error** |
| 7 | 3 | In a sample of 10 items the degree of freedom for student's t test is | 10 | **9** | 12 | 11 | **9** |
| 8 | 3 | If the sample size is less than 30 then those samples may be regarded as | large samples | **small samples** | parameter | attitude | **small samples** |
| 9 | 3 | The range of statistic-t is | -1 to +1 | **-∞ to +∞** | 0 to ∞ | 0 to 1 | **-∞ to +∞** |
| 10 | 3 | The distribution used to test goodness of fit is | F distribution | **$\chi^2$ distribution** | t distribution | Z distribution | **$\chi^2$ distribution** |
| 11 | 3 | Large sample theory is applicable when | **n>30** | n<30 | n=30 | n=10 | **n>30** |
| 12 | 3 | 95% of fiducial limits of population mean are | **A.M ± 1.96 S.E** | A.M ± 3.96 S.E | A.M ± 2.58 S.E | A.M ± 3.58 S.E | **A.M ± 1.96 S.E** |
| 13 | 3 | 99% of fiducial limits of population mean are | A.M ± 1.96 S.E | A.M ± 3.96 S.E | **A.M ± 2.58 S.E** | A.M ± 3.58 S.E | **A.M ± 2.58 S.E** |
| 14 | 3 | A part of the population selected for study is called as | statistic | **sample** | parameter | event | **sample** |
| 15 | 3 | Test statistic Z = | **$(X-\mu)/S.E(X)$** | $(X+\mu)/S.E(X)$ | $X\mu/S.E(X)$ | $(X^* \mu)/S.E(X)$ | **$(X-\mu)/S.E(X)$** |
| 16 | 3 | In students t test standard error of single mean = | **$S/\sqrt{n-1}$** | $S/\sqrt{n-2}$ | $S/\sqrt{n+1}$ | $S/\sqrt{n+2}$ | **$S/\sqrt{n-1}$** |
| 17 | 3 | Null hypothesis is denoted by | $H_1$ | **$H_0$** | $H_2$ | H | **$H_0$** |
| 18 | 3 | Alternative hypothesis is denoted by | $H_0$ | **$H_1$** | $H_2$ | H | **$H_1$** |
| 19 | 3 | The value of 5% level of significance is | 2.58 | **1.96** | 1.64 | 2.33 | **1.96** |
| 20 | 3 | The value of 1% level of significance is | **2.58** | 1.64 | 1.96 | 2.33 | **2.58** |
| 21 | 3 | If the sample size n=2, the students t-distribution reduces to | Normal distribution | F- distribution | **Cauchy distribution** | z – distribution | **Cauchy distribution** |
| 22 | 3 | If n, the sample size is larger than 30, the students t-distribution reduces to | **Normal distribution** | F- distribution | Cauchy distribution | Z- distribution | **Normal distribution** |
| 23 | 3 | The d.f. for students t based on a random sample of size n is | **n-1** | n | n-2 | (n-2)/2 | **n-1** |
| 24 | 3 | Degrees of freedom for statistic-chi-square in case of contingency table of order(3x3)is | 3 | **4** | 2 | 1 | 4 |
| 25 | 3 | Students 't' test is defined by the statistic | $t = x-\mu/\sqrt{}^{s}/n$ | $\sqrt{}^{t} = x-\mu/\sqrt{}^{s}/n$ | **$t = x-\mu/s/\sqrt{}^{n}$** | $t = x-\mu/s/n$ | **$t = x-\mu/s/\sqrt{}^{n}$** |
| 26 | 3 | Chi-square variate has _______ degree of freedom | 2 | **1** | 5 | 7 | **1** |
| 27 | 3 | Degree of freedom is denoted by | Φ | **ν** | $\chi^2$ | fi | **ν** |
| 28 | 3 | Students t- distribution, the name was given by | Arthur Henshalme | Beniamin | **William S.Cosset** | Karl Pearson | **William S.Cosset** |
| 29 | 3 | Chi-Square test was first used in testing statistical hypothesis by | **Karl Pearson** | Robert mayer | A.Fischer | Spearman | **Karl Pearson** |
| 30 | 3 | While testing significance of the difference of two sample means in case of small samples, the degree of freedom is calculated by | $V=n_1+n_2$ | $V=n_1+n_2-1$ | **$V=n_1+n_2-2$** | $V=n_1+n_2+2$ | **$V=n_1+n_2-2$** |
| 31 | 3 | Analysis of variance utilizes | **F- test** | Chi-square test | Z-test | t-test | **F- test** |
| 32 | 3 | Analysis of variance technique originated in | **Agricultural research** | Industrial research | Biological research | Business research | **Agricultural research** |
| 33 | 3 | Analysis of variance technique was developed by | Gosset | **R.A.Fisher** | Laplace | Karl Pearson | **R.A.Fisher** |
| 34 | 3 | Design of experiment was introduced by | **R.A.Fisher** | Spearman | Karl Pearson | Lorenz | **R.A.Fisher** |
| 35 | 3 | The variation within samples is known as | error | **block** | degrees of freedom | local control | **block** |
| 36 | 3 | CRD's statistical analysis is same as | two way classification | randomized block design | **One way classification** | completely randomized design | **One way classification** |
| 37 | 3 | If in a RBD having 5 treatments and 4 replications, a treatment is added, the increase in error difference will be | 1 | 2 | **3** | 4 | **3** |
| 38 | 3 | In a RBD with 4 blocks and 5 treatments having one missing value, the error difference will be | 12 | **11** | 10 | 9 | **11** |
| 39 | 3 | The ratio of the number of replications required in CRD and RBD for the same amount of information is | 6:04 | **10:06** | 10:08 | 6:10 | **10:06** |
| 40 | 3 | Error sum of squares in RBD as compared to CRD using the same material is: | more | **less** | equal | not comparable | **less** |
| 41 | 3 | In the analysis of data of a RBD with b block and v treatments, the error d.f. are: | b(v-1) | v(b-1) | **(b-1) (v-1)** | A=0 | **(b-1) (v-1)** |
| 42 | 3 | A randomized block design (RBD) has: | **two way classification** | one way classification | three way classification | no classification | **two way classification** |
| 43 | 3 | A CRD is also known as: | unsystematic design | **non-restrictional design** | single block design | restrictional design | **non-restrictional design** |
| 44 | 3 | CRD's are mostly used in | **field experiments** | experiments on animals | pot experiments | mathematical calculation | **field experiments** |
| 45 | 3 | In a CRD with 't' treatments and 'n' experimental units, error d.f. is equal to | **n-t** | n-t-1 | n-t+1 | t-n | **n-t** |
| 46 | 3 | The formula for obtaining a missing value in RBD by minimizing the error mean square was given by | W.G.Cochran | T.Wishart | **F.Yates** | J.W.Tukey | **F.Yates** |
| 47 | 3 | The idea of preliminary test of significance was propounded by | W.G.Cochran | **A.E.Paull** | F.Yates | J.W.Tukey | **A.E.Paull** |
| 48 | 3 | The range of F-variate is | **-∞ to ∞** | 0 to 1 | 0 to ∞ | -∞ to 0 | **-∞ to ∞** |
| 49 | 3 | A business firm's sales data classified on the basis of different salesman and sales in different region is example of | One way ANOVA | **Two way ANOVA** | T-test | Chi square test | **Two way ANOVA** |
| 50 | 3 | The ratio F is equal to | **$S_1^2/S_2^2$** | $S_1/S_2$ | $S_2/S_1$ | $S_2^2/S_1^2$ | **$S_1^2/S_2^2$** |
| 51 | 3 | If the calculated value of F is greater than the table value, the difference in sample means is | **Significant** | Not significant | Less significant | Insignificant | **Significant** |
| 52 | 3 | Coding method should be used specially | When given figures are simple or convenient | **When given figures are big or inconvenient** | When given figures are simple or inconvenient | When given figures are big or convenient | **When given figures are big or inconvenient** |
| 53 | 3 | If the calculated value of F is lesser than the table value, the difference in sample mean is | Significant | **Not significant** | Less significant | Insignificant | **Not significant** |
| 54 | 3 | In one way ANOVA the degrees of freedom is equal to | **n-1** | 1-n | v-n | X | **n-1** |
| 55 | 3 | We can determine one way ANOVA if there are | **Differences within one factor** | Differences between two factors | Similarities within one factor | Similarities between two factors | **Differences within one factor** |
| 56 | 3 | We can determine two way ANOVA if there are | Differences within one factor | **Differences between two factors** | Similarities within one factor | Similarities between two factors | **Differences between two factors** |
| 57 | 3 | In two way ANOVA, sum of squares due to error is | SSC-SSR | **SST-(SSC+SSR)** | SSC+(SST-SSR) | SST-SSR | **SST-(SSC+SSR)** |
| 58 | 3 | Mean square of residual variance is | SSC/c-1 | **SSE/(r-1)(c-1)** | SSR/(r-1)(c-1) | SSR/(r-1) | **SSE/(r-1)(c-1)** |
| 59 | 3 | The value of F in one way classification is | **between column variable/ within column variable** | within column variable/ between column variable | within row variable/ between column variable | between row variable/ within row variable | **between column variable/ within column variable** |
| 60 | 3 | In randomized block design the units are called as | **experimental units** | blocks | plots | block units | **experimental units** |

**UNIT V**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 5 | All items in any field of inquiry constitute a -------------------- | **Population** | sample | total items | infinite samples | **Population** |
| 2 | 5 | Population census conducted once in a ----------------- | century | **decade** | fortnight | two decade | **decade** |
| 3 | 5 | A definite plan for obtaining a sample from a given population is known as | Research design | **sample design** | clear plan | accurate plan | **sample design** |
| 4 | 5 | ---------------- is determined before data are collected | Research design | **sample design** | clear plan | accurate plan | **sample design** |
| 5 | 5 | The following is an example of finite universe | Number of stars in the sky | listeners of a specific radio programme | Throwing of a dice | **the population of a city** | **the population of a city** |
| 6 | 5 | The following are infinite universe except | Number of stars in the sky | listeners of a specific radio programme | Throwing of a dice | **the population of a city** | **the population of a city** |
| 7 | 5 | One of the following is the example of infinite universe | the population of a city | the number of workers in a factory | the number of students in a college | **listeners of a specific radio programme** | **listeners of a specific radio programme** |
| 8 | 5 | A --------------- results from errors in the sampling procedures | **systematic bias** | sampling error | designing error | research bias | **systematic bias** |
| 9 | 5 | When the size of the sample increases sampling error ---------- | **decreases** | increases | no change | zero | **decreases** |
| 10 | 5 | To study the economic status of a town or village the sampling design used is | probability sampling | **non-probability sampling** | based on each item | based on items at random | **non-probability sampling** |
| 11 | 5 | Probability sampling is also known as | quota sampling | judgement sampling | **random sampling** | purposive sampling | **random sampling** |
| 12 | 5 | One of the following is not a non-probability sampling | deliberate sampling | purposive sampling | judgement sampling | **chance sampling** | **chance sampling** |
| 13 | 5 | The probability of selecting sample of size 3 from a finite population of 6 elements is | **1/20** | 3/6 | 2/6 | 1/6 | **1/20** |
| 14 | 5 | Selecting every $i^{th}$ item on a list is known as | **systematic sampling** | area sampling | multi-stage sampling | cluster sampling | **systematic sampling** |
| 15 | 5 | If a 4% sample is desired, the sample selection will be | one in every 10th item | **one in every 25th item** | one in every 50th item | one in every 4th item | **one in every 25th item** |
| 16 | 5 | If a population from which a sample is to be drawn doesnot constitute a homogenous group the technique is known as | **Stratified sampling** | area sampling | multi-stage sampling | cluster sampling | **Stratified sampling** |
| 17 | 5 | In stratified sampling the different sub populations are called as | sub datas | **strata** | substitutes | random samples | **strata** |
| 18 | 5 | The most efficient and an optimal design of complex random sampling is | **Stratified sampling** | area sampling | multi-stage sampling | cluster sampling | **Stratified sampling** |
| 19 | 5 | Using proportional allocation, the sample sizes for different strata are 15,9 and 6 respectively which is in proportion to the sizes of the strat | 3000:4000:1000 | **4000:2400:1600** | 2400:1600:4000 | 1800:2200:4000 | **4000:2400:1600** |
| 20 | 5 | If clusters happens to be some geographic subdivisions it is known as | Stratified sampling | **area sampling** | multi-stage sampling | cluster sampling | **area sampling** |
| 21 | 5 | ___________ is a process of mapping aspects of a domain onto other aspects of a range according to some rule of correspondence | strata | **measurement** | sub strata | clusters | **measurement** |
| 22 | 5 | The data which are numerical in name only and donot share any properties of the numbers we deal in ordinary arithmetic is | **nominal data** | ordinal data | interval data | ratio data | **nominal data** |
| 23 | 5 | In those situations,when we can't do anything except set up inequalities, we refer to the data as | nominal data | **ordinal data** | interval data | ratio data | **ordinal data** |
| 24 | 5 | On Mohs' scale number 5&2 refers to apatite and gypsum.In this 5 > 2 refers to | **apatite is harder than gypsum** | apatite is softer than gypsum | gypsum is softer than apatite | both are equally hard | **apatite is harder than gypsum** |
| 25 | 5 | On Mohs' scale the numbers 6&9 refers to feldspar and sapphire respectively in this 6 < 9 refers to | feldspar is the most hardest | sapphire and feldspar are equally hard | **feldspar is softer than sapphire** | sapphire is softer than feldspar | **feldspar is softer than sapphire** |
| 26 | 5 | is simply a system of assigning number symbols to events in order to label them | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 27 | 5 | The assignment of number of basketball players in order to identify them is the usual example of | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 28 | 5 | In nominal scale the measure of central tendency used is | mean | median | **mode** | quartiles | **mode** |
| 29 | 5 | Generally used measure of dispersion for nominal scales is | standard deviation | mean deviation | quartile deviation | **no measure of dispersion** | **no measure of dispersion** |
| 30 | 5 | The most common test of statistical significance that can be utilized in nominal scales is | F-test | t-test | Z-test | **chi-square test** | **chi-square test** |
| 31 | 5 | For the measure of correlation ----------- can be worked out for nominal scales | **contingency coefficient** | scatter diagram | Karl-pearson's coefficient of correlation | graphs | **contingency coefficient** |
| 32 | 5 | is the least powerful level of measurement | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 33 | 5 | The lowest level of the ordered scale that is commonly used is the ----------------- | nominal scale | **ordinal scale** | interval scale | ratio scale | **ordinal scale** |
| 34 | 5 | A students rank in his graduation class involves the use of an ------------- | nominal scale | **ordinal scale** | interval scale | ratio scale | **ordinal scale** |
| 35 | 5 | Multiplication and division can only be used with this scale but not with other scales. | nominal scale | ordinal scale | interval scale | **ratio scale** | **ratio scale** |
| 36 | 5 | Measures of central tendency used in ratio scale is | median | mode | arithmetic mean | **geometric and harmonic mean** | **geometric and harmonic mean** |
| 37 | 5 | Researchers in physical sciences have the advantage to describe variables in | nominal scale | ordinal scale | interval scale | **ratio scale** | **ratio scale** |
| 38 | 5 | Researchers in behavioural sciences have the advantage to describe variables in | nominal scale | ordinal scale | **interval scale** | ratio scale | **interval scale** |
| 39 | 5 | The most precise type of scale is | nominal scale | ordinal scale | interval scale | **ratio scale** | **ratio scale** |
| 40 | 5 | refers to the extent to which a test measures what we actually wish to measure | **Validity** | Reliability | Practicality | speed | **Validity** |
| 41 | 5 | has to do with the accuracy and precision of a measurement procedure | validity | **reliability** | practicality | speed | **reliability** |
| 42 | 5 | is concerned with a wide range of factors of economy,convenience and interpretability | validity | reliability | **practicality** | speed | **practicality** |
| 43 | 5 | If the instrument contains a representative sample of the universe,the ---------- is good | Criterion related validity | **content validity** | construct validity | obstruct validity | **content validity** |
| 44 | 5 | enables the researcher to study the perceptual structure of a set of stimuli and the cognitive processes underlying the developm | rational scaling | nominal scaling | **multi-dimentional scaling** | interval scale | **multi-dimentional scaling** |
| 45 | 5 | __________ describes the procedures of assigning numbers to various degrees of opinion,attitude and other concepts | **scaling** | sampling | validity | reliability | **scaling** |
| 46 | 5 | Categorical scales are also known as | nominal scale | ordinal scale | interval scale | **rating scale** | **rating scale** |
| 47 | 5 | Comparative scales are also known as | nominal scale | **ranking scale** | interval scale | ratio scale | **ranking scale** |
| 48 | 5 | Classification without indicating order, distance or unique origin is | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 49 | 5 | _________ indicates magnitude relationship of 'more than' or 'less than' but indicate no distance or unique origin | nominal scale | **ordinal scale** | interval scale | ratio scale | **ordinal scale** |
| 50 | 5 | Which scales have both order and distance values but no unique origin | nominal scale | ordinal scale | **interval scale** | ratio scale | **interval scale** |
| 51 | 5 | The most widely used scale construction technique developed on ad hoc basis is | **arbitrary approach** | item analysis approach | consensus approach | factor scales | **arbitrary approach** |
| 52 | 5 | Qualitative description of a limited number of aspects of a thing or of traits of a person is | nominal scale | ordinal scale | interval scale | **rating scale** | **rating scale** |
| 53 | 5 | The scaling technique in the form of 'always-often-occasionally-rarely-never' is | nominal scale | ordinal scale | interval scale | **rating scale** | **rating scale** |
| 54 | 5 | Greater sensitivity of measurement is achieved in rating scale if there is | **more points on a scale** | less points on a scale | only two points | zero points | **more points on a scale** |
| 55 | 5 | If the respondents are either easy raters or hard raters, the type of error occurs is | error of central tendency | **error of leniency** | error of hallo effect | error of reliability | **error of leniency** |
| 56 | 5 | When raters are reluctant to give extreme judgments, the result is the | **error of central tendency** | error of leniency | error of hallo effect | error of reliability | **error of central tendency** |
| 57 | 5 | Composite standard method of paired comparision is given by | **J.P.Guilford** | E.L.Thorndike | M.D.Thurston | R.A.Fisher | **J.P.Guilford** |
| 58 | 5 | It is easy to construct Likert-type scale in comparison to Thurstone type because | only written examination is done | **can be performed without a panel of judges** | only oral questions are asked | It is performed with a panel of judges | **can be performed without a panel of judges** |
| 59 | 5 | An attempt to measure the psychological meanings of an object to an individual is | **semantic differential scale** | ordinal scale | interval scale | ratio scale | **semantic differential scale** |
| 60 | 5 | In which section of dissertation ,we can give our suggestions? | Introduction | Result | **Discussion** | Su mmary | **Discussion** |

## SYLLABUS

## UNIT IV

Sampling distribution and test of significance – concepts of sampling, testing of hypothesis, errors in hypothesis testing, standard errors and sampling distribution– Student's t test, F-test, Chi square test - goodness of fit. Analysis of variance – one way and two way classification. CRD, RBD Designs. Duncan's multiple range tests.

### Hypothesis:

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

### Hypothesis testing:

Hypothesis testing is to test some hypothesis about parent population from which the sample is drawn.

### Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

### Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by $H_0$ and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students".* Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

### Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis $H_0$ is called alternative hypothesis and is denoted by $H_1$ or $H_a$.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$$H_0 : \mu = 100$$

Then the alternative hypothesis could be any one of the statements.

$$H_1 : \mu \neq 100$$

$$(or)\, H_1 : \mu > 100$$

$$(or)\, H_1 : \mu < 100$$

**Errors in testing of hypothesis:**

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error
2) Type-II error

**1)Type-I error:** The type-I error is said to be committed if the null hypothesis ($H_0$) is true but our test rejects it.

**2)Type-II error:** The type-II error is said to be committed if the null hypothesis ($H_0$) is false but our test accepts it.

**Level of significance:**

The maximum probability of committing type-I error is called level of significance and is denoted by $\alpha$ .

$\alpha$ = P (Committing Type-I error)
= P ($H_0$ is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

**Power of the test:**

The probability of rejecting a false hypothesis is called power of the test and is denoted by$1-\beta$.

Power of the test =P ($H_0$ is rejected when it is false)
= 1- P ($H_0$ is accepted when it is false)
= 1- P (Committing Type-II error)
= $1-\beta$

- A test for which both $\alpha$ and $\beta$ are small and kept at minimum level is considered desirable.
- The only way to reduce both $\alpha$ and $\beta$ simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

### Critical region:

A statistic is used to test the hypothesis $H_0$. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which $H_0$ is rejected. It indicates that if the value of test statistic lies in this region, $H_0$ will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance$\alpha$. The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

### One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$against the alternative hypothesis $H_1 : \theta > \theta_0$(right tailed alternative) or $H_1 : \theta < \theta_0$(left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$------- left tailed test

### Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}c_{n}$ possible samples. If we calculate some particular statistic from each of the ${}^{N}c_{n}$

samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

### Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

$$\text{i.e.} \quad S.E\,(t)=\sqrt{Var(t)}$$

### Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z|=\left|\frac{t-E(t)}{S.E(t)}\right|>1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.
3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

### Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

$$\text{i.e. test statistic } Z=\frac{t-E(t)}{S.E(t)}$$

### Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0:\theta=\theta_0$.
2. Set up a alternative hypothesis i.e. $H_1:\theta\neq\theta_0$ or $H_1:\theta>\theta_0$ or $H_1:\theta<\theta_0$
3. Choose the level of significance i.e. $\alpha$.
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at $\alpha$% l.o.s i.e. $Z_\alpha$.
7. Compare the test statistic value with the tabulated value at $\alpha$% l.o.s. and make a decision whether to accept or to reject the null hypothesis.

## Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

**Assumption-1:** The random sampling distribution of the statistic is approximately normal.

**Assumption-2:** Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

### Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

where $\mu$ is population mean

$\mu_0$ is the value of $\mu$

Let $x_1, x_2, x_3, ................, x_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2$

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at $\alpha$ % l.o.s i.e. $Z_\alpha$

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

**Note:** if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s=\sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}$ .

**Large sample test for difference between two means:**

If two random samples of size $n_1$ and $n_2$ are drawn from two normal populations with means $\mu_1$ and $\mu_2$, variances $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1}^2/n_1\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2}^2/n_2\right)$

Therefore $\bar{x}_1-\bar{x}_2 \sim N\left(\mu_1-\mu_2, \frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}\right)$

For this test

The null hypothesis is $H_0:\mu_1=\mu_2 \Rightarrow \mu_1-\mu_2=0$

against the two sided alternative $H_1:\mu_1\neq\mu_2$

Now the test statistic $Z=\frac{t-E(t)}{S.E(t)}\sim N(0,1)$

$$=\frac{(\bar{x}_1-\bar{x}_2)-E(\bar{x}_1-\bar{x}_2)}{S.E(\bar{x}_1-\bar{x}_2)}\sim N(0,1)$$

$$\Rightarrow Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S.E(\bar{x}_1-\bar{x}_2)}\sim N(0,1)$$

$$\Rightarrow Z=\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}}\sim N(0,1)$$ [since $\mu_1-\mu_2$=0 from H0]

Now calculate $|Z|$

Find out the tabulated value of Z at $\alpha$% l.o.s i.e. $Z_\alpha$

If $|Z|>Z_\alpha$, reject the null hypothesis $H_0$

If $|Z|<Z_\alpha$, accept the null hypothesis $H_0$

**Note:** If $\sigma_1{}^2$ and $\sigma_2{}^2$ are unknown then we can consider $S_1{}^2$ and $S_2{}^2$ as the estimate value of $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively..

