Semester - III



KARPAGAM ACADEMY OF HIGHER EDUCATION

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

Coimbatore – 641 021.

SYLLABUS

I7MBP304 BIOSTATISTICS AND RESEARCH METHODOLOGY 4H – 4C

Instruction Hours / week: L: 4 T: 0 P: 0 Marks: Internal: 40 External: 60 Total: 100

End Semester Exam: 3 Hours

SCOPE:

This course has been intended to provide the learner insights into helpful areas of Statistics which plays an essential role in present, future use and applications of Biology.

OBJECTIVE:

Students get an idea about collection, interpretation and presentation of statistical data.

UNIT-I

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.

UNIT – II

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

Test of significance: Tests based on Means only-Both Large sample and Small sample tests - Chi square test - goodness of fit. Analysis of variance – one way and two way classification. CRD, RBD Designs.

$\mathbf{UNIT} - \mathbf{IV}$

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.

$\mathbf{UNIT} - \mathbf{V}$

Sampling Design: Meaning – Concepts – Steps in sampling – Criteria for good sample design. Scaling measurements – Techniques – Types of scale.

SUGGESTED READINGS

REFERENCES

1 .Jerrold H. Zar. (2003). *Biostatistical Analysis*. (4th ed.). Pearson Education (P) Ltd, New Delhi.

2. Kothari. C.R. (2004). *Research Methodology – Methods and Techniques*. (2nd ed.). New Age International Pvt. Ltd, New Delhi.



KARPAGAM (Deemed to be University Established Under Section 3 of UGC Act 1956) (Deemed to be University) (Deemed to be University Established Under Section 3 of UGC Act 1956) Coimbatore – 641 021.

DEPARTMENT OF MATHEMATICS

LESSON PLAN

Name of the faculty: Y.Sangeetha

Class: II M.Sc Microbiology

Subject: Biostatistics and Research Methodology

Subject Code: 17MBP304

S.No	Lecture	Topic to be covered	Support Material						
1	Duration								
	Unit – I								
1.	1	Definitions, Scope of Biostatistics, Variables in biology	R1: Ch 1: Pg: 1-2						
2.	1	Collection, Classification and tabulation of data	R3: Ch 4: Pg: 31, Ch 6: Pg: 56-59,73-76						
3.	1	Diagrammatic representation of data: Bar diagram and Pie diagram	R3: Ch 7: Pg: 84-97						
4.	1	Graphical representation of data: Histogram and Ogives	R3: Ch 8: Pg: 101-110						
5.	1	Measures of Central tendency: Problems on arithmetic Mean	R3: Ch 9: Pg: 121-133						
6.	1	Problems on Median	R3: Ch 9: Pg: 146-158						
7.	1	Problems on Mode	R3: Ch 9: Pg: 146-158						
8.	1	Measures of dispersion: Problems on Range and standard Deviation	R3: Ch 9: Pg: 234-239, 249-259						
9.	1	Problems on Coefficient of Variation	R3: Ch 10: Pg: 272-282						
10.	1	Recapitulation and discussion of possible question							
Total N	No of Hours	s Planned For Unit I – 10 hours							
		Unit – II							

1.	1	Correlation: Meaning & definition,	R3: Ch 12: Pg: 359-364
2.	1	Types of correlation	R3: Ch 12: Pg:361-362
3.	1	Scatter diagram	R3: Ch 12:Pg:363-364
4.	1	Problems on Karl Pearson's correlation coefficient	R3: Ch 12: Pg: 366-375
5.	1	Problems on Rank correlation	R3: Ch 12: Pg: 385-389
6.	1	Regression: Uses of regression	R3: Ch 13: Pg: 425-427
7.	1	Problems on Regression in two variables	R3: Ch 13: Pg: 431-443
8.	1	Problems in Regression coefficient	R3: Ch 13: Pg: 445-460
9.	1	Recapitulation and discussion of possible question	
Total N	No of Hour	s Planned For Unit II – 9 hours	
		Unit – III	
1.	1	Tests of significance: Tests based on Means	R2: Ch 9: Pg: 195-202
2.	1	Problems on Large sample tests	R2: Ch 9: Pg: 202-210
3.	1	Problems on small sample tests	R2: Ch 9: Pg: 211-213
4.	1	Problems on Chi square test	R2: Ch 10: Pg: 233-234
5.	1	Chi square test -goodness of fit	R2: Ch 10: Pg: 235-236
6.	1	Analysis of variance – one way classification	R2: Ch 11: Pg: 256-262
7.	1	Analysis of variance - two way	R2: Ch 11: Pg: 263-270
8.	1	CRD designs	R2: Ch 3: Pg: 42-44
9.	1	RBD designs	R2: Ch 3: Pg: 44-45
10.	1	Recapitulation and discussion of possible question	
Total	No of Hou	rs Planned For Unit III–10 hours	
		Unit – IV	
1.	1	Research: Scope	R2: Ch 1: Pg: 1-2
2.	1	Significance of Research	R2: Ch 1: Pg: 1-2
3.	1	Types of Research	R2: Ch 1: Pg: 2-3

4.	1	Research Process	R2: Ch 1: Pg: 10-19
5	1		D2. C1. 1. D.: 20.21
5.	1	Characteristics of good research	R2: Ch 1: Pg: 20-21
6.	1	Problems in research	R2: Ch 1: Pg: 24-26
7.	1	Identifying research problems	R2: Ch 1: Pg: 26-29
8.	1	Research designs	R2: Ch 1: Pg: 31-33
9.	1	Features of good designs	R2: Ch 1: Pg: 31-33
10.	1	Recapitulation and discussion of possible question	
Total N	No of Ho	ours Planned For Unit IV – 10 hours	
		Unit – V	
1.	1	Sampling design: Meaning and concepts	R2: Ch 4: Pg: 55-56
2.	1	Steps in sampling	R2: Ch 4: Pg: 56-58
3.	1	Criteria for good sample designs	R2: Ch 4: Pg: 56-58
4.	1	Scaling measurements	R2: Ch 5: Pg: 71-72
5.	1	Techniques and Types of scale	R2: Ch 5: Pg: 75-92
6.	1	Recapitulation and discussion of possible questions	
7.	1	Discussion of previous ESE question papers	
8.	1	Discussion of previous ESE question papers	
9.	1	Discussion of previous ESE question papers	
Total N	o of Ho	urs Planned For Unit V–9 hours	

REFERENCES:

R1. Jerrold H. Zar. (2003). Biostatistical Analysis (4th ed.). Pearson Education (P) Ltd, New Delhi.

R2. Kothari.C.R. (2004). Research Methodology – Methods and Techniques. (2nd ed). New Age International Pvt. Ltd, New Delhi.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IBATCH-2017-2019

<u>UNIT-I</u>

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.

STATISTICS

Statistics - Meaning and Definitions

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.