**Large sample test for single standard deviation (or) test for significance of standard deviation:**

Let $x_1, x_2, x_3, .................., x_n$ be a random sample of size n drawn from a normal population with mean $\mu$ and variance $\sigma^2$,

for large sample, sample standard deviation s follows a normal distribution with mean $\sigma$ and variance $\sigma^2 / 2n$ i.e. $s \sim N\left(\sigma, \sigma^2 / 2n\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$

against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{s - \sigma}{\sigma / \sqrt{2n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at $\alpha$% l.o.s i.e. $Z_\alpha$

If $|Z| > Z_\alpha$, reject the null hypothesis $H_0$

If $|Z| < Z_\alpha$, accept the null hypothesis $H_0$

**Large sample test for difference between two standard deviations:**

If two random samples of size $n_1$ and $n_2$ are drawn from two normal populations with means $\mu_1$ and $\mu_2$, variances $\sigma_1^{\ 2}$ and $\sigma_2^{\ 2}$ respectively

Let $s_1$ and $s_2$ be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, \sigma_1^{\ 2} / 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, \sigma_2^{\ 2} / 2n_2\right)$

Therefore $s_1 - s_2 \sim N\left(\sigma_1 - \sigma_2, \dfrac{\sigma_1^{\ 2}}{2n_1} + \dfrac{\sigma_2^{\ 2}}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$

against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{\sigma_1^{\ 2}}{2n_1} + \frac{\sigma_2^{\ 2}}{2n_2}}} \sim N(0,1) \qquad \text{[since } \sigma_1 - \sigma_2 = 0 \text{ from } H_0\text{]}$$

Now calculate $|Z|$

Find out the tabulated value of Z at $\alpha\%$ l.o.s i.e. $Z_\alpha$

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

### Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

Then $E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p$

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at $\alpha\%$ l.o.s i.e. $Z_\alpha$

If $|Z| > Z_\alpha$ , reject the null hypothesis $H_0$

If $|Z| < Z_\alpha$ , accept the null hypothesis $H_0$

**Large sample test for single proportion (or) test for significance of proportion:**

let $x_1$and $x_2$ be the number of persons processing a given attribute in a random sample of size $n_1$and $n_2$ then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$

$$S.E(\hat{p}_1) = \sqrt{\frac{p_1q_1}{n_1}} \text{ and } S.E(\hat{p}_2) = \sqrt{\frac{p_2q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When $p$ is not known $p$ can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at $\alpha$ % l.o.s i.e. $Z_\alpha$

If $|Z| > Z_\alpha$, reject the null hypothesis $H_0$

If $|Z| < Z_\alpha$, accept the null hypothesis $H_0$

- **As $\sigma$ is unknown,**

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

**Step 2: If $\mu_0$ falls into the above confidence intervals, then do *not* reject $H_0$. Otherwise, reject $H_0$.**

Example 1:

The average starting salary of a college graduate is $19000 according to government's report. The average salary of a random sample of 100 graduates is $18800. The standard error is 800.

(a) Is the government's report reliable as the level of significance is 0.05.

(b) Find the p-value and test the hypothesis in (a) with the level of significance $\alpha = 0.01$.

(c) The other report by some institute indicates that the average salary is $18900. Construct a 95% confidence interval and test if this report is reliable.

[solutions:]

(a)

$$H_0 : \mu = \mu_0 = 19000 \text{ vs. } H_a : \mu \neq \mu_0 = 19000,$$
$$n = 100, \bar{x} = 18800, s = 800, \alpha = 0.05$$

Then,

$$|z| = \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| = \left|\frac{18800 - 19000}{800/\sqrt{100}}\right| = |-2.5| = 2.5 > z_{\alpha/2} = z_{0.025} = 1.96.$$

Therefore, reject $H_0$.

(b)

$$\text{p-value} = P(|Z| > |z|) = P(|Z| > 2.5) = 2 \cdot P(Z > 2.5) = 0.0124 > 0.01$$

Therefore, *not* reject $H_0$.

(c)

$$H_0 : \mu = \mu_0 = 18900 \text{ vs } H_a : \mu \neq \mu_0 = 18900,$$

A 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 18800 \pm z_{0.025}\frac{800}{\sqrt{100}} = 18800 \pm 1.96 \cdot 80 = [18643.2, 18956.8].$$

Since $\mu_0 = 18900 \in [18643.2, 18956.8]$, Therefore, *not* reject $H_0$.

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let $\alpha = 0.05$. Please test the hypothesis

$$H_0 : u = 40 \ \ vs. \ \ H_a : u \neq 40.$$

based on

(a) classical hypothesis test
(b) p-value
(c) confidence interval.

[solution:]

$$\bar{x} = 38, \ s = 7, \ u_0 = 40, \ n = 49, \ z = \frac{\bar{x} - u_0}{s / \sqrt{n}} = \frac{38 - 40}{7 / \sqrt{49}} = -2.$$

(a)

$$|z| = 2 > 1.96 = z_{0.025}$$

we reject $H_0$.

(b)

$$p - value = P(|Z| > |z|) = P(|Z| > 2) = 2 * (1 - 0.9772) = 0.0456 < 0.05 = \alpha$$

we reject $H_0$.

(c)

$100 \times (1 - \alpha)\% = 95\%$ confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 38 \pm z_{0.025} \frac{7}{\sqrt{49}} = 38 \pm 1.96 = [36.04, 39.96].$$

Since $40 \notin [36.04, 39.96]$, we reject $H_0$.

### *Hypothesis Testing for the Mean (Small Samples)*

For samples of size less than 30 and when $\sigma$ is unknown, if the population has a normal, or nearly normal, distribution, the *t*-distribution is used to test for the mean $\mu$.

| **Using the t-Test for a Mean $\mu$ when the sample is small** | | |
|---|---|---|
| **Procedure** | **Equations** | **Example 4** |

| | | |
|---|---|---|
| State the claim mathematically and verbally. Identify the null and alternative hypotheses | State $H_0$ and $H_a$ | $H_0 : \mu \geq 16500$<br>$H_a : \mu < 16500$<br>$n = 14, \bar{x} = 15700, s = 1250$ |
| Specify the level of significance | Specify $\alpha$ | $\alpha = 0.05$ |
| Identify the degrees of freedom and sketch the sampling distribution | $d.f = n - 1$ | $d.f. = 13$ |
| Determine any critical values. If test is left tailed, use One tail, $\alpha$ column with a negative sign. If test is right tailed, use One tail, $\alpha$ column with a positive sign. If test is two tailed, use Two tails, $\alpha$ column with a negative and positive sign. | Table 5 (*t*-distribution) in appendix B | The test is left-tailed. Since test is left tailed and $d.f = 13$, the critical value is $t_0 = -1.771$ |
| Determine the rejection regions. | The rejection region is $t < t_0$ | The rejection region is $t < -1.771$ |
| Find the standardized test statistic | $t = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} \approx \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ | $t = \dfrac{15700 - 16500}{1250/\sqrt{14}} \approx -2.39$ |
| Make a decision to reject or fail to reject the null hypothesis | If t is in the rejection region, reject $H_0$,<br>Otherwise do not reject $H_0$ | Since $-2.39 < -1.771$,<br>reject $H_0$ |
| Interpret the decision in the context of the original claim. | | Reject claim that mean is at least 16500. |

**Chi-Square Tests and the F-Distribution**

### *Goodness of Fit*

DEFINITION A **chi-square goodness-of-fit test is** used to test whether a frequency distribution fits an expected distribution.

The test is used in a multinomial experiment to determine whether the number of results in each category fits the null hypothesis:

$H_0$: The distribution fits the proposed proportions

$H_1$: The distribution differs from the claimed distribution.

To calculate the test statistic for the chi-square goodness-of-fit test, you can use observed frequencies and expected frequencies.

DEFINITION The **observed frequency O** of a category is the frequency for the category observed in the sample data.

The **expected frequency E** of a category is the calculated frequency for the category. Expected frequencies are obtained assuming the specified (or hypothesized) distribution. The expected frequency for the *i*th category is

$$E_i = np_i$$

where *n* is the number of trials (the sample size) and $p_i$ is the assumed probability of the *i*th category.

The Chi-square Goodness of Fit Test: The sampling distribution for the goodness-of-fit test is a chi-square distribution with $k-1$ degrees of freedom where k is the number of categories. The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequency of each category and *E* represents the expected frequency of each category. To use the chi-square goodness of fit test, *the following must be true*.

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be $\geq 5$.

| **Performing the Chi-Square Goodness-of-Fit Test (p 496)** | | |
|---|---|---|
| Procedure | Equations | Example **(p 497)** |
| Identify the claim. State the null and alternative hypothesis. | State $H_0$ and $H_1$ | $H_0$:<br>Classical 4%<br>Country 36% |

| | | |
|---|---|---|
| | | Gospel 11%<br>Oldies 2%<br>Pop 18%<br>Rock 29% |
| Specify the significance level | Specify $\alpha$ | $\alpha = 0.01$ |
| Determine the degrees of freedom | d.f. = #categories - 1 | $d.f. = 6 - 1 = 5$ |
| Find the critical value | $\chi^2_\alpha$: Obtain from Table 6 Appendix B | $\varphi^2_{0.01}(d.f = 5) = 15.086$ |
| Identify the rejection region | $\chi^2 \geq \chi^2_\alpha$ | $\chi^2 \geq 15.086$ |
| Calculate the test statistic | $\chi^2 = \sum \frac{(O-E)^2}{E}$ | Survey results, n = 500<br>Classical O= 8 E = .04*500 = 20<br>Country O = 210 E = .36*500 = 180<br>Gospel O = 7 E = .11*500 = 55<br>Oldies O = 10 E = .02*500 = 10<br>Pop O = 75 E = .18*500 = 90<br>Rock O= 125 E = .29*500 = 145<br><br>Substituting $\chi^2 = 22.713$ |
| Make the decision to reject or fail to reject the null hypothesis | Reject if $\chi^2$ is in the rejection region<br>Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$ | Since 22.713 > 15.086 we reject the null hypothesis<br>Equivalently<br>$P(X \geq 22.713) < 0.01$so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.) |
| Interpret the decision in the context of the original claim | | Music preferences differ from the radio station's claim. |

### *Using Minitab to perform the Chi-Square Goodness-of-Fit Test (Manual p 237)*

The data from the example above (Example 2 p 497) will be used.

Enter Three columns: Music Type: Classical, etc, Observed: 8 etc, Distribution 0.04, etc. (Note the names of the columns 'Music Types', 'Observed' and 'Distribution' are entered in the gray row at the top.)

Select **Calc->Calculator**, **Store the results in** C4, and calculate the **Expression** C3*500, click **OK**, Name C4 ‘Expected’ since it now contains the expected frequencies

| Music Type | Observed | Distribution | Expected |
|---|---|---|---|
| Classical | 8 | 0.04 | 20 |
| Country | 210 | 0.36 | 180 |
| Gospel | 72 | 0.11 | 55 |
| Oldies | 10 | 0.02 | 10 |
| Pop | 75 | 0.18 | 90 |
| Rock | 125 | 0.29 | 145 |

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: Click **Calc->Calculator**. **Store the results in** C5 and calculate the **Expression** (C2-C4)**2/C4. Click on **OK** and C5 should contain the calculated values.

| |
|---|
| 7.2000 |
| 5.0000 |
| 41.8909 |
| 0.0000 |
| 2.5000 |
| 2.7586 |

Next add up the values in C5 and the sum is the test statistic as follows: Click on **Calc->Column Statistics**. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

**Sum of C5**

```
Sum of C5 = 22.7132
```

Next calculate the P-value: Click on **Calc->Probability Distributions->Chi-square**. Select

**Cumulative Probability** and enter 5 **Degrees of Freedom** Enter the value of the test statistic `22.7132` for the **Input Constant**. Click **OK**.

The following is displayed on the Session Window.

**Cumulative Distribution Function**

```
Chi-Square with 5 DF

      x  P( X <= x )

22.7132     0.999617
```

P(X≤`22.7132`) = `0.999617` So the P-value = 1 – 0.999617 = 0.000383. This is less that α = 0.01 so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

### *Chi-Square with M&M's*

| |
|---|
| $H_0$: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24% |
| Significance level: α = 0.05 |
| Degrees of freedom: number of categories – 1 = 5 |
| Critical Value: $\chi^2{}_{0.05}(d.f.=5)=11.071$ |
| Rejection Region: $\chi^2 \geq 11.071$ |
| Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where *O* is the actual number of M&M's of each color in the bag and *E* is the proportions specified under $H_0$ times the total number. |
| Reject $H_0$ if the test statistic is greater than the critical value (1.145) |

### *Section 10.2 Independence*

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c* contingency table** shows the observed frequencies for the two variables.

The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where $f_{ij}$ is the frequency that A equals $A_i$ and B equals $B_j$.

| | **A$_1$** | **A$_2$** | **A$_3$** | **A$_4$** | **A** |
|---|---|---|---|---|---|
| **B$_1$** | $f_{11}$ | $f_{12}$ | $f_{13}$ | $f_{14}$ | $f_{1.}$ |
| **B$_2$** | $f_{21}$ | $f_{22}$ | $f_{23}$ | $f_{24}$ | $f_{2.}$ |
| **B$_3$** | $f_{31}$ | $f_{32}$ | $f_{33}$ | $f_{34}$ | $f_{3.}$ |
| **B** | $f_{.1}$ | $f_{.2}$ | $f_{.3}$ | $f_{.4}$ | $f$ |

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum of\ row\ i) * (sum of\ column\ j)}{sample\ size}\ ($$

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

| | <= 39 | 40 - 49 | 50 - 59 | 60 - 69 | >= 70 | Total |
|---|---|---|---|---|---|---|
| Small/midsize | 42 | 69 | 108 | 60 | 21 | 300 |
| Large | 5 | 18 | 85 | 120 | 22 | 250 |
| Total | 47 | 87 | 193 | 180 | 43 | 550 |

| | <= 39 | 40 - 49 | 50 - 59 | 60 - 69 | >= 70 | Total |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| Small/midsize | $\frac{300*47}{550} \approx 25.64$ | $\frac{300*87}{550} \approx 47.45$ | $\frac{300*193}{550} \approx 105.27$ | $\frac{300*180}{550} \approx 98.18$ | $\frac{300*43}{550} \approx 23.45$ | 300 |
| Large | $\frac{250*47}{550} \approx 21.36$ | $\frac{250*87}{550} \approx 39.55$ | $\frac{250*193}{550} \approx 87.73$ | $\frac{250*180}{550} \approx 81.82$ | $\frac{250*43}{550} \approx 19.55$ | 250 |
| Total | 47 | 87 | 193 | 180 | 43 | 550 |

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be $\geq 5$

The sampling distribution for the test is a chi-square distribution with

$$(r-1)(c-1)$$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequencies and *E* represents the expected frequencies.

To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

| **Performing a Chi-Square Test for Independence (p 507)** | | |
|---|---|---|
| Procedure | Equations | Example2 **(p 507)** |
| Identify the claim. State the | State $H_0$ and $H_1$ | $H_0$: CEO's ages are |

| | | |
|---|---|---|
| null and alternative hypotheses. | | independent of company size<br>$H_1$: CEO's ages are dependent on company size. |
| Specify the level of significance | Specify $\alpha$ | $\alpha = 0.01$ |
| Determine the degrees of freedom | $d.f. = (r-1)(c-1)$ | $d.f. = (2-1)(5-1) = 4$ |
| Find the critical value. | $\chi^2_\alpha$ :Obtain from Table 6, Appendix B | $\chi^2_\alpha \geq 13.277$ |
| Identify the rejection region | $\chi^2 \geq \chi^2_\alpha$ | $\chi^2 \geq 13.277$ |
| Calculate the test statistic | $\chi^2 = \sum \frac{(O-E)^2}{E}$ | $\sum \frac{(O-E)^2}{E} \approx 77.9$<br>Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above |
| Make a decision to reject or fail to reject the null hypothesis | Reject if $\chi^2$ is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$ | Since 77.9 > 13.277 we reject the null hypothesis Equivalently<br>$P(X \geq 77.0) < \alpha$<br>so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.) |
| Interpret the decision in the context of the original claim | | CEO's ages and company size are dependent. |

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

   *Solution:*

   There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion $p$ = 5/20 = 0.25. Using the PROB-VALUE method the steps in this test are:

   1) $H_0$: $\pi = 0.5$ and $H_A$: $\pi$ ¹ 0.5
   2) We will use the $Z$-distribution
   3) We will use the 5%-level, thus $\alpha = 0.05$
   4) The test statistic is $z = (0.25 - 0.5)/\sqrt{0.25/20} = -2.24$
   5) Table A-4 shows that P(|$Z$| > 2.24) » 0.025.
   6) Because PROB-VALUE <$\alpha$, we reject $H_0$. We conclude $\pi$ is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

   *Solution:*

   Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

$$P(X \geq 8) = 0.1208 \quad + 0.0537 \quad + 0.0161 \quad + 0.0029 \quad + 0.002 \quad = 0.1937$$

Adopting the 5% uncertainty level, we see that PROB-VALUE >$\alpha$. Thus we fail to reject $H_0$. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

   *Solution:*

   (a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

   (b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference. We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

| High Density | Low Density | Sparsely Settled |
|---|---|---|
| 1.84 | 2.04 | 1.07 |

| | | |
|---|---|---|
| 3.06 | 2.28 | 2.31 |
| 3.62 | 4.01 | 0.91 |
| 4.91 | 1.86 | 3.28 |
| 3.49 | 1.42 | 1.31 |

*Solution:*

We will use the multi-sample Kruskal-Wallis test with an uncertainly level $\alpha$ = 0.1. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5} + \frac{39^2}{5} + \frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the $\chi$2 distribution with $3 - 1 = 2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

| | Distance (km) | | | Distance (km) | |
|---|---|---|---|---|---|
| Person | 1996 | 2006 | Person | 1996 | 2006 |
| 1 | 8.6 | 8.8 | 7 | 7.7 | 6.5 |
| 2 | 7.7 | 7.1 | 8 | 9.1 | 9 |
| 3 | 7.7 | 7.6 | 9 | 8 | 7.1 |
| 4 | 6.8 | 6.4 | 10 | 8.1 | 8.8 |
| 5 | 9.6 | 9.1 | 11 | 8.7 | 7.2 |
| 6 | 7.2 | 7.2 | 12 | 7.3 | 6.4 |

Has the length of the journey to work changed over the decade?

*Solution:*

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $H_0 : \eta = 0$ and $H_A : \eta \neq 0$. We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

S = {-,+,+,+,+,0,+,-,+,-,+,+}

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with π = 0.5. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the α = 10% level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

8. One hundred randomly sampled residents of a city subject to periodic flooding are classified according to whether they are on the floodplain of the major river bisecting the city or off the floodplain. These households are then surveyed to determine whether they currently have flood insurance of any kind. The survey results are as follows:

| | On the Floodplain | Off the Floodplain |
|---|---|---|
| Insured | 50 | 10 |
| No Insurance | 15 | 25 |

Test a relevant hypothesis.

*Solution:*

We will do a $\chi^2$ test for a relationship between insurance and house location. The null hypothesis is no relationship (independence). Augmenting the data with expected frequencies, we have:

| | On the Floodplain | Off the Floodplain |
|---|---|---|
| Insured | 50<br>(39) | 10<br>(21) |
| No Insurance | 15<br>(26) | 25<br>(14) |

The corresponding $\chi^2$ value is 22.16. Table A-8 shows that with 1 degree of freedom, $P(\chi^2 > 20)$ is zero to 3 decimal places. Thus for any reasonable level of uncertainty (any $\alpha < 0.0005$), we can reject the null hypothesis.

9. The occurrence of sunshine over a 30-day period was calculated as the percentage of time the sun was visible (i.e., not obscured by clouds). The daily percentages were:

| Day | Percentage of sunshine | Day | Percentage of sunshine | Day | Percentage of sunshine |
|---|---|---|---|---|---|
| 1 | 75 | 11 | 21 | 21 | 77 |
| 2 | 95 | 12 | 96 | 22 | 100 |
| 3 | 89 | 13 | 90 | 23 | 90 |
| 4 | 80 | 14 | 10 | 24 | 98 |
| 5 | 7 | 15 | 100 | 25 | 60 |
| 6 | 84 | 16 | 90 | 26 | 90 |
| 7 | 90 | 17 | 6 | 27 | 100 |
| 8 | 18 | 18 | 0 | 28 | 90 |
| 9 | 90 | 19 | 22 | 29 | 58 |
| 10 | 100 | 20 | 44 | 30 | 0 |

If we define a sunny day as one with over 50% sunshine, determine whether the pattern of occurrence of sunny days is random.

*Solution:*

For this we can use the number-of-runs test. Rather than calculate runs across two samples, here we will simply note if a day has 50% or more sunshine. The sample becomes

S={+,+,+,+,-,+,+,-,+,+,-,+,+,-,+,+,-,-,-,-,+,+,+,+,+,+,+,+,-}

We see that the sample consists of 12 runs. There are $n_x = 21$ sunny days, and $n_y = 9$ cloudy days. Because $n_x < 20$, we cannot use the normal approximation given in Table 10-5. Instead the probability is computed using combinatorial rules, and is approximately 0.4. This is far too large for rejection of the randomness hypothesis. We cannot conclude the pattern is non-random.

10. Test the normality of the DO data (a) using the Kolmogorov–Smirnov test with the ungrouped data of Table 2-4 and (b) using the $\chi^2$ test with $k$ = 6 classes of Table 2-6.

Solution:

(a) We will take the mean and standard deviation as known rather than estimated from the sample. Doing so results in calculated PROB-VALUES that are smaller than the true values (i.e., we are more likely to reject the null hypothesis). For the DO data the mean and standard deviation are 5.58 and 0.39 respectively. We sort the data, and then find the differences between the observed and expected cumulative distributions. The table below shows the results for a few of the 50 observations:

| $x_i$ | $S(x_i)$ | $F(x_i)$ | $\lvert S(x_i)-F(x_i)\rvert$ |
|---|---|---|---|
| 4.2 | 0.020 | 0.015 | 0.005 |
| 4.3 | 0.040 | 0.023 | 0.017 |
| 4.4 | 0.060 | 0.032 | 0.028 |

| | | | |
|---|---|---|---|
| … | … | … | … |
| 5.9 | 0.780 | 0.692 | 0.088 |
| … | … | ... | … |
| 6.7 | 0.960 | 0.960 | 0.000 |
| 6.8 | 0.980 | 0.972 | 0.008 |
| 6.9 | 1.000 | 0.981 | 0.019 |

The maximum difference is 0.088. Table A-9 shows that with 50 degrees of freedom, the corresponding PROB-VALUE is about 0.6. We obviously cannot reject the hypothesis of normality.

(b) Here we will take the mean and standard deviation as unknown, to be estimated from the sample. In other words, we estimate two parameters from the sample. In building the $\chi^2$ table, we combine the first two and the last two categories in Table 2-6 to ensure at least 2 expected frequencies per cell. This reduces the number of categories 4, as seen in the table below:

| Group | Minimum | Maximum | $O_j$ | $E_j$ | $(O_j-E_j)^2/E_j$ |
|---|---|---|---|---|---|
| 1 | 4.000 | 4.990 | 9 | 3.3 | 10.13 |
| 2 | 5.000 | 5.490 | 10 | 17.0 | 2.89 |
| 3 | 5.500 | 5.990 | 20 | 21.7 | 0.14 |
| 4 | 6.000 | 6.990 | 11 | 7.0 | 2.24 |

The observed Chi-square value is 15.4. With $k - p - 1 = 4 - 2 - 1 = 1$ degrees of freedom, Table A-8 shows that the PROB-VALUE is less than 0.0005. We therefore reject the null hypothesis.

Note that with only 4 classes, we can obtain only a rough idea of the distribution of DO. The 4 classes given in Table 2-6 do not yield a distribution that is at all similar to the normal distribution. In practice one would need many classes (and observations) for the $\chi^2$ test to be reliable.

## Analysis of Variance (ANOVA)

## I. Introduction

In Regression, the decomposition of the total sum of squares (SST) into the "explained" sum of squares (SSR) and the "unexplained" sum of squares (SSE) took place in the Analysis of Variance or ANOVA table. However, ANOVA also refers to a statistical technique used to test for diffferences between the means for several populations. While the procedure is related to regression, in ANOVA the independent variable(s) are qualitative rather than quantitative. In both regression and ANOVA the dependent variable is quantitative.

**Example 1:** As city manager, one of your responsibilities is purchasing. The city is looking to buy lightbulbs for the city's streetlights. Aware that some brands' lightbulbs might outlive other brands' lightbulbs, you decide to conduct an experiment. Seven lightbulbs each are purchased from four brands (GE, Dot, West, and a generic) and placed in streetlights. The lifetime of each of the 28 lightbulbs is then recorded in the file "**Lightbulbs**."

In this example, the lifetime of a lightbulb, in thousands of hours, is the quantitative dependent variable of interest. The company marketing the lightbulb, i.e., the brand-name, is the qualitative independent variable. The variable "brand name" has four possible values (or four "levels" in the terminology of ANOVA). The letter "*k*" will be used for the number of "levels" of the independent variable or "factor". Here, $k = 4$ for the four brands being tested. We say, "the factor *brand-name* has four levels: GE, Dot, West, and generic."

The "populations" referred to in these notes are simply the different levels of the factor. So that, in this example, we are interested in whether the mean lifetimes for the four populations of lightbulbs differ. Since we cannot know with certainty, however, the true mean lifetime of all lightbulbs carrying a certain brand-name, we rely upon statistics to determine if the differences observed *between* samples drawn from the four brands are statistically significant. (Non-significant differences are those that can plausibly be attributed to chance, i.e., sample-to-sample, variation alone.)

## II. The (one-way) ANOVA Model

In order to perform tests of statistical significance, a model is assumed. The model used in ANOVA is similar in many respects to the model employed in regression. In fact, You may find it useful in these notes to make analogies between the model and formulas in ANOVA and the the corresponding model and formulas in regression. In the model below, recall that the dependent (or **response**) variable is quantitative as in regression, but the independent (or **factor**) variable is now qualitative. We begin with a model in which a single independent variable is used to describe the dependent variable. This **One-Way** ANOVA is analogous to simple regression.

**Terminolgy**

Although regression and analysis of variance are closely related, historically they developed separately. As a result they each adopted their own terminolgy. Unfortunately, this often means that similar things are referred to differently in the two procedures. Below is a list of some of the names used in ANOVA and what they refer to.

- Response: the dependent variable
- Factor(s): the independent variable(s)
- Levels: the possible values of a factor
- Treatments: another name for levels in one-way ANOVA, but there will be a distinction between levels and treatments when we discuss two-way ANOVA later. The term treatments derives from medicine, where the different treatments were the drugs or procedures being tested on patients, and agriculture, where the treatments were the different fertilizers or pesticides being tested on crops.
- The □ $_i$ are called the “factor-level means” or the “treatment means” in one-way ANOVA and represent the true mean value of the response variable for the i$^{th}$ population of treatments.

**Example 1 (continued):** For the lightbulb problem,

- the response is the lifetime of a particular lightbulb (in thousands of hours)
- the factor is the brand-name
- there are four levels or treatments: GE, Dot, West, and generic
- □□ $_{GE}$ is the mean lifetime of all GE bulbs, □□ $_{Dot}$ is the mean lifetime of all Dot bulbs, etc.

## IV. Hypothesis Test

As usual, we rely on a hypothesis test to determine if the sample means for the k samples drawn (one from each population) differ enough for the difference to be statistically significant (more than would likely occur due to random chance alone).

**Example 1 (continued):** It is important that the student understand why probability is important here. It is not unusual for one manufacturer to source a product marketed under many brand-names. For example, there are only a handfull of companies manufacturing denim jeans, but there are dozens of brand-name jeans available to the consumer. Similarly, not all lightbulbs are manufactured by the companies marketing them. It is not inconceivable, therefore, that all four brands of lightbulbs being tested by the city come off of the same assembly line. Yet, when tested, they would still yield four *different* sample means simply because of sample-to-sample variation. As city manager, you might be more than a little embarassed to discover that the brand that you've touted as superior to all others is actually different in name only! Lawsuits have been lost for far less.

**Hypotheses:**

- **$H_0$:**, i.e., all population means are equal. This is equivalent to saying that the *k* treatments have no differential effect upon the value of the response.
- **$H_A$:** At least two of the means differ. This says that different treatments produce different values of the response variable, on average.

**Test Statistic:**

**F =** $\frac{MSR}{MSE}$ **,** where***MSR***= the**<u>M</u>ean <u>S</u>quare** for **<u>T</u>reatments**,
and ***MSE*** = the **<u>M</u>ean <u>S</u>quare** for **<u>E</u>rror**

Note: What I'm calling ***MSR*** is often called *MST* in the literature. I've chosen to continue the use of *MSR* to highlight the similarity between regression and analysis of variance. MSE remains the same for both regression and analysis of variance. Formulas for the mean squares are given later in the notes.

**Logic:**

The analysis of variance uses the ratio of two ***variances***, *MSR* and *MSE*, to determine whether population ***means*** differ; hence the name "analysis of variance." Recall that one of the assumptions of the model is that the variance $\square^{\square}$ is the same for all populations. ***MSE*** provides an unbiased estimate of $\square^{\square}\square$ in ANOVA just as it does in regression (see regression notes). If the population means are all equal, which is the null hypothesis, it can be shown that *MSRalso* provides an unbiased estimate of $\square^{\square}$. If all of the population means are equal, therefore, we would expect **F** to be nearly equal to **1** since *MSR* and *MSE* should yield similar estimates of the variance $\square^{\square}$.

If some population means differ from others, however, *MSR* will tend to be bigger than *MSE* resulting in an **F - Ratio** substantially larger than **1**. Thus we reject **$H_0$** for large values of **F**, just as in regression.

## V. The ANOVA Table: Sums of Squares and Degrees of Freedom

### *A. Introduction*

At the heart of any analysis of variance is the ANOVA Table. The formulas for the sums of squares in ANOVA are simplified if the *k* samples are all of the same size $n_S$. In the interests of simplicity, therefore, the following discussion assumes that all *k* samples contain the same number of observations $n_S$.

### *B. Notation*

- The index i represents the i^th population or treatment, where i ranges from 1 to *k*
- The index j represents the j^th obsevation within a sample, where j ranges from 1 to $n_S$
- *n* is the total number of observations from all samples
- $y_{ij}$ is the value of the j^th observation in the i^th sample
- $\bar{y}_i$ is the mean of the i^th sample
- $\bar{\bar{y}}$ (read "y double-bar") is the mean of all *n* observations, $\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_S} y_{ij}$, or the mean of the sample means (hence the "double-bar" in the name), $\bar{\bar{y}} = \frac{\bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_k}{k}$

### *C. Sums of Squares*

**Sum of Squares for Treatments**, $SSR = n_S \sum_{i=1}^{k}\left(\bar{y}_i - \bar{\bar{y}}\right)^2$ is the "Between Group" variation, where the *k* "groups" or populations are represented by their sample means. If the sample means differ substantially then SST will be large.

**Sum of Squares for Error,** $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_S}\left(y_{ij} - \bar{y}_i\right)^2$ is the "Within Group" variation and represents the random or sample-to-sample variation

**TotalSum of Squares,** $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_S}\left(y_{ij} - \bar{\bar{y}}\right)^2$ is the total variation in the values of the response variables over all *k* samples. (Note: SST is the same as in regression)

### *D. Degrees of Freedom*

Degrees of freedom for treatments, $\mathbf{df_{SSR} = k\text{-}1}$. Rather than memorizing this formula, just imagine the number of dummy variables that you would have to create to conduct the equivalent analysis in regression. Since you always leave one possibility out in regression, you would need to create *k* – 1 dummy variables. Since the resulting regression model would have *k* – 1 independent variables, SSR (SST here) would have *k* – 1 degrees of freedom.

Degrees of freedom for error, $\mathbf{df_{SSE} = \mathit{n} - \mathit{k}}$.

Total degrees of freedom, $\mathbf{df_{SST} = \mathit{n} - 1}$. This is the same result obtained in regression.

Note: The two component degrees of freedom sum to the total degrees of freedom, just as in regression.

### E. *Mean Squares*

**<u>M</u>ean <u>S</u>quare for <u>T</u>reatments,** $MSR = \frac{SSR}{k - 1}$ is equivalent to MSR in regression

**<u>M</u>ean <u>S</u>quare for <u>E</u>rror,** $MSE = \frac{SSE}{n - k}$ is the same as MSE in regression. As in regression, MSE is an unbiased estimator of the common population variance □$^{□}$.

### F. *F – Ratio*

The statistic used to test the null hypothesis □$_{□□}$□ □ □ □$_{□□}$□ □ □ □ □ □ □ □$_{□}$□$_{k}$is□ $F = \frac{MSR}{MSE}$ . As mentioned earlier, if the null hypothesis is correct then this ratio should be close to one. If some of the sample means differ substantially, however, the ratio will be much larger. Large values of F therefore correspond to strong evidence for rejecting $H_0$. Statgraphics reports a *P*-value for the test.

### G. *Summary*

The ANOVA Table below summarizes some of the information in this section

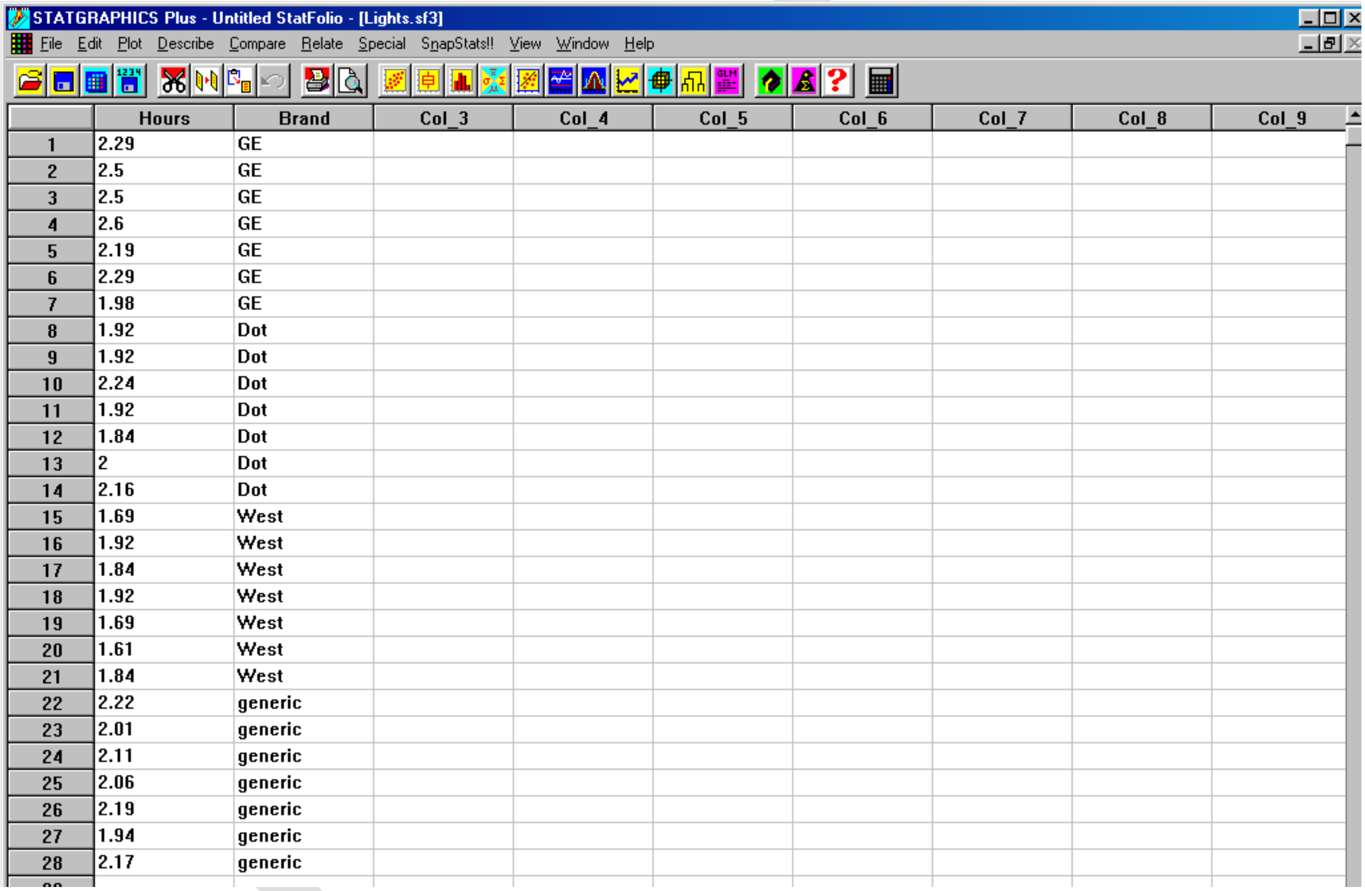
**ANOVA Table for One-Way Analysis of Variance**

| *Source* | *Sum of Squares* | *Df* | *Mean Square* | *F-Ratio* | *P-Value* |
|---|---|---|---|---|---|
| **Between** groups | ***SSR*** | *k* - 1 | ***MSR*=*SSR*/(***k*-1) | ***F = MSR/MSE*** | |
| **Within** groups | ***SSE*** | *n* - *k* | ***MSE = SSE***/(n-k) | | |
| **Total** (Corr.) | ***SST*** | *n* - 1 | | | |

### VI. Using Statgraphics

To perform a one-way analysis of variance in Statgraphics, follow *Compare* >*Analysis of Variance* >*One-Way ANOVA* and enter the response and factor into the dependent variable and factor fields, respectively.

**Example 1 (continued):** For the lightbulb problem, the spreadsheet might look like the one below. Notice that the qualitative factor Brand doesn't need to be numeric. Statgraphics will treat the factor in ANOVA as qualitative, so there is no need to recode it as a numeric variable. For the same reason there is no need to create dummy variables as in regression.



| | Hours | Brand | Col_3 | Col_4 | Col_5 | Col_6 | Col_7 | Col_8 | Col_9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.29 | GE | | | | | | | |
| 2 | 2.5 | GE | | | | | | | |
| 3 | 2.5 | GE | | | | | | | |
| 4 | 2.6 | GE | | | | | | | |
| 5 | 2.19 | GE | | | | | | | |
| 6 | 2.29 | GE | | | | | | | |
| 7 | 1.98 | GE | | | | | | | |
| 8 | 1.92 | Dot | | | | | | | |
| 9 | 1.92 | Dot | | | | | | | |
| 10 | 2.24 | Dot | | | | | | | |
| 11 | 1.92 | Dot | | | | | | | |
| 12 | 1.84 | Dot | | | | | | | |
| 13 | 2 | Dot | | | | | | | |
| 14 | 2.16 | Dot | | | | | | | |
| 15 | 1.69 | West | | | | | | | |
| 16 | 1.92 | West | | | | | | | |
| 17 | 1.84 | West | | | | | | | |
| 18 | 1.92 | West | | | | | | | |
| 19 | 1.69 | West | | | | | | | |
| 20 | 1.61 | West | | | | | | | |
| 21 | 1.84 | West | | | | | | | |
| 22 | 2.22 | generic | | | | | | | |
| 23 | 2.01 | generic | | | | | | | |
| 24 | 2.11 | generic | | | | | | | |
| 25 | 2.06 | generic | | | | | | | |
| 26 | 2.19 | generic | | | | | | | |
| 27 | 1.94 | generic | | | | | | | |
| 28 | 2.17 | generic | | | | | | | |

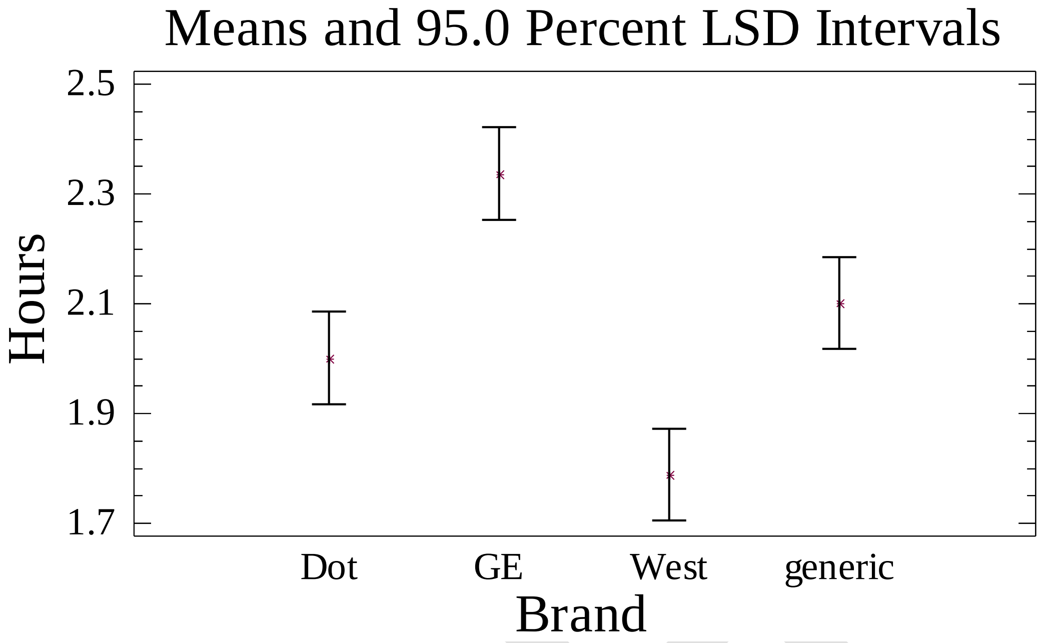
This leads to the ANOVA Table below. Looking at the *P*-value for the *F*-test, we conclude that there is strong evidence that at least two of the mean lifetimes differ.

```
ANOVA Table for Hours by Brand

                         Analysis of Variance
-------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio
-------------------------------------------------------------------
Between groups                1.08917      3     0.363057      15.62
Within groups                0.557714     24    0.0232381
-------------------------------------------------------------------
Total (Corr.)                 1.64689     27
```

Once the city manager has detected a difference in mean lifetimes, he/she would naturally wish to determine which brand's lightbulbs are superior. Statgraphics has a graphical option called a "Means Plot" which graphs 95% confidence intervals for the mean lifetimes of the four brands. If the 95% confidence intervals for two brands don't overlap then the city manager may conclude, at the 5% level of significance, that the true mean lifetimes for the two brands differ. If, on the other hand, the intervals *do* overlap the manager cannot draw a statistically significant conlusion at the 5% level of significance. (Remember, it's quite possible that the two brands' bulbs come off of the same assembly line, so don't try to force conclusions that can't be supported statistically!)

Below is the *Means Plot*. There is clearly evidence, at the 5% level of significance, that the GE bulbs last longer, on average, than bulbs from the other brands. Similarly, there is evidence, at the 5% level, that the West bulbs fail sooner, on average, than bulbs from the other brands. The sample differences between the Dot and generic bulbs, however, may be due to chance alone. (We don't actually *know* that Dot and generic bulbs are interchangeable, but the sample doesn't provide strong enough evidence to discount the possibilty.)



## VII. Two-Way ANOVA

When the effects of two qualitative factors upon a quantitative response variable are investigated, the procedure is called two-way ANOVA. Although a model exists for two-way analysis of variance, similar to the multiple regression model, it will not be covered in this class. Neither will we cover the details of the ANOVA Table. Nevertheless, there are some new considerations in two-way ANOVA stemming from the presence of the second factor in the model.

**Example 2:** The EPA (Environmental Protection Agency) tests public bodies of water for the presence of *coliform* bacteria. Aside from being potentially harmful to people in its own right, this bacteria tend to proliferate in polluted water, making the presence of *coliform* bacteria a surrogate for polution. Water samples are collected off public beaches, and the number of *coliform* bacterial per cc is determined. (See the file "**Bacteria**."

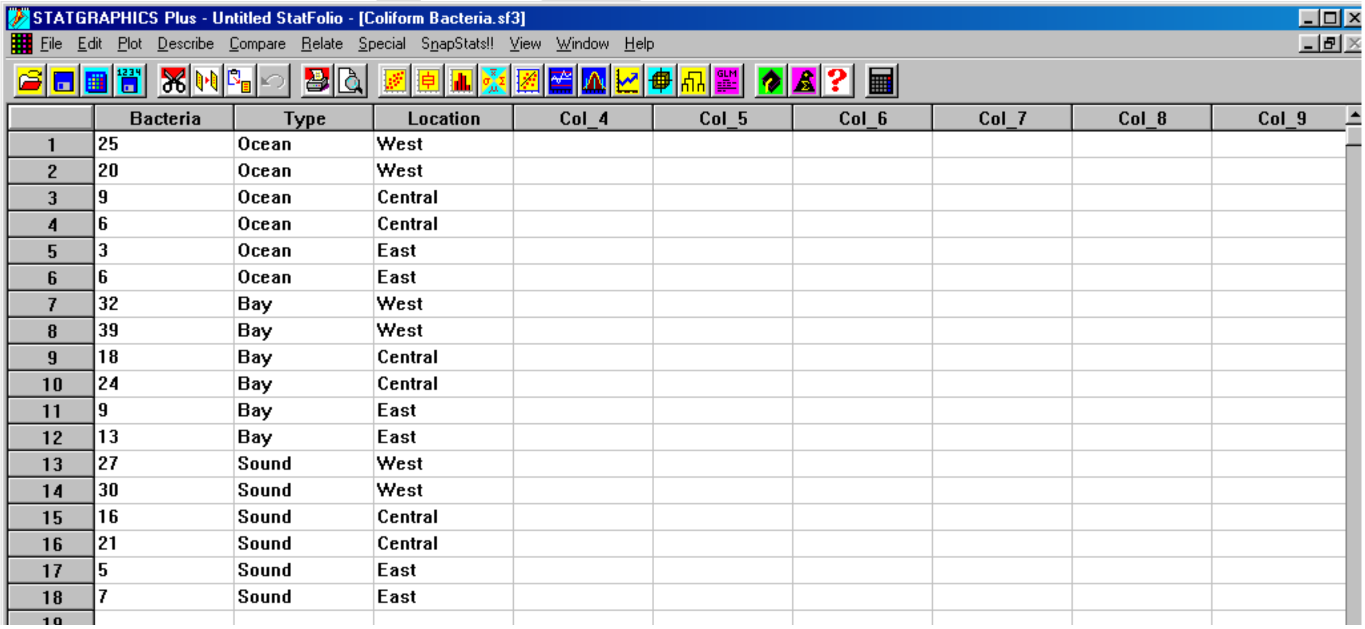
The EPA is interested in determining the factors that affect *coliform* bacterial formation in a particular county. The county has beaches adjacent to the ocean, a bay, and a sound. The EPA beleives that the amount of "flushing" a beach gets may affect the ability of polution to accumulate in the waters off the beach. The EPA also believes that the geographical location of the beach may be significant. (There could be several reasons for this: the climate may be different in different parts of the county, or the land-use may vary across the county, etc.)