Population: A population is any entire collection of people, animals, plants or things on which we may collect data. It is the entire group of interest, which we wish to describe or about which we wish to draw conclusions. In the above figure the life of the light bulbs manufactured say by GE, is the concerned population.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

Qualitative and Quantitative Variables: Any object or event, which can vary in successive observations either in quantity or quality is called a "variable." Variables are classified accordingly as quantitative or qualitative. A qualitative variable, unlike a quantitative variable does not vary in magnitude in successive observations. The values of quantitative and qualitative variables are called "Variates" and" Attributes", respectively.

Variable: A characteristic or phenomenon, which may takes different values, such as weight, gender since they are different from individual to individual.

Randomness: Randomness means unpredictability. The fascinating fact about inferential statistics is that, although each random observation may not be predictable when taken alone, collectively they follow a predictable pattern called its distribution function. For example, it is a fact that the distribution of a sample average follows a normal distribution for sample size over 30. In other words, an extreme value of the sample mean is less likely than an extreme value of a few raw data.

Sample: A subset of a population or universe.

An Experiment: An experiment is a process whose outcome is not known in advance with certainty.

Statistical Experiment: An experiment in general is an operation in which one chooses the values of some variables and measures the values of other variables, as in physics. A statistical experiment in contrast is an operation in which one takes a random sample from a population and infers the values of some variables. For example, in a survey, we "survey" i.e. "look at" the situation without aiming to change it, such as in a survey of political opinions. A random sample from the relevant population provides information about the voting intentions.

In order to make any generalization about a population, a random sample from the entire population; that is meant to be representative of the population, is often studied. For each population, there are many possible samples. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a set of data would give information about the overall population mean μ .

It is important that the investigator carefully and completely defines the population before collecting the sample, including a description of the members to be included.

Example: The population for a study of infant health might be all children born in the U.S.A. in the 1980's. The sample might be all babies born on 7^{th} of May in any of the years.

An experiment is any process or study which results in the collection of data, the outcome of which is unknown. In statistics, the term is usually restricted to situations in which the researcher has control over some of the conditions under which the experiment takes place.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

Example: Before introducing a new drug treatment to reduce high blood pressure, the manufacturer carries out an experiment to compare the effectiveness of the new drug with that of one currently prescribed. Newly diagnosed subjects are recruited from a group of local general practices. Half of them are chosen at random to receive the new drug, the remainder receives the present one. So, the researcher has control over the subjects recruited and the way in which they are allocated to treatment.

Design of experiments is a key tool for increasing the rate of acquiring new knowledge. Knowledge in turn can be used to gain competitive advantage, shorten the product development cycle, and produce new products and processes which will meet and exceed your customer's expectations.

Primary data and Secondary data sets: If the data are from a planned experiment relevant to the objective(s) of the statistical investigation, collected by the analyst, it is called a Primary Data set. However, if some condensed records are given to the analyst, it is called a Secondary Data set.

Random Variable: A random variable is a real function (yes, it is called" variable", but in reality it is a function) that assigns a numerical value to each simple event. For example, in sampling for quality control an item could be defective or non-defective, therefore, one may assign X=1, and X = 0 for a defective and non-defective item, respectively. You may assign any other two distinct real numbers, as you wish; however, non-negative integer random variables are easy to work with. Random variables are needed since one cannot do arithmetic operations on words; the random variable enables us to compute statistics, such as average and variance. Any random variable has a distribution of probabilities associated with it.

Probability: Probability (i.e., probing for the unknown) is the tool used for anticipating what the distribution of data should look like under a given model. Random phenomena are not haphazard: they display an order that emerges only in the long run and is described by a distribution. The mathematical description of variation is central to statistics. The probability required for statistical inference is not primarily axiomatic or combinatorial, but is oriented toward describing data distributions.

Sampling Unit: A unit is a person, animal, plant or thing which is actually studied by a researcher; the basic objects upon which the study or experiment is executed. For example, a person; a sample of soil; a pot of seedlings; a zip code area; a doctor's practice.

Parameter: A parameter is an unknown value, and therefore it has to be estimated. Parameters are used to represent a certain population characteristic. For example, the population mean μ is a parameter that is often used to indicate the average value of a quantity.

Within a population, a parameter is a fixed value that does not vary. Each sample drawn from the population has its own value of any statistic that is used to estimate this parameter. For example, the mean of the data in a sample is used to give information about the overall mean μ in the population from which that sample was drawn.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

Statistic: A statistic is a quantity that is calculated from a sample of data. It is used to give information about unknown values in the corresponding population. For example, the average of the data in a sample is used to give information about the overall average in the population from which that sample was drawn.

A statistic is a function of an observable random sample. It is therefore an observable random variable. Notice that, while a statistic is a "function" of observations, unfortunately, it is commonly called a random "variable" not a function.

It is possible to draw more than one sample from the same population, and the value of a statistic will in general vary from sample to sample. For example, the average value in a sample is a statistic. The average values in more than one sample, drawn from the same population, will not necessarily be equal.

Statistics are often assigned Roman letters (e.g. \bar{x} and s), whereas the equivalent unknown values in the population (parameters) are assigned Greek letters (e.g., μ , σ).

The word estimate means to esteem, that is giving a value to something. A statistical estimate is an indication of the value of an unknown quantity based on observed data.

More formally, an estimate is the particular value of an estimator that is obtained from a particular sample of data and used to indicate the value of a parameter.

Example: Suppose the manager of a shop wanted to know μ , the mean expenditure of customers in her shop in the last year. She could calculate the average expenditure of the hundreds (or perhaps thousands) of customers who bought goods in her shop; that is, the population mean μ . Instead she could use an estimate of this population mean μ by calculating the mean of a representative sample of customers. If this value were found to be Rs.25, then Rs.25 would be her estimate.

Functions

The following are the important functions of statistics.

- * Collection
- * Numerical Presentation
- * Diagrammatic Presentation
- * Condensation
- * Comparison
- * Forecasting

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IBATCH-2017-2019

- * 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 & SECONDARY DATA

Methods of collecting primary data

- Primary data are those which are called as first information .
- There are five types of data.
- 1.Direct personal interview
- 2. Indirect oral interview

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

- 3. Information through agencies
- 4.Mailed Questionnaire method
- 5. Schedules sent through Enumerators

Secondary Data

- Secondary data can be compiled from published sources and unpublished sources
- Published source:
- The government collect some datas and they publish it .
- Eg: Census of India
- Unpublished Sources:
- All the statistical data are not published.
- Eg:Bank collect certain particulars from the loaner.

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.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology

COURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

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 allinclusive classes are classes that together contain all the data.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology **COURSE NAME: Biostatistics and Research Methodology** COURSE CODE: 17MBP304 BATCH-2017-2019

<u>UNIT: I</u>

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	а	sample	:	of	10	stude	ents	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 \	N=(25.3 -	18.1)/4 = 1	1.8 which	is rounded	l-up to	2. T	he freque	ncy table	is as foll	ows:

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.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

Rules for construction

* Title

* Proportion between width and height

* Size

- * scale
- * Index
- * Suitable Diagram
- * Simplicity
- * Neatness
- * Foot-Note and source
- * Identification numbers.

Graphic Presentation of Data

- Use initial exploratory data-analysis techniques to produce a pictorial representation of the data.
- Resulting displays reveal patterns of behavior of the variable being studied.
- The method used is determined by the type of data and the idea to be presented.
- No single correct answer when constructing a graphic display.

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.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: I BATCH-2017-2019

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

Histograms are typically used to display the frequency distribution of a continuous variable. The Y-axis indicates the frequency of occurance of the value/s specified on the X-axis. The X-axis represents an ordered range of values. E.g.



Notes: From the above histogram we might conclude that a score of 20 occurred three times, however, in most instances the box identified with a score of 20 covers a range of possible scores, in this case those scores being greater than 15 but not greater than 25. It should also be noted that in the case of a histogram there is no gap between the bars.

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.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

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.

Bar Charts

Simple Bar charts are often used to compare data from two or more groups. The Y-axis is typically interval level data representing one variable or measure, and the X-axis is used to denote different categories or nominal data (for example comparing men and women's average income, or comparing two or more participant's word fluency scores. The height of the bar indicates the 'group's' value on the Y-axis.



CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

Notes: The chart has a title (Annual Income by Gender) indicating that annual income is being 'broken down' by gender. The Y-axis indicates the measure of interest (Mean Income) and the units of measurement (£k). The categorisation variable (Gender) is indicated and the levels of this variable are also indicated (Men & Women). All charts and tables should include a figure or table number to identify them subsequently. Note that the actual values associated with each category have also been added to aid clarity.

Clustered bar charts are typically used to illustrate the distribution of a continuous variable (e.g. income) across two or more categorical variables. E.g.



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

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IBATCH-2017-2019

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 X

Individual series

—

 $X = \Sigma X / N$

Example

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

. 36, 40, 250, 42, 8, 000, 75, 10, 70, 36

Solution

Here N=10

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

X =1061 / 10 =106.1

Discrete series

—

 $X = \Sigma f X / \Sigma f$

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology **COURSE NAME: Biostatistics and Research Methodology** COURSE CODE: 17MBP304 UNIT: I BATCH-2017-2019 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 Х f fΧ 2 10 20 75 3 25 4 30 120 5 25 125 6 <u>10</u> 60 $\Sigma f = 100 \Sigma f X = 400$

X = 400 / 100 = 4.

Continuous series

_

 $X = \Sigma fm / \Sigma f$ where m represents the mid value.

Midvalue= (upper boundary + lower boundary)/2.

LASS: II M.S	c Microb	iology	COUR	SE NAI	ME: Bio	ostatistic	cs and Research Methodology
COURSE CO	DE: 17M	BP304	UN	IIT: I		BATCH-2	2017-2019
Example							
Calculate the n	nean for tl	ne followi	ng.				
Marks	: 20-30	30-4	40-50	50-60	60-70	70-80	
No.of student	: 5	8	12	15	6	4	
Solution							
C.I	f	m fr	n				
20-30	5	25 2	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 <u>3</u>	<u>00</u>				
Σ f =	= 50	Σ f m= 2	460				
-							
X = 2460 /	50 = 49.2						

Median

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

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IBATCH-2017-2019

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

Discrete series

Position of the median = $(N+1)/2^{\text{th}}$ item.

Example

LASS: II M.Sc Microbiol	ogy COURSE NAM	E: Biostatistics and Research Methodology
COURSE CODE: 17MBP	304 UNIT: I	BATCH-2017-2019
Find the median for the follo	wing.	
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 <u>3</u> 40		
N= 40		
(N+1)/2 = (40+1)/2 = 20.5	5 th item	
= (20 th item +21 st i	item) /2 =(15+17) /2	
= 16.		
Continuous sorias		
continuous series		
M = L + [((N/2) - c.f) x]	<u>i</u>]	
f.		
Where L- lower boundary, f-	frequency, i-size of class int	erval,
c.f- cumulative frequency.		

COURSE CODE: 17MBP.	304 UNIT	°∙Т ВАТСН-2	
		DATCH 2	017-2019
Example			
Calculate the median height	given below.		
Height : 145-150 1	50-155 155-160	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 <u>1</u>	30		
Σ f = 30			
Position of the median = $N/2$	2 th item = 30 / 2 =15.		
M = L+ <u>[((N/2) –c.f) x i</u>]		
f.			
ï			
= 155+ <u>[(15-7)x5]</u>	= 155+(40/10) = 15	59.	
10			
Mode			

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

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IBATCH-2017-2019

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



Here mode represents highest frequency .

Mode =20

Continuous series

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

LASS: II M.Sc Microbiology	COURSENAM	E: Biostatistics and Research Methodology
COURSE CODE: 17MBP304	UNIT: I	BATCH-2017-2019
Where L- lower boundary , f ₁ -frequ	uency of the modal cl	ass, f_0 – frequency of the preceeding modal class,
f_{2} - frequency of the succeeding mo	dal class, i-size of cla	ss interval , c.f- cumulative frequency.
2 1 7 0		, , ,
Example		
Determine the mode		
Marks : 0-10 10-20 20	0-30 30-40 40-50	
No.of student : 5 20	35 20	12
Solution		
Marks No. of student		
0-10 5		
0-10 5		
10-20 20		
20-30 35		
30-40 20		
40-50 12		Y
$Z = L + [i(f_1 - f_0) / (2f_1 - f_0 - f_2)]$		
= 20+[10(35-20)/(2(35)-20-20)]	= 20+5	
= 25.		
For stated set of		
Empirical relation		

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

Meaure 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)

LASS: II M.So	Microbi	ology	COURSE NAME: Biostatistics and Research Methodolog			
COURSE CO	DE: 17MI	3P304	UNIT: I	BATCH-2	2017-2019	
Continuous	series					
Range = $L - S$,	where L-M	lidvalue of	f largest boundary	/ & S-Midvalue of	smallest boundary	
Example						
Calculate the ra	nge.					
Marks	: 20-30	30-40 40	0-50 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 8	k S=25					
Range - I - 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 $\boldsymbol{\sigma}$.