As luck would have it, there is at least one beach for each combination of type (ocean, bay, sound) and location (west, central, east) within the county. Because of this, the EPA decides to sample a beach at each of the 9 possible combinations of type and location and conduct a two-way analysis of variance for *coliform* bacterial count. Two independent samples are taken at each beach to allow for an estimation of the natural variation in *coliform* bacterial count (this "repetition" is needed for the computation of MSE, which estimates the sample-to-sample variance in bacterial counts).

## VIII. Two-Way ANOVA Using Statgraphics

To perform a two-way analysis of variance in Statgraphics, follow *Compare >Analysis of Variance >Multifactor ANOVA* and enter the response and factors into the dependent variable and factor fields, respectively.

**Example 2 (continued):** Since data from such a study often appears in the form of a two-way table, with one factor as the row variable, the second as the column variable, and the observations as values in the row-by-column cells, it is important to remember that each variable must have its own column in the spreadsheet as in the example below. (This may require that you re-format the original spreadsheet prior to beginning the analysis.)



| | Bacteria | Type | Location | Col_4 | Col_5 | Col_6 | Col_7 | Col_8 | Col_9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 25 | Ocean | West | | | | | | |
| 2 | 20 | Ocean | West | | | | | | |
| 3 | 9 | Ocean | Central | | | | | | |
| 4 | 6 | Ocean | Central | | | | | | |
| 5 | 3 | Ocean | East | | | | | | |
| 6 | 6 | Ocean | East | | | | | | |
| 7 | 32 | Bay | West | | | | | | |
| 8 | 39 | Bay | West | | | | | | |
| 9 | 18 | Bay | Central | | | | | | |
| 10 | 24 | Bay | Central | | | | | | |
| 11 | 9 | Bay | East | | | | | | |
| 12 | 13 | Bay | East | | | | | | |
| 13 | 27 | Sound | West | | | | | | |
| 14 | 30 | Sound | West | | | | | | |
| 15 | 16 | Sound | Central | | | | | | |
| 16 | 21 | Sound | Central | | | | | | |
| 17 | 5 | Sound | East | | | | | | |
| 18 | 7 | Sound | East | | | | | | |
| 19 | | | | | | | | | |

The default ANOVA Table below has separate rows for the factors Type (called factor A) and Location (called factor B). A test of the significance of each factor is performed and the corresponding p-value displayed. It appears that both the type of beach and its location affect *coliform* bacterial count.

But does the effect of the beach type on bacteria count depend upon its location within the county? If the particular pairings of factor levels are important, the factors are said to "interact."

```
Source                     Sum of Squares     Df     Mean Square     F-Rat
---------------------------------------------------------------------------
MAIN EFFECTS
 A:Type                           364.778      2         182.389       16.
 B:Location                       1430.11      2         715.056       62.

RESIDUAL                          148.222     13         11.4017
---------------------------------------------------------------------------
TOTAL (CORRECTED)                 1943.11     17
```

Before interpreting the results in the ANOVA table above, we should consider the role that interaction plays. If the effect of beach type on bacteria formation depends on the location of the beach then it is better to investigate the *combinations* of the levels of the factors type and location for their affect on bacteria. It will come as no surprise to you that there is a hypothesis test for interactions.

**$H_0$:** The factors Type and Location do *not* interact.
**$H_A$:** The factors Typ and Location *do* interact

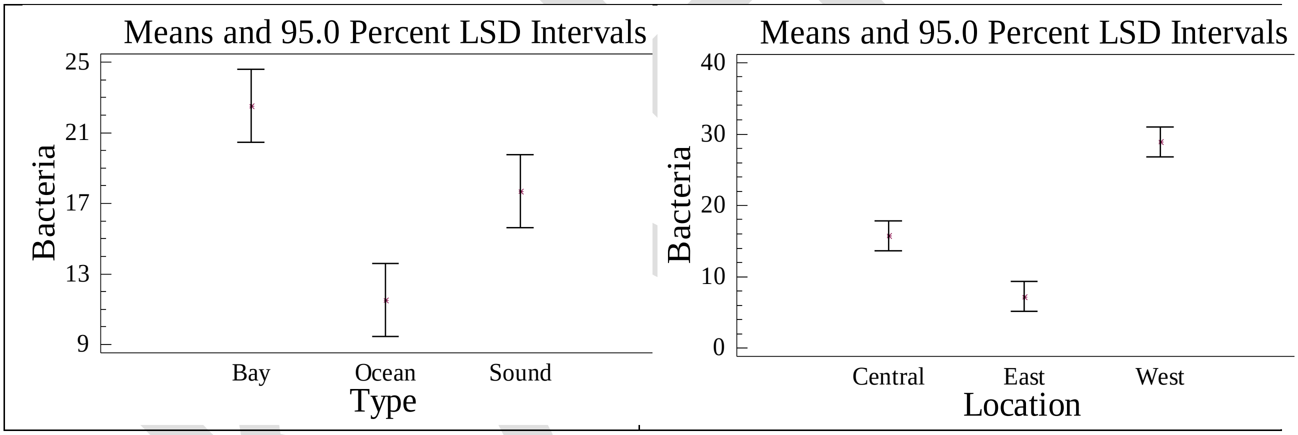
To check for interaction, use the right mouse button and *Analysis Options* and enter "2" for the *Maximum Order Interaction*. The resulting output for our example below shows a *P*-value of 0**.**3047 for the test for interactions. Thus the evidence for interaction is not particularly strong. The practical effect of discounting interaction is that we are able to return to the previous output (the one without interactions) and interpret the *P*-values for the factors Type and Location separately. Since the *P*-values for both factors are significant, we conclude that factors affect bacteria growth.

```
Source                     Sum of Squares     Df    Mean Square    F-Rat
--------------------------------------------------------------------------
MAIN EFFECTS
 A:Type                           364.778      2        182.389       18.
 B:Location                       1430.11      2        715.056       70.

INTERACTIONS
 AB                               57.2222      4        14.3056        1.

RESIDUAL                             91.0      9        10.1111
--------------------------------------------------------------------------
TOTAL (CORRECTED)                 1943.11     17
```

Having determined that the type of beach and the beach’s location are both significant, we next investigate the nature of the relationship between these factors and bacteria count. Once again we turn to the means plots under *Graphical Options*. Statgraphics dfaults to a means plot for the factor Type because this was the first factor entered in the *Input Dialog Box*. To get a means plot for the factor Location, use *Pane Options* to select it. The two means plots appear below.



Individually, these means plots are interpreted as in one-way ANOVA. There is evidence, at the 5% level of significance, that the mean bacteria count at ocean beaches is less than for other types, and that the mean count is highest at bay beaches. Similarly, the mean count is lowest in the east and greatest in the west, with all differences being statistically significant at the 5% level of significance. Furthermore, ***because interactions were judged not-significant***, we can add the main effects together and say that the least polluted beaches tend to be located in the east on the ocean, while the most polluted tend to be in the west on bays. We could not have added the separate (or main) effects in this way if there had been significant interact, for in that case the effect upon bacteria count at a particular type of beach (ocean, for example) may be very different for different locations.

**Example 3:** The last two examples are based on a marketing study. A new apple juice product was entering the marketplace. It had three distinct advantages relative to existing apple juices. First, it was not a concentrate and was therefore considered to be of higher “quality” than many similar products. Second, as one of the first juices packaged in cartons, it was cheaper than competing products. Third, partly because of the packaging, it was more convenient. The director of marketing for the company would like to know which advantage should be emphasized in advertisements. The director would also like to know whether local television or newspapers are better for sales.

Consequently, six cities with similar demographics are chosen, and a different combination of “Marketing Strategy “ and “Media” is tried in each. The unit sales of apple juice for the ten weeks immediately following the start of the ad campaigns are recorded for each city in the file **Apple Juice (two-way)**. The two-way table below describes the city assignments for the six possible combinations of levels for the two factors. Below the assignment table is the ANOVA Table for interactions.

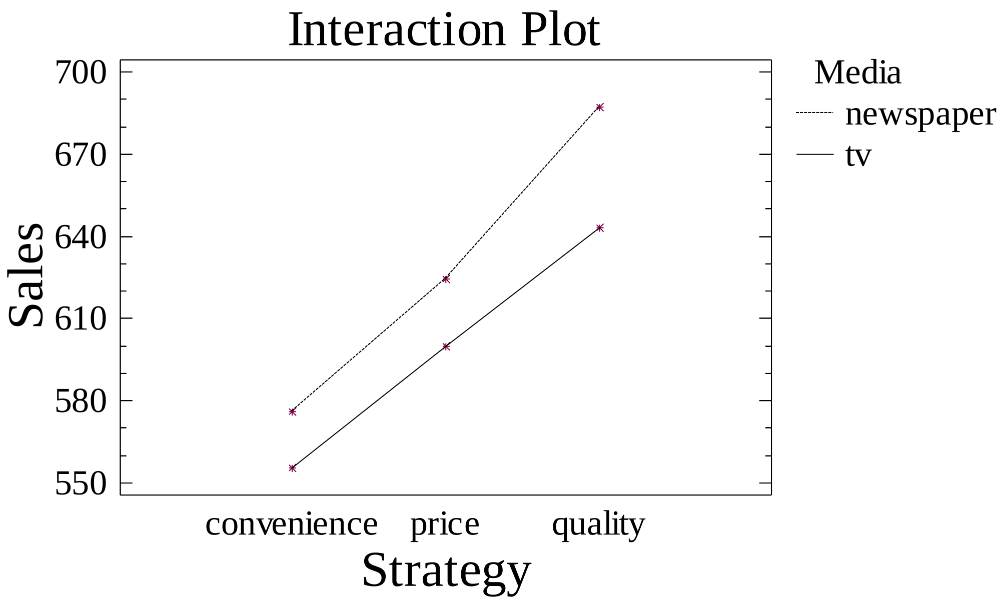
| | Convenience | Quality | Price |
|---|---|---|---|
| Local Television | City 1 | City 3 | City 5 |
| Newspaper | City 2 | City 4 | City 6 |

```
Source                    Sum of Squares     Df     Mean Square    F-Rat
------------------------------------------------------------------------
MAIN EFFECTS
 A:Strategy                      98838.6      2         49419.3       5.
 B:Media                         13172.0      1         13172.0       1.

INTERACTIONS
 AB                              1609.63      2         804.817       0.

RESIDUAL                        501137.0     54         9280.31
------------------------------------------------------------------------
TOTAL (CORRECTED)               614757.0     59
```

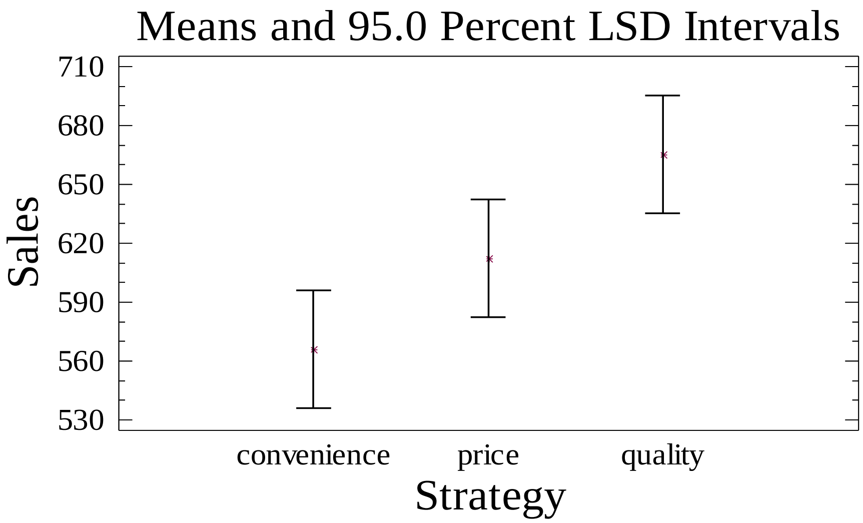
Interactions are not significant to the model (p-value equals 0.9171), a fact which is reinforced by looking at the *Interaction Plot* under *Graphical Options*. Note that the two curves are almost parallel, a sign that interactions are not significant.



Removing interactions, we obtain the ANOVA Table below, from which we conclude that the marketing strategy is significant, but the media used probably isn't. Since only marketing strategy apppears to affect sales, we'll restrict ourselves to the means plot for the factor Strategy below. Only the difference in mean sales when emphasizing quality versus emphasizing convenience is statistically significant at the 5% level of significance.

```
Source                   Sum of Squares     Df    Mean Square    F-Rat
----------------------------------------------------------------------
MAIN EFFECTS
 A:Strategy                     98838.6      2        49419.3        5.
 B:Media                        13172.0      1        13172.0        1.

RESIDUAL                       502746.0     56        8977.61
----------------------------------------------------------------------
TOTAL (CORRECTED)              614757.0     59
```



**Example 4:** This is just the apple juice problem revisited (see file "**Apple Juice – Remix**"). By a judicious rearrangement of sales figures, I've created a marketing study in which interactions are significant. (See the two-way table below for the new assignments.) The comparison of the interaction plots for this example and example 3 should help to clarify the role of interactions in the interpretation of ANOVA output. The small *P*-value of 0**.**0474 for the hypothesis test of interactions implies that certain combinations of marketing strategy and media are important to sales.

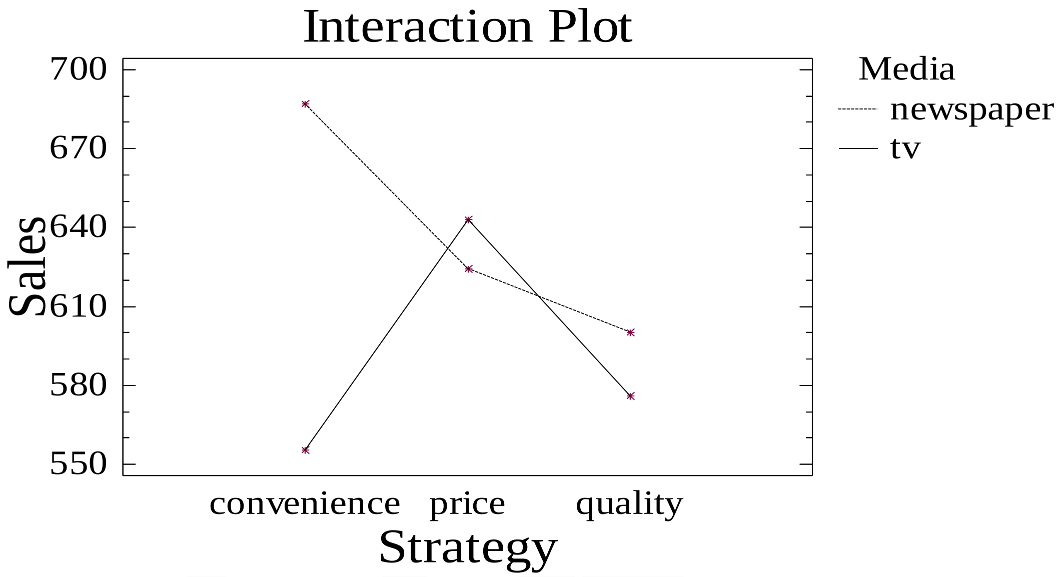
| | Convenience | Quality | Price |
|---|---|---|---|
| Local Television | City 1 | City 2 | City 3 |
| Newspaper | City 4 | City 5 | City 6 |

```
Source                    Sum of Squares     Df    Mean Square    F-Rat
-----------------------------------------------------------------------
MAIN EFFECTS
 A:Strategy                      22393.6      2        11196.8       1.
 B:Media                         31327.4      1        31327.4       3.

INTERACTIONS
 AB                              59899.3      2        29949.6       3.

RESIDUAL                        501137.0     54        9280.31
-----------------------------------------------------------------------
TOTAL (CORRECTED)               614757.0     59
```

Looking at he interaction plot, notice that emphasizing convenience lead to both the lowest and highest mean sales, depending upon whether local television or newspapers were used. Thus, it wouldn’t make sense to talk about the effect of emphasizing convenience without consideration of the media used, i.e., we should only interpret levels of the two factors taken together (the combinations). Therefore, we will not investigate the means plots for Strategy and Media. From the interaction plot, it appears that the most effective campaign would emphasize convenience in newspapers. The least effective combination is to emphasize convenince on local television. (Note: Since the interaction plot doesn’t display confidence intervals for the six possible combinations, we cannot attach a particular significance level to our conclusions as we could with the means plots.)



**Hypothesis testing**

Hypothesis testing
Up till now, we have dealt mainly with descriptive statistics, but as we’ve mentioned before, we also have inferential statistics at our disposal. These are statistics that allow us to make statements about one or more population based upon one or more samples that we’ve taken from the population. This allows us to test various hypotheses. There are many ways in which to do this, and we will cover only a few in this course.
In this chapter, we will go through the rationale behind hypothesis testing, and how we go about determining whether to reject, or fail to reject a null hypothesis, and in the process, decide whether our sample statistic is significantly different from some other measurement important to our experimental objectives.

Null vs alternate hypothesis

Hypothesis testing is a systematic model to summarise the evidence in order to decide between possible hypotheses.

Inferential stats are based upon the idea of a null hypothesis and an alternative hypothesis. The null hypothesis (written as HO:) is a statement written in such a way that there is no difference between two items. When we test the null hypothesis, we will determine a P value, which provides a numerical value for the likelihood that the null hypothesis is true. If it is unlikely that the null hypothesis is true, then we will reject our null hypothesis in favour of an alternate hypothesis (written as HA), and this states that the two items are not equal.

One can use the analogy of the criminal justice system. If one is arrested, the null hypothesis is that one is innocent of the crime. The state has the burden of showing that this null hypothesis is not likely to be true. If the state does that, then the judge rejects the null hypothesis and accepts the alternate hypothesis that you are guilty. Thus the state has to show that you are not innocent, in order to reject the null hypothesis. It is similar in statistics – your statistical test must show that the two items are different.

In the trial, if the null hypothesis is not rejected, your innocence has not been proven. It is just that the state failed to support your guilt. You are never proven innocent in terms of the trial: the state simply failed to show your guilt. The same is true in stats: if you fail to show that the two items are different, then you fail to reject the null hypothesis, but you have not proven that the null hypothesis is true. In other words, you can reject the null hypothesis, but it is incorrect to say you have "accepted" the null hypothesis – this implies that you have shown the null hypothesis to be correct, and that is not the case. Philosophically, it is important to understand these concepts.

Remember that we always state our hypotheses in terms of population parameters.

Example

When analysing data collected from a sample, we want to use that data to answer a biological question. We want to use our sample estimates to make some inferences about the biological population under study. Lets use an example the average height of students at UWC versus the average height of students at Wits. We want to use the estimates of our samples of these two populations in order to determine if there is a difference in height between students attending the two different universities.

For our null hypothesis, we will state that with respect to the parameter of the average height, there is no difference between the two groups.

When we examine our sample estimates (our form of descriptive statistics), we see that the two groups are different. But does this mean that the populations are different? There are two possibilities: the first is that the populations are different, and this is why their estimates are different. In other words, our null hypothesis is wrong and should be rejected. The second possibility is that the populations are the same, and the difference seen in the samples is just due to random error. In other words, our initial assumption (the null hypothesis) is correct, and so we fail to reject the null hypothesis. And so we have to decide which of the two possibilities is the correct one.

Sample difference

We must now ask: how much difference is there in our sample?

We need to quantify the difference. Inferential statistics are numbers that quantify differences.

We will ask questions like: is the difference big or small? A small difference could happen just by random sampling error, so if the difference is small we will assume the second of the possibilities we mentioned (ie that the populations are the same). A big difference is unlikely to occur just by chance, so if the difference is big, then we will assume the first of the possibilities (ie that the populations are different).

Alpha possibility

To determine if we have a small or big difference, we ask what is the probability of obtaining this much difference just by chance if we have sampled populations that are not different (ie, if our null hypothesis is correct)? This probability is called “alpha (α) probability”.

A small difference has a large probability (>0.05) of occurring. A big difference has a small probability (≤0.05) of occurring.

Since the inferential statistic quantifies the difference, we must determine the probability of finding the particular value of the statistic. The sampling distribution of the statistic allows us to determine probability. If the alpha (α) probability of the statistic is > 0.05, then we fail to reject the null hypothesis. If the alpha (α) probability of the statistic is ≤0.05, then we reject the null hypothesis.

Type I, II errors

When we reject or fail to reject a null hypothesis, we hope we are making the right decision. However, there is always some probability of us being wrong.

There are two possible ways in which we could be wrong:

We might reject a null hypothesis that we should have rejected – in other words, we concluded that there is a difference when there really isn’t. Statisticians call this a Type I (one) error.

We might also fail to reject a null hypothesis that we should have rejected – in other words, we failed to find a difference that actually does exist. Statisticians call this a Type II (two) error. When we fail to reject a null hypothesis, the probability that we have committed a Type II error is called the beta (β) probability. The ability of a statistical test to avoid making a Type II error is called the power of a test. Power therefore refers to how well a test can detect a difference when it actually exists. A powerful test is one that can detect small differences.

When we make scientific conclusions, we want to have both α and β as small as possible. They are inversely related – in other words, as the one goes up, the other goes down. Statisticians have shown both theoretically and empirically, that you can minimize both α and β by using a value of 0.05. If you use a smaller α, the β goes up too high. This is why statisticians generally recommend that null hypotheses be rejected at the α = 0.05 value. Although it might seem like an arbitrary value to us biologists, there are actually good mathematical reasons for using 0.05.

The only way to simultaneously decrease both α and β is to increase your sample size.

Reasoning of hypothesis testing

These are the steps we generally follow when hypothesis testing. We make a statement in both the null and sometimes also alternative hypothesis form, about a parameter. Then we select the experimental units that comprise our sample, and gather data that allow us to calculate a sample parameter. We use these parameters in order to decide whether or not to reject the null hypothesis. In order to do that though, we need to know what probability level we want to use, and we then use our chosen probability level in order to decide whether or not to reject the null hypothesis.

Hypothesis stating

When we state our hypotheses, we need to decide whether we are going to be applying a one-sided or two-sided test, and then state our hypotheses accordingly.

Setting criterion

Next we set our criterion, in other words, we chose an appropriate probability level. As already discussed, we generally use an alpha level of 0.05. This will determine the regions of the distribution of our parameter (in this instance, we are looking at sample means) in which our sample mean will fall, and whether it means we reject or fail to reject the null hypothesis.

Z scores

You will already have encountered the idea of Z scores and how they are used in order to determine the probability that your parameter falls within either the expected, or the unexpected range of possibilities. We will quickly review what we need to know:

We are able to convert our sample mean for example, into a score that fits in somewhere on the standard distribution graph, and we call this the Z score. Z scores are a special application of the transformation rules. The z score for an item, indicates how far and in what direction, that item deviates from its distribution's mean, expressed in units of its distribution's standard deviation. The mathematics of the z score transformation are such that if every item in a distribution is converted to its z score, the transformed scores will necessarily have a mean of zero and a standard deviation of one. So we are able to calculate a Z score we call Z-critical, and this is the Z score that defines the boundary of the region you will use in order to reject, or fail to reject, the null hypothesis. The Z-test score is the value you will have calculated from your sample value.