LASS:] COUR	II M.Sc Microbiology SE CODE: 17MBP304	COURSE NAM UNIT: I	E: Biostatistics and Research Methodology BATCH-2017-2019
ndivi	dual series		
Example	$\sigma = \sqrt{(\Sigma x^2 / N) - (\Sigma x / P)}$	N) ²	
Calculat	e standard deviation for th	e following data.	
40,41,4	5,49,50,51,55,59,60,60.		
Solutior	ı		
х	X ²		
40	1600		
41	1681		
45	2025		
49	2401		
50	2500		
51	2601		
55	3025		
59	3481		
60	3600		
<u>60</u>	<u>3600</u>		
510 X	E x ² = 26504		
σ=√ (Σ	$\Sigma x^2 / N) - (\Sigma x / N)^2$		

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: I BATCH-2017-2019 = 7.09

- 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

Х	f		fx	x ²		b	κ ²
0	1		0	()		0
1	2		2	1	1		2
2	4		8		4		16
3	3		9		9		27
4	0		0		16		0
5	<u>2</u>		<u>10</u>		25		<u>50</u>
	$\Sigma f = 12$	Σ fx=	29		Σf	x ² =	95

LASS: II WI.SC WICTODIOlogy	COURSE NAME: Biostatistics and Research Meth				
COURSE CODE: 17MBP304	UNIT: I	BATCH-2017-2019			
σ = $\sqrt{(\Sigma fx^2 / Σ f)}$ –($\Sigma fx / Σ f)^2$					
$=\sqrt{(95/12) - (29/12)^2}$					
= 1.44					
Continuous series					
$\sigma = \sqrt{(\Sigma \text{ fm}^2 / \Sigma \text{ f}) - (\Sigma \text{ fm} / \Sigma \text{ f})^2}$					
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 <u>2025</u>				

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

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

= 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:

Part-B(6 mark)

1.Explain the methods of collection of data?

2. Explain about the Classification of data

CLASS: II M.Sc Microbiology			gy 04	COUR	SE NA	ME: Bi	ostatisti	statistics and Research Methodology			
COURSECO		<u>IDP5</u>	04				DAICH-	2017-20	19		
3. Calculate the	e standard	l devia	ation for	the follo	owing da	ata.					
Annua	l Profit (F	Rs):	20-40)	40-60)	60-80		80-100		
No. Of	Banks:		10		14		25		48		
Annua	l Profit (F	Rs):	100-1	20	120-1	40	140-1	60			
No. Of	Banks:		33		24	Ļ	16	5			
4. Calculate the	e mean an	nd mee	lian for	the follo	wing da	ta.					
Class In	terval: 20	0-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100		
Frequen	cy:	4	14	20	51	32	17	6	4		
5. Calculate the	e median	for the	e followi	ng data.							
Class I	nterval:	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16		
Freque	ncy :	45	50	45	70	30	25	20	18		
6. Draw a suita	ble Pie d	liagrai	n to repi	esent th	e follow	ing sub	mitted as	a part of	the budge	et	
proposal	of the go	vt. of	India fo	r the vea	or 1995 -	- 96.		•			
Proposa	L CI		1.			,	D (
	Item of I	Expen	diture				Percent	age			
	1. Inte	erest						26			
	2. De	fence						13			
	3. Sul	bsidie	S					6			
	4. Oth	er nor	n plan ex	penditu	re			10			
	5. Sta	ites sh	are of ta	xes and	duties			15			
	6. No	n-plai	n assista	nce to st	ate and	UT govt	•	6			
	7. Sta	ite and	l UT pla	n assista	ince			10			
		. 1	lon					14			

LASS: II M.Sc Microbiology	COURSE NAM	COURSE NAME: Biostatistics and Research Methodolog					
COURSE CODE: 17MBP304	UNIT: I	BATCH-201	.7-2019				
7. Calculate the mode.							
Marks : 0-20	20-40	40-60 6	60-80 80-10	0			
No. of students: 5	5 20	35	20 12				
8. Calculate the standard deviatio	n for the following dat	a.					
X : 0-10	.0-20 20-30	30-40	40-50				
F: 2	5 9	3	1				
9. Draw a Histogram and hence f	nd the modal wage.						
Weekly wage(in Rs)]:	310 330 350	370 390					
(Midvalue)							
No of labourary	25 50 75	60 15					
No. of fabourers .	25 50 75	00 15					
10. Calculate the coefficient of va	riation for the following	ng data.					
X: 6 9 1	2 15 18						
F: 7 12 1	.3 10 8						
	Part-C (10 ma	nrk)					
1. Calculate the standard deviation	on for the following da	ta.					
Annual Profit (Rs): 2	20-40 40-60	60-80	80-100				
No. Of Banks:	10 14	25	48				
1	100-120 120-14	140-160					
	33 24	16					

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIBATCH-2017-2019

UNIT-II

SYLLABUS

Correlation – Meaning and definition - Scatter diagram –Karl pearson's correlation coefficient. Rank correlation.

 $Regression: Regression in two \ variables - Regression \ coefficient \ problems - uses \ of \ regression.$

CORRELATION ANALYSIS

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	

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: II BATCH-2017-2019

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)
- \blacktriangleright 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 $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$. 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

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology **COURSE NAME: Biostatistics and Research Methodology** COURSE CODE: 17MBP304

BATCH-2017-2019

UNIT: II

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
KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology **COURSE NAME: Biostatistics and Research Methodology** COURSE CODE: 17MBP304 BATCH-2017-2019

UNIT: II

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.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIBATCH-2017-2019

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



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

~		
Subject	Age x	Weight Level y
Ů)	
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	ху	\mathbf{x}^2	y^2
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIBATCH-2017-2019

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

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	ху	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	ху	\mathbf{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

CLASS: II M.Sc Microbiology **COURSE NAME: Biostatistics and Research Methodology** COURSE CODE: 17MBP304 UNIT: II BATCH-2017-2019

Step 6: Use the following formula to work out the correlation coefficient. The answer is: $1.3787 \times 10-4$ 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:

•	/		
o = 1	l -	$6 \Sigma d^2$	
r -		\mathbf{T}^2 1	
	_N(.	N⁻ -1))	

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: 2	1 36	42	37	25	
Y: 4	47 40	37	42	43.	For the data given above , calculate the rank
correlation co	efficient.				

Solution :

		RANK			
X	Y	Х	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				$\sum d = 0$	$\Sigma d^2 = 38$

Total

 $\rho = 1 - \left(\frac{6 \sum d^{2}}{N(N^{2} - 1)}\right)$ $= 1 - \left(\frac{6 \times 38}{5 (5^{2} - 1)}\right)$ = 1 - 1.9

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIBATCH-2017-2019

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 Economi	cs: 50	60	65	70	75	40	70	80
Marks in Statistics	: 80	71	60	75	90	82	70	50
Solution: Let X - Marks in E	Economics							
Y - Marks in S	Statistics	_						
		R/	ANK					
X Y		Х		Y		d		\mathbf{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	$\sum d^2 = 113.5$

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

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.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIBATCH-2017-2019

- 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 <u>sum of squares</u> 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.

0

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,

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

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,

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

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

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIBATCH-2017-2019

Methods of forming the regression equations

- Regression equations on the basis of <u>normal equations</u>.
- 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.

Х	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
∑x=0	∑y=0	∑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,

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

 $\sum \! xy = a {\sum \! x} + b {\sum \! x^2}$

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIBATCH-2017-2019

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

(1)	the two regression equations										
(ii)	The correlation coefficient between the variables X and y										
(iii)	The value	e of Y	when X	= 30							
X : 25	28	35	32	31	36	29	38	34	32		
		10					•		•		
Y: 43	46	49	41	36	32	31	30	33	39		

CLASS: II M.Sc Microbiology COURSE CODE: 17MBP304 COURSE NAME: Biostatistics and Research Methodology UNIT: II BATCH-2017-2019

Solution

Х	Y	x = X - X	$Y = Y - Y^{$	xy	\mathbf{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, $(Y - Y) = b_{yx} (X-X)$

 $(Y - 38) = -0.6643(X-32) \Longrightarrow Y = 59.26-0.6643X$

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

 $(X - 32) = -0.2337Y + 8.88 \Rightarrow X = 40.88 - 0.233 Y$

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

(iv) $Y = 59.26 - 0.6643 \times 30 = 39$

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.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIBATCH-2017-2019

- 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.

CLASS: II M.Sc Microbiology COURSE CODE: 17MBP304

yCOURSE NAME: Biostatistics and Research Methodology4UNIT: IIBATCH-2017-2019

POSSIBLE QUESTIONS :

Part-B(6 mark)

1.	Calculate the	correla	tion coe	fficient	from th	e follo	wing dat	a.			
	X: 57	58	59	59	60	61	62	64			
	Y: 67	68	65	68	72	72	69	71			
2.		Find k	Karl Pea	rson s co	oefficie	nt of c	orrelation	n from t	the follo	owing d	ata.
	Wages	: 100	101	102	102	100	99	97	98	96	95
	Cost of living	g : 98	99	99	97	95	92	95	94	90	91
3.	Find the Kar	l Pearso	n's coe	fficient	of corre	elation	from the	marks	secured	d by 10	students
	in Accountant	cy and	statistic	s.							
	Marks in Acc	countanc	cy (X):	45 70	65 30) 90	40 50	75 85	5 60		
	Marks in stat	istics	(Y):	35 90	70 4	0 95	40 60	80 8	0 50		
4.	Marks obtain	ed by 8	student	s in Acc	ountan	cy (x)	and statis	stics(y)	are give	en belov	V.
	Compute ran	k correl	ation								
	X : 15	20	28	12	40	60	20	80			
	Y : 40	30	50	30	20	10	30	60			
5.	The heights of	of Father	rs (X) a	nd those	of thei	r Sons	(Y) are g	iven be	low. Ca	lculate	
	Spearman's r	ank cor	relation	coeffici	ent.						
	X:180	155	170	174	160	172	166	172	172		
	Y:170	165	180	180	164	169	170	170	174		
6.	Find the rank	correla	tion coe	efficient	for the	percer	ntage of r	narks se	ecured b	by a gro	up of 8
	students in E	conomic	es and S	tatistics	•						
	Marks in Eco	onomics	: 50	60	65	70	75	40	70	80	
	Marks in Star	tistics:	80	71	60	75	90	82	70	50	
7.	Calculate the	two reg	gression	equatio	ns from	the fo	ollowing	data.			
	X:10 12	13 1	2 16	15							
	Y: 40 38	43 43	5 37	43							
8.	From the data	a given	below f	ind the t	wo regi	ression	lines.				
	X: 10	12	13	12	16	15					
	Y: 40	38	43	45	37	43.					
	1) Estimate Y	when 2	X = 20.								
0	11) Estimate 2	when the fell	$\mathbf{Y} = 2\mathbf{S}.$	data	V	V					
9.	You are give	n the Iol	llowing	data	А 5	1 25					
	Standard dev		anto hat		J V l-V	23					
	Eind the rear		o offici	onto	ΛαΙ	0.0					
10	Vou are give	ession c n tha fol	lowing	doto							
10.	Tou are give	ii the io	Nowing X	uala.	v						
	Arithmetic m	ean	20		20						
	Standard dev	iation	5		25						
	Correlation c	oefficie	nt betw	een X ar	$\frac{2}{10}$ $\mathbf{Y} = 0$	0.66					
	Find the two	regress	ion equ	ations							
	- 110 110 100		-on equ								

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology **COURSE NAME: Biostatistics and Research Methodology** COURSE CODE: 17MBP304 UNIT: II BATCH-2017-2019 Part-C(10 mark) 1. You are given the following data: Х Y Arithmetic mean 36 85 Standard deviation 11 8 Correlation coefficient between X and Y = 0.66i) Find the two regression equations

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019

UNIT-III

SYLLABUS

Test of significance: Tests based on Means only-Both Large sample and Small sample tests - Chi square test - goodness of fit. Analysis of variance – one way and two way classification. CRD, RBD Designs.

Hypothesis :

A hypothesis a statement about the population parameter. In other words a hypothesis is a conclusion which is usually drawn on a logical bases.

Statistical hypothesis:

It is an idea which may or may not be true about a population or about the probability distribution which we want to test on the basis of random sample.

Testing of Hypothesis:

It is a procedure that helps us to claim the likelihood of hypothesis population parameter, being correct by making use of the sample statistics.

Test of Significance:

The testing of hypothesis discloses the fact that whether the different between sample statistics and the corresponding population parameter is significant or not.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

Procedure for testing a Hypothesis:

1.Setting up a Hypothesis:

Hypothesis testing is a scientific method of choosing between two claims H_0 and H_a

a) Null Hypothesis:

The null hypothesis asserts that there is no different between sample statistic and the population parameter. H_0 is called the null hypothesis. It says that the educated earn no better than others.

 $H_0: \mu = \mu_0$

b) Alternative Hypothesis:

The negation of Null hypothesis is called alternative hypothesis. H_a is called the alternative hypothesis, which says that the educated earn more.

Example:

In this example, the alternative is one-sided. If the claim were that the educated earn less than others, then the alternative would be set up as H_a μ < 30000.

Some principles:

1) H₀ is presumed true unless overwhelming evidence rejects it. *E.g.*, H₀: defendant is not guilty!

2)Sample data give test statistics for μ , p or σ^2

3)We reject the Null H_0 if statistic falls in Rejection Region

4) Null is usually a zero value (hence the name null).

5) Instead of saying ACCEPT one says FAILS TO REJECT.

6) Absolute certainty does not exist.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

If there is a standard value for μ the null is H₀ μ =std value (or true value). e.g. H₀ μ =10 and two-sided alternative is H_a $\mu \neq$ 10, where it could be larger than 10 or smaller than 10.

Need Skill to decide (1) appropriate statistical parameter μ , p, etc. (2) appropriate Null (3) One-sided or two-sided. When one of the alternatives is selected, there can be error.

Type I (α) and Type II (β) errors

Definition

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

Data Sets

There are two types of data sets you will use when studying statistics. These data sets are called *populations* and *samples*.