Test statistics

We have more than one type of test statistic that we can use, other than the Z-test score. We also have the T-test score for example, and we will discuss this at greater length in the next chapter. As we've already seen, these test scores allow us to convert our original measurement from our data set, into units that feature on the standard distribution, and this allows us to look up the probability of our score occurring randomly for example, in the table.

Setting a criterion

Z scores are especially informative when the distribution to which they refer is normal. In every normal distribution, the distance between the mean and a given Z score cuts off a fixed proportion of the total area under the curve. Statisticians have provided us with tables such as table B2 in your textbook by Zar, indicating the value of these proportions for each possible Z. If your Z-test value falls beyond either of the two areas cut off by Z-critical, then you can reject the null hypothesis in favour of your alternate hypothesis. Should your Z-test value fall within the area under the peak of the curve, then you have to conclude that your sample (or samples) have yielded a statistic that is not among those cases that would only occur alpha proportion of the time, if the hypothesis tested is true. You then fail to reject the null hypotheis.

Making a decision

In this particular instance, we have selected alpha to equal 0.05. In other words, we want to know whether our sample mean lies in the null distribution region of 95%, or does it fall in a more extreme part of the distribution, where there is a greater than 95% chance of the mean being drawn from the population. In order to answer the question, we convert our sample mean to a Z-score, and observe where it falls in the Z, or standard, distribution. If it falls outside the region of Z = 1.65 ( the critical region), then we are able to reject the null hypothesis.

One-tailed tests

When rejecting the null hypothesis in favour of the alternative hypothesis. We have more than one type of alternative hypothesis to select, depending on our particular experiment. We have both non-directional alternative hypotheses which we call two-tailed tests and we will discuss these later on in this chapter, and we have directional hypotheses, or one-tailed tests. In a one-tailed test, the direction of deviation from the null value is clearly specified. We place all of alpha in the one tail in a one-tailed test.

One-tailed tests should be approached with caution though: they should only be used in the light of strong previous research, theoretical or logical considerations. You need to have a very good reason beforehand that the outcome will lie in a certain direction.

One application in which one-tailed tests are used is in industrial quality control settings. For example, a company making statistical checks on the quality of its medical products is only interested in whether their product has fallen significantly below an acceptable standard. They are not usually interested in whether the product is better than average, this is obviously good, but in terms of legal liabilities and general consumer confidence in their product, they have to watch that their product quality is not worse than it should be. Hence one-tailed tests are often used.

Right-tailed tests

If using a one-tailed test, it can be either a right-tailed or a left-tailed test, and this just refers to the expected direction of your result. If you expect your sample mean to fall in the region beyond Z critical, then we refer to it as a right tailed test. If the sample mean is indeed larger than Z critical, then we can reject the null hypothesis. If it is less, then we may not.

Left-tailed tests

With a left-tailed test, we are expecting our sample mean to be less than Z critical, and we want to know whether it differs significantly from the population mean of 100 in the example we have here. If the sample mean is indeed less than Z critical, we are able to reject the null hypothesis.

Two-tailed tests
A two-tailed test requires us to consider both sides of the Ho distribution, so we split alpha, and place half in each tail. With a two-tailed test, we want to know whether our sample mean is significantly bigger or smaller than the population mean.

Two-tailed hypothesis testing
If our chosen alpha is 0.05 therefore, we will divide that into half, and use that figure to calculate our Z critical score, which will indicate the position on the distribution curve where, should our Z test score be observed to be either larger than, or smaller than, the Z critical, we can reject the null hypothesis.

One- and two-tail comparison
Here we have a comparison between a one-tail, and two-tail test, using the same alpha value.

## POSSIBLE QUESTIONS
## UNIT III

**PART A (20 X 1 = 20 Marks)**
**Question number 1 – 20 Online examination**

**PART-B (5 x 2 = 10 Marks)**

1. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 10 | 12 | 13 | 11 | 10 | 14 | 15 | 13 |
| B | 9 | 11 | 10 | 12 | 13 | | | |
| C | 11 | 10 | 15 | 14 | 12 | 13 | | |

Given, table value of F for (2,16) d.f at 5% level of significance is 3.63. Carry out the analysis of variance and state your conclusion.
2. Explain one-way classification in ANOVA.
3. What are the criterions for a uniformly most powerful test?
4. Describe power functions and OC functions
5. Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 10 | 12 | 13 | 11 | 10 | 14 | 15 | 13 |

| | | | | | | |
|---|---|---|---|---|---|---|
| B | 9 | 11 | 10 | 12 | 13 | |
| C | 11 | 10 | 15 | 14 | 12 | 13 |

Given, table value of F for (2,16) d.f at 5% level of significance is 3.63. Carry out the analysis of variance and state your conclusion.

6. Describe power functions and OC functions

**PART-C (5 x 6 = 30 Marks)**

1. Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha$ = 5%.

| | **Compact cars** | **Midsize cars** | **Full-size cars** |
|---|---|---|---|
| | 643 | 469 | 484 |
| | 655 | 427 | 456 |
| | 702 | 525 | 402 |
| $\bar{X}$ | 666.67 | 473.67 | 447.33 |
| S | 31.18 | 49.17 | 41.68 |

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics and Research Methodology
UNIT IV

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | **UNIT IV** | | | | | |
| 1 | 4 | Research in common parlance refers to | **A search for knowledge** | a search for attitude | a search for aptitude | a search for anything | **A search for knowledge** |
| 2 | 4 | A scientific and systematic search for pertinent information on a specific topic is known as | knowledge | **research** | information | attitude | **research** |
| 3 | 4 | The meaning of research as " a careful investigation specially through search for new facts in any branch of knowledge is given by | Redman | Mory | Spearman | **The advanced Learner's dictionary of current English** | **The advanced Learner's dictionary of current English** |
| 4 | 4 | Research as a " systematized effort to gain new knowledge" is given by | **Redman and Mory** | Spearman | Karl Pearson | The advanced Learner's dictionary of current English | **Redman and Mory** |
| 5 | 4 | ----------------- is the persuit of truth with the help of study | knowledge | **research** | information | attitude | **research** |
| 6 | 4 | In social science and business research we quite often use the term ----------- for descriptive research studies | ***Ex post facto research*** | *in post facto research* | *Ex pre facto research* | *in pre facto research* | ***Ex post facto research*** |
| 7 | 4 | The main characteristic of *Ex post facto research* method is that ------------ has no control over the variables | The sample | **the researcher** | the engineer | the citizen | **the researcher** |
| 8 | 4 | Studying of frequency of shopping, preferences of people, or similar data are studied by ----------- | Analytical research | applied research | Fundamental research | **descriptive research** | **descriptive research** |
| 9 | 4 | The use of facts already available and analyse these to make a critical evaluation of the material is coming under | | | | | |
| 10 | 4 | Finding a solution for an immediate problem facing a society or an industrial/business organization is | Analytical research | **applied research** | Fundamental research | descriptive research | **applied research** |
| 11 | 4 | Which of the following is not a quantitative approach | inferential approach | experimental approach | simulation approach | **qualitative approach** | **qualitative approach** |
| 12 | 4 | The purpose of -------- to research is to form a data base from which to infer the relationship of population | **inferential approach** | experimental approach | simulation approach | qualitative approach | **inferential approach** |
| 13 | 4 | Inferential approach is otherwise called as | **survey research** | deep research | qualitative research | vague research | **survey research** |
| 14 | 4 | Research approach in which artificial environment is created and data generated isknown as | inferential approach | experimental approach | **simulation approach** | qualitative approach | **simulation approach** |
| 15 | 4 | Building models for understanding future conditions is | inferential approach | experimental approach | **simulation approach** | qualitative approach | **simulation approach** |
| 16 | 4 | Research that is concerned with subjective assessment of attitudes,opinions and behaviour is known as | inferential approach | experimental approach | simulation approach | **qualitative approach** | **qualitative approach** |
| 17 | 4 | Increased amount of research make progress ------------ | **possible** | impossible | easier | difficult | **possible** |
| 18 | 4 | ____________ inculcates scientific and inductive thinking and it promotes the development of logical habits of thinking | knowledge | **research** | attitude | aptitude | **research** |
| 19 | 4 | Government's budgets should be based on | **people's need and availability of resources** | poverty status | business status of the country | economic status | **people's need and availability of resources** |
| 20 | 4 | ___________ provides the basis for nearly all Government policies in our economic system | **Research** | information | knowledge | money | **Research** |
| 21 | 4 | Investigating the structure and development of a market for the purpose of formulating efficient policies for purchasing pro | **market research** | business research | operational research | motivational research | **market research** |
| 22 | 4 | Application of mathematical ,logical and analytical techniques to the solution of business problems of cost minimization and | market research | business research | **operational research** | motivational research | **operational research** |
| 23 | 4 | ------------ has become an integral tool of business policy these days | business analysis | **market analysis** | operational analysis | motivational analysis | **market analysis** |
| 24 | 4 | Determining why people behave as they do is mainly concerned with market characteristics is | market research | business research | operational research | **motivational research** | **motivational research** |
| 25 | 4 | To those students who are to write a master's or PhD thesis research may means a | source of livelihood | **careerism** | the outlet of new ideas and insights | the generalization of new theories | **careerism** |
| 26 | 4 | To professionals in research methodology ,research may mean a | **source of livelihood** | careerism | the outlet of new ideas and insights | the generalization of new theories | **source of livelihood** |
| 27 | 4 | To philosophers and thinkers,research may means | source of livelihood | careerism | **the outlet of new ideas and insights** | the generalization of new theories | **the outlet of new ideas and insights** |
| 28 | 4 | To analysts and intellectuals,research may mean | source of livelihood | careerism | the outlet of new ideas and insights | **the generalization of new theories** | **the generalization of new theories** |
| 29 | 4 | To literary men and women research may mean | source of livelihood | **the development of new styles and creative work** | the outlet of new ideas and insights | the generalization of new theories | **the development of new styles and creative work** |
| 30 | 4 | ______________ is the fountain of knowledge for the sake of knowledge and an important source of providing guidelines | attitude | **research** | aptitude | knowledge | **research** |
| 31 | 4 | Research is a sort of formal ---------- which enables one to understand the new developments in one's field in a better way | **training** | placement | project | case study | **training** |
| 32 | 4 | A step of greatest importance in the entire research process is | development of working hypothesis | extensive literature survey | preparing the research design | **formulating the research problem** | **formulating the research problem** |
| 33 | 4 | The plan to select 12 of a city's 200 drug stores in a certain way constitutes a | **sample design** | executing the research | analytical design | operational research | **sample design** |
| 34 | 4 | Sampling of every 15th name on a list or every 10th house on one side of a street and so on is known as | random sampling | deliberate sampling | **systematic sampling** | stratified sampling | **systematic sampling** |
| 35 | 4 | The most extensively used method of collecting the datas in various economic and business surveys is | by observation | **by mailing of questionnaire** | through schedules | through personal interview | **by mailing of questionnaire** |
| 36 | 4 | In which method of collecting data the enumerators are appointed and given training | by observation | by mailing of questionnaire | **through schedules** | through personal interview | **through schedules** |
| 37 | 4 | At the end of the work there should be . | Summary | Result and discussion | **Bibliography** | Objective. | **Bibliography** |
| 38 | 4 | __________ is normally an early section in the dissertation. | Summary | Result and discussion | **Literature Review** | Objective | **Literature Review** |
| 39 | 4 | Preparatory stage for the literature review is ________. | Collecting data | General survey of related research. | Writing title | Writing contents | General survey of related research. |
| 40 | 4 | Age is an example of | **continuous variable** | non continuous variable | dependent variable | independent variable | **continuous variable** |
| 41 | 4 | The number of children in a family is an example of | continuous variable | **non continuous variable** | dependent variable | independent variable | **non continuous variable** |
| 42 | 4 | Readymade films and lectures are examples of | continuous variable | non continuous variable | dependent variable | **independent variable** | **independent variable** |
| 43 | 4 | Behavioural changes, occurring as a result of the environmental manipulations are | continuous variable | non continuous variable | **dependent variable** | independent variable | **dependent variable** |
| 44 | 4 | Independent variables that are not related to the purpose of the study,but may affect the dependent variable are termed as | continuous variable | non continuous variable | **extraneous variable** | independent variable | **extraneous variable** |
| 45 | 4 | Most of the social research comes under the category of | simple research | **descriptive research** | diagnostic research | exploratory research | **descriptive research** |
| 46 | 4 | One of the following is not an informal experimental design | before and after without control design | after only with control design | before and after with control design | **factorial design** | **factorial design** |
| 47 | 4 | The following are formal experimental designs except | latin square design | CRD | RBD | **after only with control design** | **after only with control design** |
| 48 | 4 | One way ANOVA is used to analyse | latin square design | **CRD** | RBD | after only with control design | **CRD** |
| 49 | 4 | The experimental design frequently used in agricultural research is | **latin square design** | CRD | RBD | after only with control design | **latin square design** |
| 50 | 4 | The latin square design is very similar to | one way ANOVA | **two way ANOVA** | correlation | regression | **two way ANOVA** |
| 51 | 4 | Factorial designs are mainly used in | agricultural phenomena | **economic and social phenomena** | biological research phenomena | physical research phenomena | **economic and social phenomena** |
| 52 | 4 | The statistical accuracy of the experiments is increased by | **principle of replication** | principle of randomization | principle of local control | principle of reliability | **principle of replication** |
| 53 | 4 | Rothamsted experimental station is | **centre for agricultural research in England** | centre for agricultural research in India | centre for agricultural research in America | centre for agricultural research in Canada | **centre for agricultural research in England** |
| 54 | 4 | The following are the basic principles of experimental designs except | principle of replication | principle of randomization | principle of local control | **principle of reliability** | **principle of reliability** |
| 55 | 4 | Studies which determine the frequency with which something occurs or its association with something else is called | descriptive research | **diagnostic research** | exploratory research | hypothesis testing research | **diagnostic research** |
| 56 | 4 | In a hypothesis testing research, if the independent variable intelligence is not manipulated it is called | **non experimental hypothesis testing research** | experimental hypothesis testing research | exploratory research | diagnostic research | **non experimental hypothesis testing research** |
| 57 | 4 | Apredictive statement that relates an independent variable to a dependent variable is called | **research hypothesis** | research plan | research summary | research design | **research hypothesis** |
| 58 | 4 | Summary must be | **Easy to read** | Difficult to read | Easy to understand | Difficult to understand | **Easy to read** |
| 59 | 4 | ________ makes a summary attractive to readers. | Figures | Fonts | **Title** | Format | **Title** |
| 60 | 4 | Summary should be written in ________ | **Easy terminology** | figures | tables | structures | **Easy terminology** |

## UNIT-V
## SYLLABUS

Research: Scope and significance – Types of Research – Research Process – Characteristics of good research – Problems in Research – Identifying research problems. Research Designs – Features of good designs.
Sources of information: Journals, eJournals, books, biological abstracts, preparation of index cards, review writing, article writing – structure of article, selection of journals for publication – Impact factor – citation index and H index. Proposal writing for funding. IPR and patenting. Concepts and types.

### OBJECTIVES OF RESEARCH

The purpose of research is to discover answers to questions through the application of scientific procedures. The main aim of research is to find out the truth which is hidden and which has no discovered as yet. Though each research study has its own specific purpose, we may think of research objectives as falling into a number of following broad groupings:

- To gain familiarity with a phenomenon or to achieve new insights into it (studies with this object in view are termed as *exploratory* or *formulative*research studies);
- To portray accurately the characteristics of a particular individual, situation or a (studies with this object in view are known as *descriptive* research studies);
- To determine the frequency with which something occurs or with which it as so with something else (studies with this object in view are known as *diagnostic* re: studies);
- To test a hypothesis of a causal relationship between variables (such studies are *hypothesis-testing* research studies).

### MOTIVATION IN RESEARCH

What makes people to undertake research? This is a question of fundamental importance

possible motives for doing research may be either one or more of the following:

1. Desire to get a research degree along with its consequential benefits;

2. Desire to face the challenge in solving the unsolved problems, i.e., concern over practical problems initiates research;
3. Desire to get intellectual joy of doing some creative work;
4. Desire to be of service to society;
5. Desire to get respectability.

However, this is not an exhaustive list of factors motivating people to undertake research st

Many more factors such as directives of government, employment conditions, curiosity abot things, desire to understand causal relationships, social thinking and awakening, and the like n

well motivate (or at times compel) people to perform research operations.

**TYPES OF RESEARCH**

The basic types of research are as follows:

i. Descriptive vs. Analytical: Descriptive researchincludes surveys and fact-finding enquiries of different kinds. The major purpose of description of the state of affairs as it exists at present. In social science and business research we quite often use the term Ex post facto researchfor descriptive research studies. The main characteristic of this method is that the researcher has no control over the variables; he can only report what has happened or what is happening. Most ex post facto researchprojects are used for descriptive studies in which the researcher seeks to measure such items as, for example, frequency of shopping, preferences of people, or similar data. Ex post facto studiesalso include attempts by researchers to discover causes even when they cannot control the variables. The methods of research utilized in descriptive research are survey methods of all kinds, including comparative and correlation methods. In analytical research,on the other hand; the researcher has to use facts or information already available, and analyze these to make a critical evaluation of the material.

ii. Applied vs. Fundamental: Research can either be applied (or action) research or fundamental (to basic or pure) research. Applied research aims at finding a solution for an

immediate problem facing a society or an industrial business organization, whereas fundamental research is mainly concerned with generalizations and with the formulation of a theory. "Gathering knowledge for knowledge's sake is termed 'pure' or 'basic' research." Research concerning some natural phenomenon or relating to pure mathematics are examples of fundamental research. Similarly, research studies, concerning human behavior carried on with a view to make generalizations about human behavior, are also examples of fundamental research, but research aimed at certain conclusions (say, a solution) facing a
concrete social or business problem is an example of applied research. Research to identify
social, economic or political trends that. May affect a particular institution or the copy research (research to find out whether certain cornmunications will be read and understood) or the marketing research or evaluation research are examples of applied research. Thus, the central aim of applied research is to discover a solution for some pressing practical problem, whereas basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge.

iii. Quantitative vs. Qualitative: Quantitative research is based on the measurement of quantity
or amount. It is applicable to phenomena that can be expressed in terms of quantity. Qualitative research, on the other hand, is concerned with qualitative phenomenon, i.e.,
phenomena relating to or involving quality or kind. For instance, when we are interested "in investigating the reasons for human behavior (i.e., why people think or do certain things), we quite often talk of 'Motivation Research', an important type of qualitative research. This type of research aims at discovering the underlying motives and desires, using in depth interviews for the purpose. Other techniques of such research are word association tests, sentence completion tests, story completion tests and similar other projective techniques. Attitude r opinion research i.e., research\h designed to find out how people feel or what they think about a particular subject or

institution is also qualitative research. Qualitative research is specially important in the behavioral sciences where the aim is to discover the underlying motives of human behavior. Through such research we can analyze the various factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing. It may be stated, however, that to apply qualitative research in Practice is relatively a difficult job and therefore, while doing such research, one should seek guidance from experimental psychologists.

iv. Conceptual vs. Empirical: Conceptual research is that related to some abstract idea(s) or

theory. It is generally used by philosophers and thinkers. to develop new concepts or to

reinterpret existing ones. On the other hand, empirical research relies on experience or

observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. We can also call it as experimental type of research. In such a research it is necessary to get at facts firsthand, at their source, and actively to go about doing certain things to stimulate the production of desired information. In such a research, the researcher must first provide himself with a working hypothesis or guess as to the probable results. He then works to get enough facts (data) to prove or disprove his hypothesis. He then sets up experimental designs which he thinks will manipulate the persons or the materials concerned so as to bring forth the desired information. Such research is thus characterized by the experimenter's control over the variables under study and his deliberate manipulation of one of them to study its effects. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis.

v. Some Other Types of Research: All other types of research are variations of one or more
of the above stated approaches, based on either the purpose of research, or the time required to accomplish research, on the environment in which research is done, or on the
basis of some other-similar factor. Form the point of view of time; we can think of research
either a one-time research O n autodial research. In the former case the research is' confine to a single time-period, whereas in the latter case the research is carried on over
several time-periods. Research can be field-setting research or Laboratory research or
simulation research, depending upon the environment in which it is to be carried out. Research can as well be understood as clinical or diagnostic research such research follow case-study methods or in depth approaches to reach the basic causal relations. Such studies usually go deep into the causes of things or events that interest us using very small samples and very deep probing data gathering devices. The research may be exploratory or it may be formalized. The objective of exploratory research is the development of hypotheses rather than their testing, whereas formalized research studies are those with substantial structure and with specific hypotheses to be tested. Historical research1 is that which utilizes historical sources like documents, remains, etc. to study events or ideas of the past; including the philosophy of persons and groups at any remote point of time Research can also be classified a conclusion-oriented and decision-oriented. While doing conclusion oriented research, a researcher is free to pick up a problem, redesign the enquiry as he proceeds and is prepared to conceptualize as he wishes. Decision-oriented research is always for the need of a decision maker and the researcher in this case is not free to embark upon research according to his own inclination. Operations research is an example
of decision oriented research since it is a scientific method of providing executive

departments. With a quantitative basis for decisions regarding operations under their control.

**Importance of Knowing How Research is done**

The study of research methodology gives the student the necessary training in gathering material and arranging or card-indexing them, participation in the fieldwork when required, and also training in techniques for the collection of data appropriate to particular problems, in the use of statistics, questionnaires and controlled experimentation and in recording evidence, sorting it out and interpreting it. In fact, importance of knowing the methodology or how research is done stems from the following considerations: -

i. For one who is preparing himself for a career of carrying out research, the importance of knowing research methodology and research techniques is obvious since the same constitute the tools of his trade. The knowledge of methodology provides good training specially to the new research worker and enables him to do better research. It helps him to develop disciplined thinking or a 'bent of mind' to observe the field objectively. Hence, those aspiring for careerism in research must develop the skill of using research techniques and must thoroughly understand the logic behind them.

ii. Knowledge of how to do research will inculcate the ability to evaluate and use research results with reasonable confidence. In other' words, we can state that the knowledge of research methodology is helpful in various fields such as government or business administration, community development and social work where persons are increasingly called upon to evaluate and use research results for action.

The chart indicates that the research process consists of a number of closely related activities, as shown through I to VII. But such activities overlap continuously rather than following a strictly prescribed sequence. At times, the first step determines the nature of the last step to be undertaken. If subsequent procedures have not been taken into account in the early stages, serious difficulties may arise which may even prevent the completion of the study. One should remember that the various steps involved in a research process are not mutually exclusive; nor they are separate and distinct. They do not necessarily follow each other in any specific order and the researcher has to be constantly anticipating' at each step in the research process the requirements of the subsequent steps. However, the following order concerning various provides a useful procedural guideline regarding the research process: (1) formulating the research problem; (2) extensive literature survey; (3) developing the hypothesis; (4) preparing the research design; (5) determining sample design; (6) collecting the data; (7) execution of the project; (8) analysis of data; (9) hypothesis testing; (10) generalisations and interpretation, and (11) preparation of the report or presentation of the results, i.e., formal write-up of conclusions reached.

A brief description of the above stated steps will be helpful.

**1. Formulating the research problem:** There are two types of research problems, viz., those which relate to stales of nature and those which relate to relationships between variables. At the very outset the researcher must single out the problem he wants to study, i.e., he must decide the general area of interest or aspect of a subject-matter that he would like to inquire into. Initially the problem may be stated in a broad general way and then the ambiguities, if any, relating to the problem be resolved. Then, the feasibility of a particular solution has to be considered before a working formulation of the problem can be set up. The formulation of a general topic into a specific research problem, thus, constitutes the first step in a scientific enquiry. Essentially two steps are involved in formulating the research problem, viz., understanding the problem thoroughly, and rephrasing the same into meaningful terms from an analytical point of vie

The best way of understanding the problem is to discuss it with one's own colleagues or with those having some expertise in the matter. In an academic institution the researcher can seek the help from a guide who is usually an experienced man and has several research problems in mind. Often, the guide puts forth the problem in general terms and it is up to the researcher to narrow it down and phrase the problem in

operational terms. n private business, units or in governmental organisations, the problem is usually earmarked by the administrative agencies with whom the researcher can discuss as to how the problem originally came about and what considerations are involved in its possible solutions.

The researcher must at the same time examine all available literature to get himself acquainted with the selected problem. He may review two types of literature-the conceptual literature concerning the concepts and theories, and the empirical literature consisting of studies made earlier which are similar to the one propose 1'The basic outcome of this review will be the know ledge as to what data and other materials are available for operational purposes which will enable the researcher to specify his own research problem in a meaningful context. After this the researcher rephrases the problem into analytical or operational terms i.e., to put the problem in as specific terms as possible. This task of formulating, or defining, a research problem is a step of greatest importance in the entire research process. The problem to be investigated must be defined unambiguously for that will help discriminating relevant data from irrelevant ones. Care must, however, be taken to-verify the objectivity and validity of the background facts concerning the problem. Professor W.A. Neiswanger correctly states that the statement of the objective is of basic importance because it determines the data which are to be collected, the characteristics of the data which are relevant, relations which are to be explored, the choice of techniques to be used in these explorations and the form of the final report. If there are certain pertinent terms, the same should be clearly defined along with the task of formulating the problem. In fact, formulation of the problem often follows a sequential pattern where a number of formulations are set up, each formulation more specific than the preceeding one, each one phrased in more analytical terms, and each more realistic in terms of the available data and resources

.

**2. Extensive literature survey:** Once the problem is formulated, a brief summary of it should be
written down. It is compulsory for a research worker writing a thesis for a Ph.D. degree to write a synopsis of the topic and submit it to the necessary Committee or the Research Board for approval. At this juncture the researcher should undertake extensive literature survey connected with the problem. For this purpose, the abstracting and indexing journals and published or unpublished bibliographies are the first place to go to. Academic journals, conference proceedings, government reports, books etc., must be tapped depending on the nature of the problem. In this process, it should be remembered that one source will lead to another. The earlier studies, if any, which are similar to the study in hand, should be carefully studied. A good library will be a great help to the researcher at this stage

**3. Development of working hypotheses:** After extensive literature survey, researcher should
state in clear terms the working hypothesis or hypotheses. Working hypothesis is tentative assumption made in order to draw out and test its logical or empirical consequences. As such the manner in which research hypotheses are developed is particularly important since they provide the focal point for research. They also affect the manner in which tests must be conducted in the analysis of data and indirectly the quality of data which is required for the analysis In most types of research, the development of

working hypothesis plays' an important role. Hypothesis should be very specific and limited to the piece of research in hand because it has to be tested. The role of the hypothesis is to guide the researcher by delimiting the area of research and to keep him on the right track. It sharpens his thinking and focuses attention on the more important facets of the problem. It also indicates the type of data required and the type of methods of data analysis to be used. How does one go about developing working hypotheses? The answer is by using the following approach:

- Discussions with colleagues and experts about the problem, its origin and the objectives in seeking a solution;
- Examination of data and records, if available, concerning the problem for possible trends, peculiarities and other clues;
- Review of similar studies in the area or of the studies on similar problems; and
- Exploratory personal investigation which involves original field interviews on a limited scale with interested parties and individuals with a view to secure greater insight into the practical aspects of the problem.

Thus, working hypotheses arise as a result of a-priori thinking about the subject, examination of the available data and material including related studies and the counsel of experts and interested parties. Working hypotheses are more useful when stated in precise and clearly defined terms. It may as well be remembered that occasionally we may encounter a problem where we do not need working hypotheses, specially in the case of exploratory or formulative researches which do not aim at testing the hypothesis. But as a general rule, specification of working hypotheses in another basic step of the research process in most research problems.

4.**Preparing the research design:** The research problem having been formulated in clear cut
terms, the researcher will be required to prepare a research design, i.e., he will have to state the
conceptual structure within which research would be conducted. The preparation of such a design facilitates research to be as efficient as possible yielding maximal information. In other words, the function of research design is to provide for the collection of relevant evidence with minimal expenditure of effort, time and money. But how all these can be achieved depends mainly on the research purpose. Research purposes may be grouped into four categories, viz., (i) Exploration, (ii) Description, (iii) Diagnosis, and (iv) Experimentation. A flexible research design which provides opportunity for considering many different aspects of a problem is considered appropriate if the purpose of the

research study is that of exploration. But when the purpose happens to be an accurate description of a situation or of an association between variables, the suitable design will be one that minimises bias and maximises the reliability of the data collected and analysed.

.

There are several research designs, 'such as, experimental and non-experimental hypothesis

testing. Experimental designs can be either informal designs (such as before-and-after without control after-only with control, before-and-after with control) or formal designs (such as completely randomized design, randomized block design, Latin square design, simple and complex factorial designs), out of which the researcher must select one for his own project.

The preparation of the research design, appropriate for a particular research problem, involves usually the consideration of the following:

- The means of obtaining the information;
- The availability and skills of the researcher and his staff (if any)
- Explanation of the way in which selected means of obtaining information will be organised and the reasoning leading to the selection;
- The time available for research; and
- The cost factor relating to research, i.e., the finance available for the purpose.

**5. Determining sample design:** All the items under consideration in any field of inquiry constitute a 'universe' or 'population' complete enumeration of all the items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry when all the items are' covered no element of chance is left and highest accuracy is obtained. But in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of observations increases. Moreover, there is no way of checking the element of bias or its extent except through a resurvey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy. Not only this, census inquiry is not possible in practice under many circumstances. For instance, blood testing is done only on sample basis. Hence, quite often we select only a few items from the universe for our study purposes. The items so selected constitute what is technically called a sampl0 '

The researcher must decide the way of selecting a sample or what is popularly known as the sample design other words, a sample design is a definite plan determined before any data are actually collected for obtaining a sample from a given population. Thus, the plan to select 12 of a city's 200 drugstores in a certain we constitutes a sample design. Samples can be either probability samples or non-probability samples. With probability sample's each element has a known probability of being included in the sample but e non-probability samples do not allow the researcher to determine this probability. Probability samples are those based on. simple random sampling, systematic sampling, stratified sampling, cluster/area sampling whereas non-probability samples are those based on convenience sampling, judgement sampling and quota sampling techniques. A brief mention of the important sample designs is as follows:

**(i)** ***Deliberate sampling:***Deliberate sampling is also known as purposive or non-probability sampling. This sampling method involves purposive or deliberate selection of particularunits of the universe for constituting a sample which represents the universe. When population elements are selected for inclusion in the sample based on the ease of access, it can be called *convenience sampling*. If a researcher wishes to secure data from, say, gasoline buyers, he may select a fixed number of petrol stations and may conduct interviews at these stations. This would be an example of convenience sample of gasoline buyers. At times such a procedure may give very biased results particularly when the population is not homogeneous. On the other hand, in *judgement sampling* the researcher's judgement is used for selecting items which he considers as representative of the population. For example, a judgement sample of college students might be taken to secure reactions to a new method of teaching. Judgement sampling is used quite frequently in qualitative research where the desire happens to be to develop hypotheses rather than to generalise to larger populations.

**(ii)** ***Simple random sampling:***This type of sampling is also known as chance sampling or probability sampling where each and every item in the population has an equal chance of inclusion in the sample and each one of the possible samples, in case of finite universe, has the same probability of being selected. For example, if we have to select a sample of 300 items from a universe of 15,000 items, then we can put the names or numbers of all the 15,000 items on slips of paper and conduct a lottery. Using the random number tables is another method of random sampling. To select the sample, each item is assigned a number from 1 to 15,000. Then, 300 five digit random numbers are selected from the table. To dothis we select some random starting point and then a systematic pattern is used in proceeding through the table. We might start in the 4th row, second column and proceed down the column to the bottom of the table and then move to the top of the next column to the right. When a number exceeds the limit of the numbers in the frame, in our case over 15,000, it issimply passed over and the next number selected that does fall within the relevant range. Since the numbers were placed in the table in a completely random fashion, the resulting sample is random. This procedure gives each item an equal probability of being selected. Incase of infinite population, the selection of each item in a random sample is controlled by the same probability and that successive selections are independent of one another.

**(iii)** ***Systematic sampling:***In some instances the most practical way of sampling is to select every 15th name on a list, every 10th house on one side of a street and so on. Sampling of this type is known as systematic sampling. An element of randomness is usually introduced into this kind of sampling by using random numbers to pick up the unit with which to start.This procedure is useful when sampling frame is available in tire form of a list. In such a design the selection process starts by picking some random point in the list and then every *n*th element is selected until the desired number is secured.

**(iv)*****Stratified sampling:***If the population from which a sample is to be drawn does not constitute a homogeneous group, then stratified sampling technique is applied so as to obtain a representative sample. In this technique, the population is stratified into a number of non-overlapping subpopulations or strata and sample items are selected from each stratum. Ifthe items selected from each stratum is based on simple random sampling the entire procedure first stratification and then simple random sampling, is known as *stratified random sampling.*

**(iv)** ***Quota sampling:***In stratified sampling the cost of taking random samples from individual strata is of tell so expensive that interviewers are simply given quota to be filled from different strata, the actual selection of items for sample being left to the interviewer’s judgement. This is called quota sampling. The size of the quota for each stratum is generally proportionate to the size of that stratum in the population. Quota sampling is thus an important form of non-probability sampling. Quota samples generally happen to be judgement samples rather than random samples.

***(vi) Cluster sampling and area sampling:*** Cluster sampling involves grouping the population and then selecting the groups or the clusters rather than individual elements for inclusion in the sample. Suppose some departmental store wishes to sample its credit card holders. It has issued its cards to 15,000 customers. The sample size is to be kept say 450. For cluster sampling this list of 15,000 card holders could be formed into 100 clusters of 150 card holders each. Three clusters might then be selected for the sample randomly. The sample size must often be larger than the simple random/sample to ensure the same level of accuracy because is cluster sampling procedural potential for order bias and other source of error is usually accentuated. The clustering approach can, however, make the sampling procedure relatively easier and increase the efficiency of field work, specially: in the case of personal interviews.
*Area sampling* is quite close to cluster sampling and is often talked about when the total geographical area of interest happens to be big one. Under area sampling we first divide the total area into a number of smaller non-overlapping areas, generally called geographical clusters, then a number of these smaller areas are randomly selected, and all units in these small areas are included in the sample. Area sampling is specially helpful where we do not have the list of the population concerned. It also makes the field interviewing more efficient since interviewer can do many interviews at each location.

***(vii) Multi-stage sampling:***This is a further development of the idea of cluster sampling. Thistechnique is meant for big inquiries extending to a considerably large geographical area like an entire country. Under multi-stage sampling the first stage may be to select large primarysampling units such as states, then districts, then towns and finally certain families withintowns. If the technique of random-sampling is applied at all stages, the sampling procedure is described as multi-stage random sampling.

*(**viii) Sequential sampling:***This is somewhat a complex sample design where the ultimate sizeof the sample is not fixed in advance but is determined according to mathematical decisions on the basis of information yielded as survey progresses. This design is usually' adopted under acceptance sampling plan in the context of statistical quality control.

In practice, several of the methods of sampling described above may well be used in the same study in which case it can be called mixed sampling. It may be pointed out here that normally one should resort to random sampling so that bias can be eliminated and sampling error can be estimated. But purposive sampling is considered desirable when the universe happens to be small and a known characteristic of it is to be studied intensively. Also, there are conditions under which sample designs other than random sampling may be considered better for reasons like convenience and low costs. The sample design to be used must be decided by the researcher taking into consideration the nature of the inquiry and other related factors.

**6. Collecting the data:** In dealing with any real life problem it is often found that data at hand are inadequate, and hence, it becomes necessary to c6ilect data that are appropriate) There are several ways of collecting the appropriate data which differ considerably in context of money costs, time and other resources at the disposal of the researcher.

Primary data can be collected either through experiment or through survey. If the researcher conducts an experiment, he observes some quantitative measurements, or the data, with the help of which he examines the truth contained in his hypothesis. But in the case of a survey, data can be collected by anyone or more of the following ways:

(i) **By observation:**This method implies the collection of information by way of investigator's own observation, without interviewing the respondents. The information obtained relates towhat is currently happening and is not complicated by either the past behaviour or future intentions or attitudes of respondents. This method is no doubt an expensive method and the information provided by this method is also very limited. As such this method is notsuitable in inquiries where large samples are concerned.

(ii) **Through personal interview:** The investigator follows a rigid procedure and seeks answersto a set of pre-conceived questions through personal interviews. This method of collectingdata is usually carried out in a structured way where output depends upon the ability of the interviewer to a large extent. .

(iii) **Through telephone interviews:** This method of collecting information involves contactingthe respondents on telephone itself. This is not a very widely used method but it plays an important role in industrial surveys in developed regions, particularly, when the survey has to be accomplished in a very limited time.

(iv) **By mailing of questionnaires:** The researcher and the respondents do come in contact with each other if this method of survey is adopted. Questionnaires are mailed to the respondents with a request to return after completing the same. It is the most extensively used method in various economic and business surveys.

Before applying this method, usually a Pilot Study for testing the questionnaire is conduced which reveals the weaknesses, if any, of the questionnaire. Questionnaire to be used must be prepared very carefully so that it may prove to be effective in collecting the relevant information.

**(v) Through schedules:** Under this method the enumerators are appointed and given training. They are provided with schedules containing relevant questions. These enumerators go to respondents with these schedules. Data are collected by filling up the schedules by enumerators on the basis of replies given by respondents. Much depends upon the capability of enumerators so far as this method is concerned. Some occasional field checks on the work of the enumerators may ensure sincere work.

The researcher should select one of these methods of collecting the data taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources,available time and the desired degree of accuracy.Though he should pay attention to all these factors but much depends upon the ability and experience of the researcher. In this context *Dr A.L. Bowley*very aptly remarks that in collection of statistical data commonsense is the chief requisite and experience the chief teacher.

1. **Execution of the project:** Execution of the project is a very important step in the research process. If the execution of the project proceeds on correct lines, the data to be collected would be adequate and dependable. The researcher should see that the project is executed in a systematic manner and in time. If the survey is to be conducted by means of structured questionnaires, data can be readily machine-processed. In such a situation, questions as well as the possible answers may be coded. If the data are to be collected through interviewers, arrangements should be made for proper selection and training of the interviewers. The training may be given with the help of instruction manuals which explain clearly the job of the interviewers at each step. Occasional field checks should be made to ensure that the interviewers are doing their assigned job sincerely and efficiently. A careful watch should be kept for unanticipated factors in order to keep the survey as much realistic as possible. This, in other words, means that steps should be taken to ensure that the survey is under statistical control so that the collected information is in accordance with the pre-defined standard of accuracy. If some of the respondents do not cooperate, some suitable methods should be designed to tackle till's problem. One method of dealing with the non-response problem is to make a list of the non-respondents and take a small sub-sample of them, and then with the help of experts vigorous efforts can be made for securing response.

**8. Analysis of data:** After the data have been collected, the researcher turns to the task of analysing them. The analysis of data requires a number of closely related operations such as establishment of categories, the application of these categories to raw data through coding, tabulation and then drawing statistical inferences. The unwieldy data should necessarily be condensed into a few manageable groups and tables for further analysis. Thus, researcher should classify the raw data into some purposeful and usable categories. *Coding* operation is usually done at this stage through which the categories of data are transformed into symbols that may be tabulated and counted. *Editing* is the procedure that improves the quality of the data for coding. With coding the stage is ready for tabulation. *Tabulation* is a part of the technical procedure wherein the classified data are put in the form of tables. The mechanical devices can be made use of at this juncture. A great deal of data, specially in large inquiries, is tabulated by computers. Computers not only save time but also make it possible to study large number of variables affecting a problem simultaneously.

Analysis work after tabulation is generally based on the computation of various percentages, coefficients, etc., by applying various well defined statistical formulae. In the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to tests of significance to determine with what validity data can be said to indicate any conclusion(s}. For instance, if there are two samples of weekly wages, each sample being drawn from factories in different parts of the same city, giving two different mean values, then our problem may be whether the two mean values are significantly different or the difference is just a matter of chance. Through the use of statistical tests we can establish whether such a difference is a real one or is the result of random fluctuations. If the difference happens to be real, the inference will be that the two samples

***3. Good research is empirical:***It implies that research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.

***4. Good research is, replicable:***This characteristic allows research results to be verified by replicating the study and thereby building a sound basis for decisions.

**Problems Encountered by Researchers in India**

Researchers in India, particularly those engaged in empirical research, are facing several problems.

Some of the important problems are as follows:

1. *The lack of a scientific training in the methodology of research* is a great impedimentfor researchers in our country. There is paucity of competent researchers. Many researchers take a leap in the dark without knowing research methods. Most of the work, which goes in the name of research, is not methodologically sound. Research to many researchers and even to their guides, is mostly a scissor and paste job without any insight shed on the collated materials. The consequence is obvious, viz., the research results, quite often, do not reflect the reality or realities. Thus, a systematic study of research methodology is an urgent necessity. Before undertaking research projects, researchers should be well equipped with all the, methodological aspects. As such, *efforts should be made to provide short- duration intensive courses for meeting this requirement.*
2. There is *insufficient interaction* between the university research departments on one side and business establishments, government departments and research institutions on the other side. A great deal of primary data of non-confidential nature remains untouched / untreated by the researchers for want of proper contacts. *Efforts should be made to develop satisfactory liaison among all concerned for better and realistic researches.* There is need for developing some mechanisms of a university-industry interaction programme so that academics can get ideas from practitioners on what needs to be researched and practitioners can apply the research done by the academics.
3. Most of the business units in our country do not have the confidence that the material supplied by them to researchers will not be misused and as such they are often reluctant in supplying the needed information to researchers, The concept of secrecy seems to be sacrosanct to business organisations in the country so much so that it proves an impermeable barrier to researchers. Thus, *there is the need for generating the confidence that the information/data obtained from a business unit will not be misused.*
4. *Research studies overlapping one another are undertaken quite often for want of adequate information.* This results in duplication and fritters away resources. This problem can be solved by proper compilation and revision, at regular intervals, of a list of subjects on which and the places where the research is going on. Due attention should be given toward identification of research problems in various disciplines of applied science which are of immediate concern to the industries.

5. *There does not exist a code of conduct for researchers* and inter-university and inter- departmental rivalries are also quite common. Hence, there is need for developing a code of conduct for researchers which, if adhered sincerely, can win over this problem.
6. Many researchers in our country also face *the difficulty of adequate and timely secretarial assistance,* including computerial assistance. This causes unnecessary delays in the completion of research studies. All possible efforts be made in this direction so that efficient secretarial assistance is made available to researchers and that too well in time. University Grants Commission must play a dynamic role in solving this difficulty.
7. *Library management and functioning is not satisfactory at many places* and much ofthe time and energy of researchers are spent in tracing out the books, journals, reports, etc., rather than in tracing out relevant material from them.
   .
8. *There is also the problem that many of our libraries are not able to get copies of oldand new Acts / Rules, reports and other government publications in time.* This problem is felt more in libraries which are away in places from Delhi and/or the state capitals. Thus efforts should be made for the regular and speedy supply of all governmental publications to reach our libraries.
9. *There is also the difficulty of timely availability of published data* from various government and other agencies doing this job in our country. Researcher also faces the problem on account of the fact that the published data vary quite significantly because ofdifferences in coverage by the concerning agencies.
10. There may, at times, take place *the problem of conceptualization* and also problemsrelating to the process of data collection and related things.

## DEFINING THE RESEARCH PROBLEM

In research process, the first and foremost step happens to be that of selecting and properly defining a research problem. A researcher must find the problem and formulate it so that it becomes susceptible to research. Like a medical doctor, a researcher must examine all the symptoms (presented to him or observed by him) concerning a problem before he can diagnose correctly. To define a problem correctly, a researcher must know: what a problem is?

## WHATIS A RESEARCH PROBLEM?

A research problem, in general, refers to some difficulty which a researcher experiences in the context of either a theoretical or practical situation and wants to obtain a solution for the same. Usually we say that a research problem does exist if the following conditions are met with:

*a.* There must be an individual (or a group or an organisation), let us call it *'I'*, to whom the problem can be attributed. The individual or the organisation, as the case may be, occupies an environment, say *'N'*, which is defined by values of the uncontrolled variables, $Y_j$.

b. There must be at least two courses of action, say $C_1$ and $C_2$, to be pursued. A course of action is defined by one or more values of the controlled variables. For example, the number of items purchased at a specified time is said to be one course of action.

c. There must be at least two possible outcomes, say $O_1$ and $O_2$ of the course of action, of which one should be preferable to the other. In other words, this means that there must be at least one outcome that the researcher wants, i.e., an objective.

d. The courses of action available must provides some chance of obtaining the objective, but they cannot provide the same chance, otherwise the choice would not matter. Thus, if $p$ $(O_j \backslash I, C_1, N)$ represents the probability that an outcome $O_j$ will occur, if *I* select C in *N*, then *P* $(O_1 \backslash I, C_1, N)$ ;# *P* $( O_1 \backslash I, 2, N)$ . In simple words, we can say that the choices must have unequal efficiencies for the desired outcomes.

Over and above these conditions, the individual or the organisation can be said to have theproblem only if 'I' does not know what course of action is best, i.e., '1', must be in doubt about the solution. Thus, an individual or a group of persons can be said to have a problem which can be technically described as a research problem, if they (individual or the group), having one or more desired outcomes, are confronted with two or more courses of action that have some but not equal efficiency for the desired objective(s) and are in doubt about which course of action is best. We can, thus, state the components [1] of a research problem as under:

i. There must be an individual or a group which has some difficulty or the problem.

ii. There must be some objective(s) to be attained at. If one wants nothing, one cannot have a problem.

iii. There must be alternative means (or the courses of action) for obtaining the objective(s) one wishes to attain. This means that there must be *at least two means* available to a researcher for if he has no choice of means, he cannot have a problem.

iv. There must remain some doubt in the mind of a researcher with regard to the selection of alternatives. This means that research must answer the question concerning the relative efficiency of the possible alternatives.

v. There must be some environment(s) to which the difficulty pertains.

Thus, a research problem is one which requires a researcher to find out the best solution for the given problem, i.e., to find out by which course of action the objective 'can be attained optimally in the context of a given environment. There are several factors which may result in making the problem complicated. For instance, the environment may change affecting the efficiencies of the courses of action or the values of the outcomes; the number of alternative courses of action may be very large; persons not involved in making the decision may be affected by it and react to it favourably or unfavourably, and similar other factors. All such elements (or at least the important ones) may be thought of in context of a research problem.