Definition

A **population** is the collection of *all* outcomes, responses, measurements, or counts that are of interest.

A sample is a subset of a population.

Definition: Type I error:

Selecting H_a when H_0 should be selected. H_0 DOGS are dead, H_a DOGS are alive

Truth is DOGS are dead, still selecting H_a is Type I error

Definition:

Type II error:

Selecting H_0 when H_a should be selected. H_0 DOGS are dead, H_a DOGS are alive.Truth is DOGS are alive, still selecting H_0 means Type II error.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019

Definition:

 α denotes the probability of Type I error = This is also called the "Level of test."

Definition:

 β denotes the probability of Type II error. This is usually hard to determine since it depends on unknown parameter μ itself.

It is desirable to formulate the hypothesis so that α is the most serious consequence. The two types of errors are inversely related. There is a trade-off between α and β . The smaller we make α the larger the β we have to accept. Hence one usually chooses α largest tolerable !

Steps in the Test of Hypothesis:

- Define the hyp. to be tested in plain English.
- Select the appropriate statistical measure (such as $\mu p, \sigma^2$) to rephrase the hypothesis.
- Determine whether hyp. should be 1 or 2-sided.
- .State the hypothesis using the statistical measure selected in step 2..Specify α , the "level" of the test.
- .Select the appropriate test statistic, based on the information at hand and the assumptions you are willing to make.
- .Determine the critical value of the test statistic. Three factors for critical value: a. the type of alternative hypothesis,

(1) Two-sided (2) 1-sided left (3) 1-sided Right

b. the specification of a, the level of the test,

c. the distribution of the test statistic.

- Collect sample data and compute the value of the test statistic.
- .Make the decision. Is the value of the test statistic in the rejection region? a. If yes, reject the null hypothesis in favor of the alternative.

b.lf no, do not reject the null hypothesis.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

Z - Test for a Mean (Large Samples) Single mean:

The sample size n has to be greater than or equal to 30, or the population standard deviation is given The given population mean μ and a sample mean \overline{x} . The result of the test will be a conclusion in which we state that the population mean is, or is not significantly different from, or is less than or more than what is stated.

The test statistic for this type of hypothesis testing is $z = \frac{\overline{X} - \mu}{\frac{s}{\overline{x}}}$, the critical value z_c is to be found

from the normal distribution table.

Example 1:

The Public Health Service publishes the "Annual Data Tabulations, Continuous Air Monitoring Projects", which some years ago indicated that a large Midwestern city had an annual mean level of sulfur dioxide of 0.12 (concentration in parts per million). To change this concentration, many steel mills and other manufacturers installed antipollution equipment. In order to study the effects of these efforts to reduce pollution 36 random checks were made throughout the year. It was found that the sample mean pollution sulfur dioxide level was 0.115 with a sample standard deviation of 0.03.

At the 0.05 level of significance, does this evidence suggest that there has been a change in the sulfur dioxide level in this city?

Solution 1:

We will follow the four step procedure as outlined in "*Hypothesis Testing, An Overview*". The word "change" in the last sentence of the problem tells us to perform a two tailed test.

Step 1 : The Null Hypothesis

The sulfur dioxide concentration in this city is not significantly different from the reported level of 0.12 parts per million.

Step 2 : The Decision Rule

We accept the null hypothesis if -1.96 < z < 1.96

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

Step 3 : The Test Statistic

Using
$$z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$
 we find $z = \frac{0.115 - 0.12}{\frac{0.03}{\sqrt{36}}} = -1$.

Step 4 : The Conclusion

Combining the info from steps 2 and 3 in a picture with a normal curve we get:



We see that z is in the shaded region, i.e. in the acceptance region formulated in step 2. Thus we conclude that at the 0.05 level of significance we accept the hypothesis that the sulfur dioxide level in this city is not significantly different from the reported level. In other words, the efforts of industry to change the sulfur dioxide level from the reported level of 0.12 per million parts have not been successful.

KARPAGAM ACADEMY OF HIGHER EDUCATION					
CLASS: II M.Sc Microbiology	COURSE NAME:	Biostatistics and Research Methodology			
COURSE CODE: 17MBP304	UNIT: III	BATCH-2017-2019			

Example 2:

A machine is set to fire 30.00 decigrams of chocolate pellets into a box of cake mix as it moves along the production line. Of course, there is some variation in the weight of the pellets. A sample of 36 boxes of mix revealed that the average weight of the chocolate pellets was 30.18 decigrams with a sample standard deviation of 0.50 decigrams. At the 0.05 level of significance, is the sampled weight of the chocolate pellets of 30.18 decigrams suggesting that the mean weight of the chocolate pellets in all cake mix boxes is higher than the set 30.00 decigrams?

Solution 2:

The words "higher than" in the last sentence of the problem tells us to perform a right tailed test.

Step 1 : The Null Hypothesis

The mean weight of the chocolate pellets in the boxes of cake mix is not significantly higher than 30.00 decigrams.

Step 2 : The Decision Rule

We accept the null hypothesis if z < 1.65

Step 3 : The Test Statistic

Using
$$z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$
 we find $z = \frac{30.18 - 30.00}{\frac{0.5}{\sqrt{36}}} = 2.16$.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

Step 4 : The Conclusion

Combining the info from steps 2 and 3 in a picture with a normal curve we get:



We see that z is outside of the shaded region, i.e. outside of the acceptance region formulated in step 2. Thus we conclude that at the 0.05 level of significance there seems to be sufficient evidence to suggest that the mean weight of the chocolate pellets in all cake mix boxes is higher than the set 30.00 decigrams. In other words, the machine firing the chocolate pellets into the boxes of cake mix is not functioning correctly.

Z- test Difference between two Mean:

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019



where x_1 and x_2 are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means), σ_1 and σ_2 are the standard deviations of the two populations, and n_1 and n_2 are the sizes of the two samples.

Example 1 (two-tailed test): The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. What is the likelihood that the population means of concentrations of the element are the same for men and women?

Null hypothesis: H_0 : $\mu_1 = \mu_2$

or $H_0: \mu_1 - \mu_2 = 0$

alternative hypothesis: H_a : $\mu_1 - \mu_2$

or: H_a : $\mu_1 - \mu_2 \neq 0$

$$z = \frac{28 - 33 - 0}{\sqrt{\frac{14.1^2}{75} + \frac{9.5}{50}^2}} = \frac{-5}{\sqrt{2.65 + 1.81}} = -2.37$$

The computed *z*-value is negative because the (larger) mean for females was subtracted from the (smaller) mean for males. But because the hypothesized difference between the populations is 0, the order of the samples in this computation is arbitrary—x, could just as well have been the female sample mean and x, the male sample mean, in which case z would be 2.37 instead of –2.37. An extreme z-score in either tail of the distribution (plus or minus) will lead to rejection of the null hypothesis of no difference.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

The area of the standard normal curve corresponding to a *z*-score of -2.37 is .0089. Because this test is two-tailed, that figure is doubled to yield a probability of .0178 that the population means are the same. If the test had been conducted at a pre-specified significance level of $\alpha < .05$, the null hypothesis of equal means could be rejected. If the specified significance level had been the more conservative (more stringent) $\alpha < .01$, however, the null hypothesis could not be rejected.