## SELECTING THE PROBLEM

The research problem undertaken for study must be carefully selected. The task is a difficult one, although it may not appear to be so. Help may be taken from a research guide in this connection. Nevertheless, every researcher must find out his own salvation for research problems cannot be borrowed. A problem must spring from the researcher's mind like a plant springing from its own seed. If our eyes need glasses, it is not the optician alone who decides about the number of the lens we require. We have to see ourselves and enable him to prescribe for us the right number by cooperating with him. Thus, a research guide can at the most only help a researcher choose a subject. However, the following points may be observed by a researcher in selecting a research problem or a subject for research:

i. Subject which is overdone should not be normally chosen, for it will be a difficult task to throw any new light in such a case.

ii. Controversial subject should not become the choice of an average researcher

iii. Too narrow or too vague problems should be avoided.

iv. The subject selected for research should be familiar and feasible so that the related research material or sources of research are within one's reach. Even then it is quite difficult to supply definitive ideas concerning how a researcher should obtain ideas for his research. For this purpose, a researcher should contact an expert or a professor in the University who is already engaged in research. He may as well read articles published in current literature available on the subject and may think how the techniques and ideas discussed therein might be applied to the solution of other problems. He may discuss with others what he has in mind concerning a problem. In this way he should make all possible efforts in selecting a problem.

v. The importance of the subject, the qualifications and the training of a researcher, the costs involved, the time factor are' few other criteria that must also be considered in selecting a problem-In other words, before the final selection of a problem is done, a researcher must ask himself the following questions:
   a. Whether he is well equipped in terms of his background to carry out the research?
   b. Whether the study falls within the budget he can afford?
   c. Whether the necessary cooperation can be obtained from those who must participate in research as subjects?

   If the answers to all these questions are in the affirmative, one may become sure so far as the practicability of the study is concerned.

vi. The selection of a problem must be preceded by a preliminary study. This may not be necessary when the problem requires the conduct of a research closely similar to one that has already been done. But when the field of inquiry is relatively new and does not have available a set of well developed techniques, a brief feasibility study must always be undertaken.

If the subject for research is selected properly by observing the above mentioned points, the research will not be a boring drudgery; rather it will be love's labour. In fact, zest for work is a must. The subject or the problem selected must involve the researcher and must have an upper most place in his mind so that he may undertake all pains needed for the study

## NECESSITY OF DEFINING THE PROBLEM

Quite often we all hear that a problem clearly stated is a problem half solved. This statement signifies the need for defining a research problem. The problem to be investigated must be defined unambiguously for that will help to discriminate relevant data from the irrelevant ones. A proper definition of research problem will enable the researcher to be on the track whereas an ill-defined problem may create hurdles. Questions like: What data are to be collected? What characteristics of data are relevant and need to be studied? What relations are to be explored. What techniques are to be used for the purpose? and similar other questions crop up in the mind of the researcher who can well plan his strategy and find answers to all such questions only when the research problem has been well defined. Thus, defining a research problem properly is a prerequisite for any study and is a step of the highest importance. In fact, formulation of a problem is often more essential than its solution. It is only on careful detailing the research problem that we can work out the research design and can smoothly carry on all the consequential steps involved while doing research.

## TECHNIQUE INVOLVED IN DEFINING A PROBLEM

Let us start with the question: What does one mean when he/she wants to define a research problem? The answer may be that one wants to state the problem along with the bounds within which it is to be studied. In other words defining a problem involves the task of laying down boundaries within which a researcher shall study the problem with a pre-determined objective in view.

How to define a research problem is undoubtedly a herculean task. However, it is a task that must be tackled intelligently to avoid the perplexity encountered in a research operation. The usual approach is that the researcher should himself pose a question (or in case someone else wants the researcher to carry on research, the concerned individual, organisation or an authority should pose the question to the researcher) and set-up techniques and procedures for throwing light on the question concerned for formulating or defining the research problem. But such an approach generally does not produce definitive results because the question phrased in such a fashion is usually in broad general terms and as such may not be in a form suitable for testing.

Defining a research problem properly and clearly is a crucial part of a research study and must in no case be accomplished hurriedly. However, in practice this a frequently overlooked which causes a lot of problems later on. Hence, the research problem should be defined in a systematic manner, giving due weightage to all relating points. The technique for the purpose involves the undertaking of the following steps generally one after the other: (i) statement of the problem in a general way; (ii) understanding the nature of the problem; (iii) surveying the available literature (iv) developing the ideas through discussions; and (v) rephrasing the research problem into a working proposition.

A brief description of all these points will be helpful.

(i) **Statement of the problem in a general way:** First of all the problem should be stated in a broad general way, keeping in view either some practical concern or some scientific or intellectual this purpose, the researcher must immerse himself thoroughly in the subject matter concerning which he wishes to pose a problem. In case of social research, it is considered advisable to do some field observation and as such the researcher may undertake some sort of preliminary survey or what is often called *pilot survey. Then* the researcher can himself state the Problem or he can seek the guidance of the guide or the subject expert in accomplishing this task. Often the guide forth the problem in general terms, and it is then up to the researcher to narrow it down and the problem in operational terms. In case there is some directive from an organisational authority, the problem then can be stated accordingly. The problem stated in a broad general way may contain various ambiguities which must be resolved by cool thinking and rethinking over the problem. At the same time the feasibility of a particular solution has to be considered and the same should be kept in view while starting the problem.

**(ii) Understanding the nature of the problem:** The next step in defining the problem is to understand its origin and nature clearly. The best way of understanding the problem is to discuss it with those who first raised it in order to find out how the problem originally came about and with what objectives in view. If the researcher has stated the problem himself, he should consider once again all those points that induced him to make a general statement concerning the problem. For a better understanding of the nature of the problem involved, he can enter into discussion with those who have a good knowledge of the problem concerned or similar other problems. The researcher should also keep in view the environment within which the problem is to be studied and understood.

**(iii) Surveying the available literature:** Allavailable literature concerning the problem at hand must necessarily be surveyed and examined before a definition of the research problem IS given. This means that the researcher must be well-conversant with relevant theories in the field, reports and records as also all other relevant literature. He must devote sufficient time in reviewing of research already undertaken on related problems. This is done to find out what data and other materials, if any, are available for operational purposes. "Knowing what data are available often serves to narrow the problem itself as well as the technique that might be used". This would also help a researcher to know if there are certain gaps in the theories, or whether the existing theories applicable to the problem under study are inconsistent with each other, or whether the findings of the different studies do not follow a pattern consistent with the theoretical expectations and so on. All this will enable a researcher to take new strides in the field for furtherance of knowledge i.e., he can move up starting from the existing premise. Studies on related problems are useful for indicating the type of difficulties that may be encountered in the present study as also the possible analytical shortcomings. At times such studies may also suggest useful and even new lines of approach to the present problem.

**(iv) Developing the ideas through discussions:** (Discussion concerning a problem often produces useful information. Various new ideas can be developed through such an exercise. Hence, a researcher must discuss his problem with his colleagues and others who have enough experience in the same area or in working on similar problems. This is quite often known as an *experience survey. People* with rich experience are in a position to enlighten the researcher on different aspects of his proposed study and their advice and comments are usually invaluable to the researcher. They help him sharpen his focus of attention on specific aspects within the field. Discussions with such persons should not only be confined to the formulation of the specific problem at hand, but should also be concerned with the general approach to the given problem, techniques that might be used, possible solutions, etc.

**(v) Rephrasing the research problem:** Finally, the researcher must sit to rephrase the research

problem into a working proposition. Once the nature of the problem has been clearly understood, the environment (within which the problem has got to be studied) has been defined, discussions over the problem have taken place and the available literature has been surveyed and examined, rephrasing the problem into analytical or operational terms is not a difficult task. Through rephrasing, the researcher puts the research problem in as specific terms as possible so that it may become operationally viable and may help in the development of working hypotheses.

a. Technical terms and words or phrases, with special meanings used in the statement of the problem, should be clearly defined.
b. Basic assumptions or postulates (if any) relating to the research problem should be clearly stated.
c. A straight forward statement of the value of the investigation (i.e., the criteria for the selection of the problem) should be provided.
d. The suitability of the time-period and the sources of data available must also be considered by the researcher in defining the problem.
e. The scope of the investigation or the limits within which the problem is to be studied must be mentioned explicitly in defining a research problem

**AN ILLUSTRATION**

The technique of defining a problem outlined above can be illustrated for better understanding by taking an example as under:

Let us suppose that a research problem in a broad general way is as follows:

"Why is productivity in Japan so much higher than in India"?

In this form the question has a number of ambiguities such as: What sort of productivity is being referred to? With what industries the same is related? With what period of time the productivity is being talked about? In view of all such ambiguities the given statement or the question is much too general to be amenable to analysis. Rethinking and discussions about the problem may result in narrowing down the question to:

“What factors were responsible for the higher labour productivity of Japan's; manufacturing

industries during the decade 1971 to 1980 relative to India's manufacturing industries?" This latter version of the problem is definitely an improvement over its earlier version for the various ambiguities have been removed to the extent possible. Further rethinking and rephrasing might place the problem on a still better operational basis as shown below:

"To what extent did labour productivity in 1971 to 1980 in Japan exceed that of India in respect of 15 selected manufacturing industries? What factors were responsible for the productivity differentials between the two countries by industries?"

With this sort of formulation, the various terms involved such as 'labour productivity', 'productivity differentials', etc. must be explained clearly. The researcher must also see that the necessary data are available. In case the data for one or more industries selected are not available for the concerning time-period, then the said industry or industries will have to be substituted by other industry or industries. The suitability of the time-period must also be examined. Thus, all relevant factors must be considered by a researcher before finally defining a research problem.

## CONCLUSION

We may conclude by saying that the task of defining a research problem, very often, follows a sequential pattern-the problem is stated in a general way, the ambiguities are resolved, thinking and rethinking process results in a more specific formulation of the problem so that it may be a realistic one In terms of the available data and resources and is also analytically meaningful. All this results in a well defined research problem that is not only meaningful from an operational point of view, but is equally capable of paving the way for the development of working hypotheses and for means of solving the problem itself.

## RESEARCH DESIGN

### MEANING OF RESEARCH DESIGN

The formidable problem that follows the task of defining the research problem is the preparation of the design of the research project, popularly known as the "research design". Decisions regarding what, where, when, how much, by what means concerning an inquiry or a research study constitute a research design. "A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure”. In fact, the research design is the conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data. As such the design includes an outline of what the researcher will do from writing the hypothesis and its operational implications to the final analysis of da90re explicitly, the desing decisions happen to be in respect of:

(i)What is the study about?

(ii)Why is the study being made?
(iii)Where will the study be carried out?
(iv)What type of data is required?
(v)Where can the required data be found?
(vi)What periods of time will the study include?
(vii) What will be the sample design?
(viii) What techniques of data collection will be used?
(ix) How will the data be analysed?
(x) In what style will the report be prepared?

Keeping in view the above stated design decisions, one may split the overall research design into the following parts:

a. *The sampling design* which deals with the method of selecting items to be observed for the given study.
b. *the observational design* which relates to the conditions under which the observations are to be made;
c. *the statistical design* which concerns with the question of how many items are to be observed and how the information and data gathered are to be analysed; and
d. *the operational design* which deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

-

From what has been stated above, we can state the important features of a research design as under:

(i) It is a plan that specifies the sources and types of information relevant to the research problem.

(ii) It is a strategy specifying which a roach will be used for gathering and analysing the data.

(iii) It also includes the time and cost bud ets since most studies me done under these constraints.

In brief, research design must, at least, contain-(a) a clear statement of the research problem; (b) procedures and techniques to be used for gathering information; (c) the population to be stud and (d) methods to be used in processing and analysing data.

**NEED FOR RESEARCH DESIGN**

Research design is needed because it facilitates the smooth sailing of the various research operations thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time and money. Just as for better, economical and attractive construction of a house, we need a blueprint (or what is commonly called the map of the house) well thought out and prepared by expert architect. Similarly we need a research design or a plan in advance of data collection and analysis for our research project. Research design stands for advance planning of the methods to be adopted for collecting the relevant data and the techniques to be used in their analysis; keeping in view the objective of the research and the availability of staff, time and money. Preparation of research design should be done with great care as any error in it may upset the entire project. Research design, in fact, has a great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the entire edifice of the research work.

Even then the need for a well thought out research design is at times not realised by many. Theimportance which this problem deserves is not given to it. As a result many researches do not serve the purpose for which they are' undertaken. In fact, they may even give misleading conclusion. Thoughtlessness in designing the research project may result in rendering the research exercise futile. It is, therefore, imperative that an efficient and appropriate design must be prepared before starting research operations. The design helps the researcher to organize his ideas in a form whereby it will be possible for him to look for flaws and inadequacies. Such a design can even be given to others for their comments and critical evaluation. In the absence of such a course of action, it will be difficult for the critic to provide a comprehensive review of the proposed study.

**FEATURES OF A GOOD DESIGN:**

A good design is often characterized by adjectives like flexible, appropriate, efficient, econornical and so on. Generally, the design which minirnises bias and maxirnises the reliability of the data collected and analysed is considered a good design. The design which gives the smallest experimental error is supposed to be the best design in many investigations. Similarly, a design which yields maximal information and provides an opportunity for considering many different aspects of a problem is considered most appropriate and efficient design in respect of many research problems. Thus, the question of good design is related to the purpose or objective of the research problem and also with the nature of the problem to be studied. A design may be quite suitable in one case, but may be found wanting in one respect or the other in the context of some other research problem. One single design cannot serve the purpose of all types of research problems.

A research design appropriate for a particular research pr6blem, usually involves the consideration of the following factors:

(i) the means of obtaining information;
(ii) the availability and skills of the researcher and his staff, if any;
(iii) the objective of the problem to be studied;
(iv) the nature of the problem to be studied; and
(v) the availability of time and money for the research work

If the research study happens to be an exploratory or a formulative one, wherein the major emphasis is on discovery of ideas and insights, the research design most appropriate must be flexible enough to permit the consideration of many different aspects of a phenomenon. But when the purpose of a study is accurate description of a situation or of an association between variables (or in what are called the descriptive studies), accuracy becomes a major consideration and a research design which minimises bias and maxirnises the reliability of the evidence collected is considered a good design. Studies involving the testing of a hypothesis of a causal relationship between variables require a design which will permit inferences about causality in addition to the rninirnisation of bias and maximisation of reliability. But in practice it is the most difficult task to put a particular study in a particular group, for a given research may have in it elements of two or more of the functions of different studies. It is only on the basis of its primary function that a study can be categorised either an exploratory or descriptive or hypothesis-testing study and accordingly the choice of a research deign may be made in case of a particular study. Besides, the availability of time, money, skills of the h staff and the means of obtaining the information must be given due weightage while working the relevant details of the research design such as experimental design, survey design, sample and the like.

## IMPORTANT CONCEPTS RELATING TO RESEARCH DESIGN

,~ .

Before describing the different research designs, it will be appropriate to explain the various concepts to designs so that these may be better and easily understood.

**1. Dependent and independent variables:** A concept which can take on different quantitative is called a variable. As such the concepts like weight, height, income are all examples of variabes, Qualitative phenomena (or the attributes) are also quantified on the basis of the presence or absence of the concerning attribute(s). Phenomena which can take on quantitatively different values even in decimal points are called 'continuous variables': But all variables are not continuous. If they can only be expressed in integer values, they are non-continuous variables or in statistical language 'discrete variables'." Age is an example of continuous variable, but the number of children is an example of non-continuous variable. If one variable depends upon or is a consequence of the other variable, it is termed as a dependent variable, and the variable that is antecedent to the dependent variable is termed as an independent variable. For instance, if we say that height depends upon age, then height is a dependent variable and age isan independent variable. Further, if in addition to being dependent upon age, height also depends upon the individual's sex, then height is a dependent variable and age and sex are independent variables. Similarly, readymade films and lectures are examples of independent variables, whereas behavioural changes, occurring as a result of the environmental manipulations, are examples of dependent variables.

**2. Extraneous variable:** Independent variables that are not related to the purpose of the study, but may affect the dependent variable are termed as extraneous variables. Suppose the researcher wants to test the hypothesis that there is a relationship between children's gains in social studies achievement and their self-concepts. In this case self-concept is an independent variable and social studies achievement is a dependent variable. Intelligence may as well affect the social studies achievement, but since it is not related to the purpose of the study undertaken by the researcher, it will be termed as an extraneous variable. Whatever effect is noticed on dependent variable as a result of extraneous variable(s) is technically described as an 'experimental error'. A study must always be so designed that *the effect upon the dependent variable is attributed entirely to the independent variable(s), and not to some extraneous variable or variables.*

**3. Control:** One important characteristic of a good research design is to minimise the influence or effect of extraneous variable(s). The technical term 'control' is used when we design the study minimising the effects of extraneous independent variables. In experimental researches, the term 'control' is used to refer to restrain experimental conditions.

**4. Confounded relationship:** When the dependent variable is not free from the influence of extraneous variable(s), the relationship between the dependent and independent variables is said to be confounded by an extraneous variable(s).

**5. Research hypothesis**: When a prediction or a hypothesised relationship is to be tested by scientific methods, it is termed as research hypothesis. The research hypothesis is a predictive statement that relates an independent variable to a dependent variable. Usually a research hypothesis must contain, at least, one independent and one dependent variable.

Predictive statements which are not to' be objectively verified or the relationships that are assumed but not to be tested, are not termed research hypotheses.

**6. Experimental and non-experimental hypothesis-testing research:** When the purpo.se of research is to test a research hypothesis, it is termed as hypothesis-testing research. It can be of the experimental design or of the non-experimental design. Research in which the independent variable is manipulated is termed 'experimental hypothesis-testing research' and a research in which an independent variable is not manipulated is called 'non-experimental hypothesis-testing research'. For instance, suppose a researcher wants to study whether intelligence affects reading ability for a group of students and for this purpose he randomly selects 50 students and tests their intelligence and reading ability by calculating the coefficient of correlation between the two sets of scores. This is an example of non-experimental hypothesis-testing research because herein the independent variable, intelligence, is not manipulated. But now suppose that our researcher randomly selects 50 students from a group of students who are to take a course in statistics and then divides them into two groups by randomly assigning 25 to Group A, the usual studies programme, and 25 to Group B, the special studies programme. At the end of the course, he administers a test to each group in order to judge the effectiveness of the training programme on the student's performance-level. This is an example of experimental hypothesis-testing research because in this case the independent variable, viz., the type of training programme, is manipulated.

**7. Experimental and control groups:** In an experimental hypothesis-testing research when a group is exposed to usual conditions, it is termed a 'control group', but when the group is exposed to some novel or special condition, it is termed an 'experimental group'. In the above illustration, the Group A can be called a control group and the Group B an experimental group. If both groups A and B are exposed to special studies programmes, then both groups would be termed 'experimental groups.' It is possible to design studies which include only experimental groups or studies which include both experimental and control groups.

**8. Treatments:** The different conditions under which experimental and control groups are put are usually referred to as 'treatments'. In the illustration taken above, the two treatments are the usual studies programme and the special studies programme. Similarly, if we want to determine through an experiment the comparative impact of three varieties of fertilizers on the yield of wheat, in that case the three varieties of fertilizers will be treated as three treatments.

**9. Experiment:** The process of examining the truth of a statistical hypothesis, relating to some research problem, is known as an experiment. For example, we can conduct an experiment to examine the usefulness of a certain newly developed drug. Experiments can be of two types, viz., absolute experiment and comparative experiment. If we want to determine the impact of a fertilizer on the yield of a crop, it is a case of absolute experiment; but if we want to determine the impact of one fertilizer as compared to the impact of some other fertilizer, our experiment then will be termed as a comparative experiment. Often, we undertake comparative experiments when we talk of designs of experiments.

**10. Experimental unites):** The pre-determined plots or the blocks, where different treatments are used, are known as experimental units. Such experimental units must be selected (defined) very carefully.

## DIFFERENT RESEARCH DESIGNS

Different research designs can be conveniently described if we categorize them as: (1) research in case of exploratory research studies; (2) research design in case of descriptive and diagnostic research studies, and (3) research design in case of hypothesis-testing research studies.

We take up each category separately.

**1. Research design in case of exploratory research studies:** Exploratory research studies are also termedas formulative research studies. The main purpose of such studies is that of formulating a problem for more precise investigation or of developing the working hypotheses from an operational point of view. The major emphasis in such studies is on the discovery of ideas and insights. As such the research design appropriate for such studies must be flexible enough to provide opportunity for considering different aspects of a problem under study. Inbuilt flexibility in research design is needed because the research problem, broadly defined initially, is transformed into one with more precise meaning in exploratory studies, which fact may necessitate changes in the research procedure for gathering relevant data. Generally, the following three methods in the context of research design for such studies are talked about: (a) the survey of concerning literature; (b) the experience survey(c) the analysis of 'insight-stimulating' examples.

*The survey of concerning literature* happens to be the most simple and fruitful method of formulating precisely the research problem or developing hypothesis. Hypotheses stated by earlier workers may be reviewed and their usefulness be evaluated as a basis for further research. It may also .be considered whether the already stated hypotheses suggest new hypothesis. In this way the researcher should review and build upon the work already done by others, but in cases where hypotheses have not yet been formulated, his task is to review the available material for deriving the relevant hypotheses from it.

Besides, the bibliographical survey of studies, already made in one's area of interest may as well as made by the researcher for precisely formulating the problem. He should also make an attempt to apply concepts and theories developed in different research contexts to the area in which he is himself working. Sometimes the works of creative writers also provide a fertile ground for hypothesis - formulation and as such may be looked into by the researcher.

*Experience survey* means the survey of people who have had practical experience with the problem to be studied. The object of such a survey is to obtain insight into the relationships between variables and new ideas relating to the research problem. For such a survey people who are competent and can contribute new ideas may be carefully selected as respondents to ensure a representation of different types of experience. The respondents so selected may then be interviewed by the investigator. The researcher must prepare an interview schedule for the systematic questioning of informants. But the interview must ensure flexibility in the sense that the respondents should be allowed to raise issues and questions which the investigator has not previously considered. Generally, the experience - collecting interview is likely to be long and may last for few hours. Hence, it is often considered desirable to send a copy of the questions to be discussed to the respondents well in advance. This will also give an opportunity to the respondents for doing some advance thinking over the various issues involved so that, at the time of interview, they may be able to contribute effectively. Thus, an experience survey may enable the researcher to define the problem more concisely and help in the formulation of the research hypothesis. This survey may as well provide information about the practical possibilities for doing different types of research.

*Analysis of 'insight-stimulating' examples* is also a fruitful method for suggesting hypotheses for research. It is particularly suitable in areas where there is little experience to serve as a guide. This method consists of the Intensive study of selected instances of the phenomenon in which one is interested. For this purpose the existing records, if any, may be examined, the unstructured interviewing may take place, or some other approach may be adopted. Attitude of the investigator, the intensity of the study and the ability of the researcher to draw together diverse information into a unified interpretation are the main features which make this method an appropriate procedure for evoking insights.

Now, what sorts of examples are to be selected and studied? There is no clear cut answer to it. Experience indicates that for particular problems certain types of instances are more appropriate than others. One can mention few examples of 'insight-stimulating' cases such as the reactions of strangers, the reactions of marginal individuals, the study of individuals who are in transition from one stage to another, the reactions of individuals from different social strata and the like. In general, cases that provide sharp contrasts or have striking features are considered relatively more useful while adopting this method of hypotheses formulation.