In practice, the two-sample z-test is not often used because the two population standard deviations σ_1 and σ_2 are usually unknown. Instead, sample standard deviations and the *t*-distribution are used.

t-test Small samples:single mean:

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

where x is the sample mean, μ is a specified value to be tested, s is the sample standard deviation, and n is the size of the sample. When the standard deviation of the sample is substituted for the standard deviation of the population, the statistic does not have a normal distribution; it has what is called the *t*-distribution. Because there is a different *t*-distribution for each sample size, it is not practical to list a separate area-of-the-curve table for each one. Instead, critical *t*-values for common alpha levels (.05, .01, .001, and so forth) are usually given in a single table for a range of sample sizes. For very large samples, the *t*-distribution approximates the standard normal (*z*) distribution.

Values in the *t*-table are not actually listed by sample size but by degrees of freedom (*df*). The number of degrees of freedom for a problem involving the *t*-distribution for sample size *n* is simply n - 1 for a one-sample mean problem.

Example 1 (one-tailed test): A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score at least 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor be at least 90 percent certain that the mean score for the class on the test would be at least 70?

null hypothesis: H_0 : $\mu < 70$

alternative hypothesis: H a: $\mu \ge 70$

First, compute the sample mean and standard deviation.

CLASS: II M.Sc Microbiology	COURSE NAME	Biostatistics and Research Methodolog
COURSE CODE: 17MBP304	UNIT: III	BATCH-2017-2019
	62	
	92	
	75	475
	68 x = -	$\frac{1}{6} = 79.17$
	83 3	= 13.17
	95	
4	75	
North construction (loss of construction)		

$$t = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = \frac{9.17}{5.38} = 1.71$$

To test the hypothesis, the computed *t*-value of 1.71 will be compared to the critical value in the *t*-table. But which do you expect to be larger and which smaller? One way to reason about this is to look at the formula and see what effect different means would have on the computation. If the sample mean had been 85 instead of 79.17, the resulting *t*-value would have been larger.

Because the sample mean is in the numerator, the larger it is, the larger the resulting figure will be. At the same time, you know that a higher sample mean will make it more likely that the professor will conclude that the math proficiency of the class is satisfactory and that the null hypothesis of less-than-satisfactory class math knowledge can be rejected. Therefore, it must be true that the larger the computed *t*-value, the greater the chance that the null hypothesis can be rejected. It follows, then, that if the computed *t*-value is larger than the critical *t*-value from the table, the null hypothesis can be rejected.

A 90 percent confidence level is equivalent to an alpha level of .10. Because extreme values in one rather than two directions will lead to rejection of the null hypothesis, this is a one-tailed test, and you do not divide the alpha level by 2. The number of degrees of freedom for the problem is 6 - 1 = 5. The value in the *t*-table for $t_{10,5}$ is 1.476. Because the computed *t*-value of 1.71 is larger than the critical value in the table, the null hypothesis can be rejected, and the professor can be 90 percent certain that the class mean on the math test would be at least 70.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

Note that the formula for the one-sample *t*-test for a population mean is the same as the *z*-test, except that the *t*-test substitutes the sample standard deviation *s* for the population standard deviation σ and takes critical values from the *t*-distribution instead of the *z*-distribution. The *t*-distribution is particularly useful for tests with small samples (n < 30).

Example 2 (two-tailed test): A Little League baseball coach wants to know if his team is representative of other teams in scoring runs. Nationally, the average number of runs scored by a Little League team in a game is 5.7. He chooses five games at random in which his team scored 5 9, 4, 11, and 8 runs. Is it likely that his team's scores could have come from the national distribution? Assume an alpha level of .05.

Because the team's scoring rate could be either higher than or lower than the national average, the problem calls for a two-tailed test.

First, state the null and alternative hypotheses:

null hypothesis: H_0 : $\mu = 5.7$

alternative hypothesis: H_a : $\mu \neq 5.7$

Next compute the sample mean and standard deviation:



Next, the *t*-value:

$$t = \frac{7.4 - 5.7}{\frac{2.88}{\sqrt{5}}} = \frac{1.7}{1.29} = 1.32$$

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

Now, look up the critical value from the *t*-table. You need to know two things in order to do this: the degrees of freedom and the desired alpha level. The degrees of freedom is 5 - 1 = 4. The overall alpha level is .05, but because this is a two-tailed test, the alpha level must be divided by two, which yields .025. The tabled value for $t_{.025,4}$ is 2.776. The computed *t* of 1.32 is smaller than the *t* from the table, so you cannot reject the null hypothesis that the mean of this team is equal to the population mean. The coach can conclude that his team fits in with the national distribution on runs scored.

Chi-Square test:

$$\chi^2 = \sum (E-O)^2 / E$$

Example, "There is no association between the gender of applicants and whether or not their application is accepted".

The first step is to put the observed data (what was actually measured/recorded) into a contingency table as shown in Table 1.

Table	1:	Observ	ed fr	equer	ncies
-------	----	--------	-------	-------	-------

	Application successful	Application not successful	TOTAL
Male	23	40	63
Female	31	39	70
TOTAL	54	79	133

The second step is to calculate the expected frequencies. These are the frequencies that we would have expected to have recorded given the row/column totals. There are several different ways you can do this. The most basic involves calculating what you would observe if there was no association between the two variables. This is done by multiplying the row total by the column total and dividing the result by the table total, for each cell. This is shown in Table 2:

KARPAGAM ACADEMY OF HIGHER EDUCATIONCLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research MethodologyBATCH-2017-2019

Table 2: Calculation of expected frequencies

	Application successful	Application not successful	TOTAL
Male	(63 * 54) / 133 = 25.58	(63 * 79) / 133 = 37.42	63
Female	(70 * 54) / 133 = 28.42	(70 * 79) / 133 = 41.58	70
TOTAL	54	79	133

Be sure to check that your observed and expected values both sum up to the same total.

The third step is to calculate the chi-square statistic.

The formula for chi-square is: $\chi^2 = \sum (E-O)^2 / E$

Where E is the expected values and O is the observed values. The sigma sign means that everything that follows is summed. So '(expected – observed)² / expected' is calculated for each cell in the contingency table as shown below.

	The cheering	
The expected	The observed	
value for this	value for this	
cell	cell	
0011		1

	Application successful	Application not successful
Male	(25.58 – 23) ² / 25.28	(37.42 - 40) ² / <i>37.42</i>
	≠ 0.26	= 0.18

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

Female	(28.42 - 31) ² / 28.42	(41.58 - 39) ² / 41.58
	= 0.23	=0.16

..... and then the results from each cell are summed:

0.26 + 0.18 + 0.23 + 0.16 = <u>0.83</u>

And that is the χ^2 value.