Thus, in an exploratory of formulative research study which merely leads to insights or hypotheses, whatever method or research design outlined above is adopted, the only thing essential is that it must continue to remain flexible so that many different facets of a problem may be considered as and when they arise and come to the notice of the researcher.

**2. Research design in case of descriptive arid diagnostic research studies:** Descriptive research studies are those studies which are concerned with describing the characteristics of a particular individual, or of a group, whereas diagnostic research studies determine the frequency with which something occurs or its association with something else. The studies concerning whether certain variables are associated are examples of diagnostic research studies. As against this, studies concerned with specific predictions, with narration of facts and characteristics concerning individual, group or situation are all examples of descriptive research studies. Most of the social research comes under this category. From the point of view of the research design, the descriptive as well as diagnostic studies share common requirements and as such we may group together these two types of research studies. In descriptive as well as in diagnostic studies, the researcher must be able to define clearly, what he wants to measure and must find adequate methods. for measuring it along with a clear cut definition of 'population' he wants to study. Since the aim is to obtain complete and accurate information in the said studies, the procedure to be used must be carefully planned. The research design must make enough provision for protection against bias and must maximise reliability, with due concern for the economical completion of the research study. The design in such studies must be rigid and not flexible and must focus attention on the following:

a) Formulating the objective of the study (what the study is about and why is it being made?)
b) Designing the methods of data collection (what techniques of gathering data will be adopted?)
c) Selecting the sample (how much material will be needed?)

d) Collecting the data (where can the-required data be found and with what time period should the data be related?)
e) Processing and analysing the data.
f) Reporting the findings.

In a descriptive/diagnostic study the first step is to specify the objectives with sufficient precision to ensure that the data collected are relevant. If this is not done carefully, the study may not provide the desired information.

Then comes the question of selecting the methods by which the data are to be obtained. In other words, techniques for collecting the information must be devised. Several methods (viz., observation, questionnaires, interviewing, examination of records, etc.), with their merits and limitations, are available for the purpose and the researcher may user one or more of these methods which have been discussed in detail in later chapters. While designing data-collection procedure, adequate safeguards against bias and unreliability must be ensured. Whichever method is selected, questions must be well examined and be made unambiguous; interviewers must be instructed not to express their own opinion; observersmust be trained so that they uniformly record a given item of behaviour. It is always desirable to pre-test the data collection instruments before they are finally used for the study purposes. In other words, we can say that *"structured instruments"* are used in such studies.

In most of the descriptive/diagnostic studies the researcher takes out sample(s) and then wishes to make statements about the population on the basis of the sample analysis or analyses. More often than not, sample has to be designed. Different sample designs have been discussed in detail in a separate chapter in this book. Here we may only mention that the problem of designing samples should be tackled in such a fashion that the samples may yield accurate information with a minimum amount of research effort. Usually one or more forms of probability sampling, or what is often described as random sampling, are used.

To obtain data free from errors introduced by those responsible for collecting them, it is necessary to supervise closely the staff of field workers as they collect and record information. Checks may be set up to ensure that the data collecting staff perform their duty honestly and without prejudice. "As data are collected, they should be examined for completeness, comprehensibility, consistency and reliability.'?

The data collected must be processed and analysed. This includes steps like coding the interview replies, observations, etc.; tabulating the data; and performing several statistical computations. To the extent possible, the processing and analysing procedure should be planned in detail before actual work is started. This will prove economical in the sense that the researcher may avoid unnecessary labour such as preparing tables for which he later finds he has no use or on the other hand, re-doing some tables because he failed to include relevant data. Coding should be done carefully to avoid error in coding and for this purpose the reliability of coders needs to be checked. Similarly, the accuracy of tabulation may be checked by having a sample of the tables re-done. In case of mechanical tabulation the material (i.e., the collected data or information) must be entered on appropriate cards which is usually done by punching holes corresponding to a given code. The accuracy of punching is to be checked and ensured. Finally, statistical computations are needed and as such averages, percentages and various coefficients must be worked out. Probability and sampling analysis may as well be used. The appropriate statistical operations, along with the use of appropriate tests of significance should be carried out to safeguard the drawing of conclusions concerning the study.

Last of all comes the question of reporting the findings. This is the task of communicating the findings to others and the researcher must do it in an efficient manner. The layout of the report needs to be well planned so that all things relating to the research study may be well presented in simple and effective style.

Thus, the research design in case of descriptive/diagnostic studies is a comparative design throwing light on all points narrated above and must be prepared keeping in view the objective(s) of the study and the resources available. However, it must ensure the minimisation of bias and maximisation of reliability of the evidence collected. The said design can be appropriately referred to as a survey *design* since it takes into account all the steps involved in a survey concerning a phenomenon to studied.

The difference between research designs in respect of the above two types of research studies can be conveniently summarized in tabular form as under:

Table 3.1

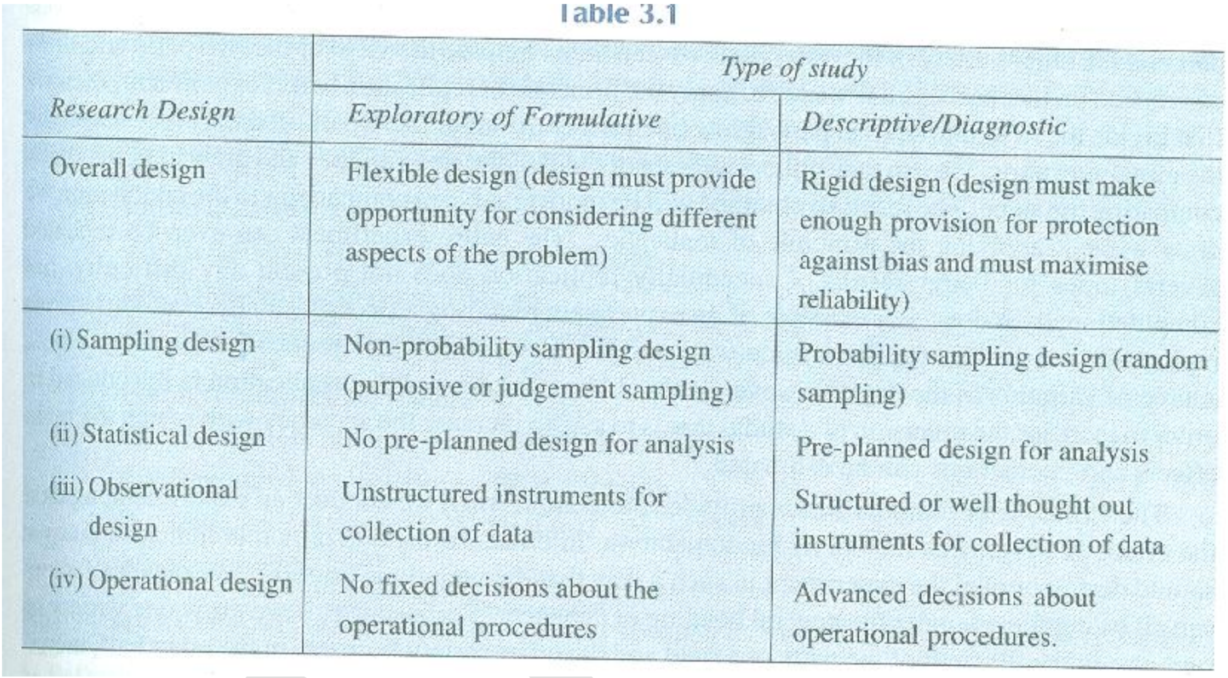


Table 3.1

| Research Design | Type of study | |
|---|---|---|
| | *Exploratory of Formulative* | *Descriptive/Diagnostic* |
| Overall design | Flexible design (design must provide opportunity for considering different aspects of the problem) | Rigid design (design must make enough provision for protection against bias and must maximise reliability) |
| (i) Sampling design | Non-probability sampling design (purposive or judgement sampling) | Probability sampling design (random sampling) |
| (ii) Statistical design | No pre-planned design for analysis | Pre-planned design for analysis |
| (iii) Observational design | Unstructured instruments for collection of data | Structured or well thought out instruments for collection of data |
| (iv) Operational design | No fixed decisions about the operational procedures | Advanced decisions about operational procedures. |

**3. Research design in case of hypothesis-testing research studies:** Hypothesis-testing research studies (generally known as experimental studies) are those where the researcher tests the hypotheses of causal relationships between variables. Such studies require procedures that will not only reduce bias and increase reliability, but will permit drawing inferences about causality. Usually experiments meet this requirement. Hence, when we talk of research design in such studies, we often mean the design of experiments.

Professor R.A. Fisher's name is associated with experimental designs. Beginning of such designs was made by him when he was working at Rothamsted Experimental Station (Centre for Agricultural Research in England). As such the study of experimental designs has its origin in agricultural research. Professor Fisher found that by dividing agricultural fields or plots into different blocks and then by conducting experiments in each of these blocks, whatever information is collected and inferences drawn from them, happens to be more reliable. This fact inspired him to develop certain experimental s for testing hypotheses concerning scientific investigations. Today, the experimental designs are being used in researches relating to phenomena of several disciplines. Since experimental designs originated in the context of agricultural operations, we still use, though in a technical sense, several terms of agriculture (such as treatment, yield, plot, block etc.) in experimental designs.

## POSSIBLE QUESTIONS
## UNIT V

### PART A (20 x 1 = 20 Marks)
### Question number 1 – 20 online examinations

### PART B (5 x 2= 20Marks)

1. Explain the steps necessary to carryout research effectively.
2. How do you define a research problem?
3. What are the problems encountered by the researchers in India?
4. What do you mean by research? Explain its significance in modern times.
5. Distinguish between Research methods and Research methodology.
6. Describe the different types of research, clearly pointing out the difference between an experiment and a survey.

### PART C (5 X 6 = 30 Marks)

7. Explain in detail the steps involved in research process
8. What are the scope and significance of research? Explain the research process.
9. How will you identify a research problem? Explain the steps involved in execution of research successfully.
10. Explain in detail the research process and the characteristics of good research.
11. Discuss research design and what are the features of good design?

**12.** Describe fully the techniques of defining a research problem.

**13.** What is research design? Discuss the basis of stratification to be employed in sampling public opinion on inflation.

**14.** Give your understanding of a good research design. Is single research design suitable in all research studies? If not, why?

**15.** Describe some of the important research designs used in experimental hypothesis – testing research study.

**16.** Write a short note on 'Experience Survey' explaining fully its utility in exploratory research studies.

**17.** What is research problem? Define the main issues which should receive the attention of the researcher in formulating the research problem.

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics and Research Methodology
UNIT V

**UNIT V**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 5 | All items in any field of inquiry constitute a -------------------- | **Population** | sample | total items | infinite samples | **Population** |
| 2 | 5 | Population census conducted once in a ----------------- | century | **decade** | fortnight | two decade | **decade** |
| 3 | 5 | A definite plan for obtaining a sample from a given population is known as | Research design | **sample design** | clear plan | accurate plan | **sample design** |
| 4 | 5 | ---------------- is determined before data are collected | Research design | **sample design** | clear plan | accurate plan | **sample design** |
| 5 | 5 | The following is an example of finite universe | Number of stars in the sky | listeners of a specific radio programme | Throwing of a dice | **the population of a city** | **the population of a city** |
| 6 | 5 | The following are infinite universe except | Number of stars in the sky | listeners of a specific radio programme | Throwing of a dice | **the population of a city** | **the population of a city** |
| 7 | 5 | One of the following is the example of infinite universe | the population of a city | the number of workers in a factory | the number of students in a college | **listeners of a specific radio programme** | **listeners of a specific radio programme** |
| 8 | 5 | A --------------- results from errors in the sampling procedures | **systematic bias** | sampling error | designing error | research bias | **systematic bias** |
| 9 | 5 | When the size of the sample increases sampling error ---------- | **decreases** | increases | no change | zero | **decreases** |
| 10 | 5 | To study the economic status of a town or village the sampling design used is | probability sampling | **non-probability sampling** | based on each item | based on items at random | **non-probability sampling** |
| 11 | 5 | Probability sampling is also known as | quota sampling | judgement sampling | **random sampling** | purposive sampling | **random sampling** |
| 12 | 5 | One of the following is not a non-probability sampling | deliberate sampling | purposive sampling | judgement sampling | **chance sampling** | **chance sampling** |
| 13 | 5 | The probability of selecting sample of size 3 from a finite population of 6 elements is | **1/20** | 3/6 | 2/6 | 1/6 | **1/20** |
| 14 | 5 | Selecting every $i^{th}$ item on a list is known as | **systematic sampling** | area sampling | multi-stage sampling | cluster sampling | **systematic sampling** |
| 15 | 5 | If a 4% sample is desired, the sample selection will be | one in every 10th item | **one in every 25th item** | one in every 50th item | one in every 4th item | **one in every 25th item** |
| 16 | 5 | If a population from which a sample is to be drawn doesnot constitute a homogenous group the technique is known as | **Stratified sampling** | area sampling | multi-stage sampling | cluster sampling | **Stratified sampling** |
| 17 | 5 | In stratified sampling the different sub populations are called as | sub datas | **strata** | substitutes | random samples | **strata** |
| 18 | 5 | The most efficient and an optimal design of complex random sampling is | **Stratified sampling** | area sampling | multi-stage sampling | cluster sampling | **Stratified sampling** |
| 19 | 5 | Using proportional allocation, the sample sizes for different strata are 15,9 and 6 respectively which is in proportion to the sizes of the stra | 3000:4000:1000 | **4000:2400:1600** | 2400:1600:4000 | 1800:2200:4000 | **4000:2400:1600** |
| 20 | 5 | If clusters happens to be some geographic subdivisions it is known as | Stratified sampling | **area sampling** | multi-stage sampling | cluster sampling | **area sampling** |
| 21 | 5 | ___________ is a process of mapping aspects of a domain onto other aspects of a range according to some rule of correspondence | strata | **measurement** | sub strata | clusters | **measurement** |
| 22 | 5 | The data which are numerical in name only and donot share any properties of the numbers we deal in ordinary arithmetic is | **nominal data** | ordinal data | interval data | ratio data | **nominal data** |
| 23 | 5 | In those situations,when we can't do anything except set up inequalities, we refer to the data as | nominal data | **ordinal data** | interval data | ratio data | **ordinal data** |
| 24 | 5 | On Mohs' scale number 5&2 refers to apatite and gypsum.In this 5 > 2 refers to | **apatite is harder than gypsum** | apatite is softer than gypsum | gypsum is softer than apatite | both are equally hard | **apatite is harder than gypsum** |
| 25 | 5 | On Mohs' scale the numbers 6&9 refers to feldspar and sapphire respectively in this 6 < 9 refers to | feldspar is the most hardest | sapphire and feldspar are equally hard | **feldspar is softer than sapphire** | sapphire is softer than feldspar | **feldspar is softer than sapphire** |
| 26 | 5 | is simply a system of assigning number symbols to events in order to label them | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 27 | 5 | The assignment of number of basketball players in order to identify them is the usual example of | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 28 | 5 | In nominal scale the measure of central tendency used is | mean | median | **mode** | quartiles | **mode** |
| 29 | 5 | Generally used measure of dispersion for nominal scales is | standard deviation | mean deviation | quartile deviation | **no measure of dispersion** | **no measure of dispersion** |
| 30 | 5 | The most common test of statistical significance that can be utilized in nominal scales is | F-test | t-test | Z-test | **chi-square test** | **chi-square test** |
| 31 | 5 | For the measure of correlation ------------ can be worked out for nominal scales | **contingency coefficient** | scatter diagram | Karl-pearson's coefficient of correlation | graphs | **contingency coefficient** |
| 32 | 5 | is the least powerful level of measurement | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 33 | 5 | The lowest level of the ordered scale that is commonly used is the ----------------- | nominal scale | **ordinal scale** | interval scale | ratio scale | **ordinal scale** |
| 34 | 5 | A students rank in his graduation class involves the use of an ------------- | nominal scale | **ordinal scale** | interval scale | ratio scale | **ordinal scale** |
| 35 | 5 | Multiplication and division can only be used with this scale but not with other scales. | nominal scale | ordinal scale | interval scale | **ratio scale** | **ratio scale** |
| 36 | 5 | Measures of central tendency used in ratio scale is | median | mode | arithmetic mean | **geometric and harmonic mean** | **geometric and harmonic mean** |
| 37 | 5 | Researchers in physical sciences have the advantage to describe variables in | nominal scale | ordinal scale | interval scale | **ratio scale** | **ratio scale** |
| 38 | 5 | Researchers in behavioural sciences have the advantage to describe variables in | nominal scale | ordinal scale | **interval scale** | ratio scale | **interval scale** |
| 39 | 5 | The most precise type of scale is | nominal scale | ordinal scale | interval scale | **ratio scale** | **ratio scale** |
| 40 | 5 | refers to the extent to which a test measures what we actually wish to measure | **Validity** | Reliability | Practicality | speed | **Validity** |
| 41 | 5 | has to do with the accuracy and precision of a measurement procedure | validity | **reliability** | practicality | speed | **reliability** |
| 42 | 5 | is concerned with a wide range of factors of economy,convenience and interpretability | validity | reliability | **practicality** | speed | **practicality** |
| 43 | 5 | If the instrument contains a representative sample of the universe,the ---------- is good | Criterion related validity | **content validity** | construct validity | obstruct validity | **content validity** |
| 44 | 5 | enables the researcher to study the perceptual structure of a set of stimuli and the cognitive processes underlying the developm | rational scaling | nominal scaling | **multi-dimentional scaling** | interval scale | **multi-dimentional scaling** |
| 45 | 5 | __________ describes the procedures of assigning numbers to various degrees of opinion,attitude and other concepts | **scaling** | sampling | validity | reliability | **scaling** |
| 46 | 5 | Categorical scales are also known as | nominal scale | ordinal scale | interval scale | **rating scale** | **rating scale** |
| 47 | 5 | Comparative scales are also known as | nominal scale | **ranking scale** | interval scale | ratio scale | **ranking scale** |
| 48 | 5 | Classification without indicating order, distance or unique origin is | **nominal scale** | ordinal scale | interval scale | ratio scale | **nominal scale** |
| 49 | 5 | _________ indicates magnitude relationship of 'more than' or 'less than' but indicate no distance or unique origin | nominal scale | **ordinal scale** | interval scale | ratio scale | **ordinal scale** |
| 50 | 5 | Which scales have both order and distance values but no unique origin | nominal scale | ordinal scale | **interval scale** | ratio scale | **interval scale** |
| 51 | 5 | The most widely used scale construction technique developed on ad hoc basis is | **arbitrary approach** | item analysis approach | consensus approach | factor scales | **arbitrary approach** |
| 52 | 5 | Qualitative description of a limited number of aspects of a thing or of traits of a person is | nominal scale | ordinal scale | interval scale | **rating scale** | **rating scale** |
| 53 | 5 | The scaling technique in the form of 'always-often-occasionally-rarely-never' is | nominal scale | ordinal scale | interval scale | **rating scale** | **rating scale** |
| 54 | 5 | Greater sensitivity of measurement is achieved in rating scale if there is | **more points on a scale** | less points on a scale | only two points | zero points | **more points on a scale** |
| 55 | 5 | If the respondents are either easy raters or hard raters, the type of error occurs is | error of central tendency | **error of leniency** | error of hallo effect | error of reliability | **error of leniency** |
| 56 | 5 | When raters are reluctant to give extreme judgments, the result is the | **error of central tendency** | error of leniency | error of hallo effect | error of reliability | **error of central tendency** |
| 57 | 5 | Composite standard method of paired comparision is given by | **J.P.Guilford** | E.L.Thorndike | M.D.Thurston | R.A.Fisher | **J.P.Guilford** |
| 58 | 5 | It is easy to construct Likert-type scale in comparison to Thurstone type because | only written examination is done | **can be performed without a panel of judges** | only oral questions are asked | It is performed with a panel of judges | **can be performed without a panel of judges** |
| 59 | 5 | An attempt to measure the psychological meanings of an object to an individual is | **semantic differential scale** | ordinal scale | interval scale | ratio scale | **semantic differential scale** |
| 60 | 5 | In which section of dissertation ,we can give our suggestions? | Introduction | Result | **Discussion** | Su mmary | **Discussion** |

# KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)
(Established Under Section 3 of UGC Act 1956)
Coimbatore - 641021.
(For the candidates admitted from 2017 onwards)

**DEPARTMENT OF BIOCHEMISTRY**

**M.Sc., Biochemistry** **2018-2020**

**Semester III**

**18BCP305A** **CORE ELECTIVE – III** **4H-4C**

## BIOSTATISTICS AND RESEARCH METHODOLOGY

**Instruction hours/week: L:4 T:0 P:0** **Marks:** Internal: **40** External:**60** Total:**100**

**End Semester Exam:** 3 Hours

### Course objectives

- To learn the various methods of collecting datas and their interpretations.
- To understand the various methods of central tendency and dispersion and to get clear cut idea of using specific method in a particular situation.
- To learn the correlation and regression analysis to compare the datas to interpret the results
- To learn the sampling distribution and the test of significance and applying of specific statistical tool to interpret the results.

### Course outcomes (CO's)

After learning this course the students will be

1. Familiar with the research process and how to plan for a research.
2. Able to interpret the result findings in a easy and use appropriate statistical methods
3. Able to analyse the data to say whether it is significant or not.

### UNIT I: Introduction to Biostatistics

Definition and scope of Biostatistics- Statistical survey-organizing , planning and executing the survey; Sources of data-primary and secondary data, Collection of data- Methods of data collection; Classification and tabulation of data- Graphical and diagrammatic representation.

Measures of central tendency – Arithmetic mean, median, mode, quartiles, deciles and percentiles. Measures of dispersion- Range, quartile deviation, mean deviation and standard deviation, Coefficient of variation.

## UNIT II: Correlation and Regression

Correlation: Meaning and definition - Scatter diagram –Karl Pearson’s correlation coefficient.  Rank correlation.
Regression: Regression in two variables – Regression coefficient problems – uses of regression**.**

## UNIT III: Probability

Probability- Definition, concepts, theorems (proofs of the theorems not necessary) and calculations of probability-simple problems, theoretical distributions-Binomial, Poisson and Normal distribution – simple problems

## UNIT IV: Sampling distribution and test of significance

Sampling distribution and test of significance – concepts of sampling, testing of hypothesis, errors in hypothesis testing, standard errors and sampling distribution– Student’s t test, F-test, Chi square test - goodness of fit. Analysis of variance – one way and two way classification. CRD, RBD Designs. Duncan’s multiple range tests.

## UNIT V: Introduction to Research

Research: Scope and significance – Types of Research – Research Process – Characteristics of good research – Problems in Research – Identifying research problems. Research Designs – Features of good designs.
Sources of information: Journals, eJournals, books, biological abstracts, preparation of index cards, review writing, article writing – structure of article, selection of journals for publication – Impact factor – citation index and H index. Proposal writing for funding. IPR and patenting. Concepts and types.

## SUGGESTED READINGS

1. Gupta, S.P., (2007). Statistical Methods, Sultan Chand & Co, New Delhi.

2. Kothari, C.R., (2009). Research Methodology – Methods and Techniques, 3rd edition, New Age International Pvt. Ltd, New Delhi.

3. Sundar Rao, P.S.S., and Richard, J., (2006). Introduction to Biostatistics and ResearchMethods, PHI Publication, New Delhi.

4. Sandhu, T., (1990). Research Techniques in Biological Sciences, Anmol Publishers, New Delhi.