Calculate the degrees of freedom:

(number of rows -1) x (number of columns -1)=(n-1)x(c-1)

Here there are two rows and two columns, so the degrees of freedom is 1.

The final step is to see whether the chi-square value, given the degrees of freedom, is statistically significant. This can be done by comparing the χ^2 value against a table of critical values that have been derived from the chi-square distribution. Alternatively most software packages will tell you the exact p-value so you can see instantly whether it is below the standard threshold of 0.05 or not.

In this example, $\chi^2 = 0.83$ which is greater than 0.0199, the critical value for 5% significance with 1 degree of freedom. Hence the result is not significant and we cannot reject the null hypothesis that there is no association between the gender of applicants and whether or not their application is accepted or not. What this means is that the likelihood that the difference between the observed and expected values is due to chance rather than any genuine affect is greater than 5%.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

When presenting results in reports it is usual to give one table and put the expected frequencies in brackets, like this:

	Application	Application not	TOTAL
	successful	successful	
Male	23 (25.58)	40 (37.42)	63
Female	31 (28.42)	39 (41.58)	70
TOTAL	54	79	133

The results are then written as $\chi^2 = 0.08$, df = 1, p<0.05 (or p=NS if the result was not significant).

One way ANOVA:

A One-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

Assumptions:

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

Hypotheses:

The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

In the following, lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size.

Grand Mean:

$$\overline{X}_{GM} = \frac{\sum x}{N}$$

$$\overline{X}_{GM} = \frac{\sum n \overline{x}}{\sum n}$$

The grand mean of a set of samples is the total of all the data values divided by the total sample size. This requires that you have all of the sample data available to you, which is usually the case, but not always. It turns out that all that is necessary to find perform a one-way analysis of variance are the number of samples, the sample means, the sample variances, and the sample sizes.

Another way to find the grand mean is to find the weighted average of the sample means. The weight applied is the sample size.

Total Variation:

The total variation (not variance) is comprised the sum of the squares of the differences of each mean with the grand mean.

There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means aren't the same.

$$SS(T) = \sum (x - \overline{X}_{GM})^2$$

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

Between Group Variation:

The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.

The variance due to the interaction between the samples is denoted MS(B) for Mean Square Between

groups. This is the between group variation divided by its degrees of freedom. It is also denoted by ${}^{\mathcal{S}}$

$$SS(B) = \sum n (\bar{x} - \bar{X}_{GM})^2$$

Within Group Variation:

The variation due to differences within individual samples, denoted SS(W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

$$SS(W) = \sum df \cdot s^2$$

Mean Square Within groups:

The variance due to the differences within individual samples is denoted MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom. It is also $s_{x,w}^2$

denoted by $\overset{\circ}{}^{\mu}$. It is the weighted average of the variances (weighted with the degrees of freedom).

KARPAGAM ACADEMY OF HIGHER EDUCATIONCLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

F test statistic:

A F variable is the ratio of their respective degrees of freedom.



Two independent chi-square variables divided by

Also recall that the F test statistic is the ratio of two sample variances, well, it turns out that's exactly what we have here. The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).

Summary Table:

	SS	df	MS	F
Between	SS(B)	k-1	SS(B) k-1	MS(B) MS(W)
Within	SS(W)	N-k	SS(W) N-k	
Total	SS(W) + SS(B)	N-1		

Each Mean Square is just the Sum of Squares divided by its degrees of freedom, and the F value is the ratio of the mean squares. Do not put the largest variance in the numerator, always divide the

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

between variance by the within variance. If the between variance is smaller than the within variance, then the means are really close to each other and you will fail to reject the claim that they are all equal. The degrees of freedom of the F-test are in the same order they appear in the table.

Decision Rule:

The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with k-1 numerator and N-k denominator degrees of freedom.

Oneway Anova Example:

Consider an experiment to study the effect of three different levels of some factor on a response (e.g. three types of fertilizer on plant growth). If we had 6 observations for each level, we could write the outcome of the experiment in a table like this, where a_1 , a_2 , and a_3 are the three levels of the factor being studied.

a 1	a 2	<i>a</i> ₃	
6	8	13	
8	12	9	
4	9	11	
5	11	8	
3	6	7	
4	8	12	

The null hypothesis, denoted H_0 , for the overall F-test for this experiment would be that all three levels of the factor produce the same response, on average. To calculate the F-ratio:

Step 1: Calculate the mean within each group:

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019

$$\overline{Y}_{1} = \frac{1}{6} \sum Y_{1i} = \frac{6+8+4+5+3+4}{6} = 5$$

$$\overline{Y}_{2} = \frac{1}{6} \sum Y_{2i} = \frac{8+12+9+11+6+8}{6} = 9$$

$$\overline{Y}_{3} = \frac{1}{6} \sum Y_{3i} = \frac{13+9+11+8+7+12}{6} = 10$$

Step 2: Calculate the overall mean:

$$\overline{Y} = \frac{\sum_i \overline{Y}_i}{a} = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3}{a} = \frac{5 + 9 + \overline{10}}{3} = 8$$

where *a* is the number of groups.

Step 3: Calculate the "between-group" sum of squares:

$$S_B = n(\overline{Y}_1 - \overline{Y})^2 + n(\overline{Y}_2 - \overline{Y})^2 + n(\overline{Y}_3 - \overline{Y})^2$$
$$= 6(5-8)^2 + 6(9-8)^2 + 6(10-8)^2 = 84$$

where *n* is the number of data values per group.

The between-group degrees of freedom is one less than the number of groups

$$f_b = 3 - 1 = 2$$

so the between-group mean square value is

$$MS_B = 84 / 2 = 42$$

Step 4: Calculate the "within-group" sum of squares. Begin by centering the data in each group

$$a_1$$
 a_2 a_3
 $6-5=1$ $8-9=-1$ $13-10=3$
 $8-5=3$ $12-9=3$ $9-10=-1$
 $4-5=-1$ $9-9=0$ $11-10=1$

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

5 - 5 = 0 11 - 9 = 2 8 - 10 = -2 3 - 5 = -2 6 - 9 = -3 7 - 10 = -3

4 - 5 = -18 - 9 = -112 - 10 = 2

The within-group sum of squares is the sum of squares of all 18 values in this table

$$\begin{split} S_W &= 1+9+1+0+4+1+1+9+0+4+9+1+9+1+1+4+9+4\\ &= 68 \end{split}$$

The within-group degrees of freedom is

$$f_W = a(n-1) = 3(6-1) = 15$$

Thus the within-group mean square value is

$$MS_W = S_W / f_W = 68/15 \approx 4.5$$

Step 5: The F-ratio is

$$F = \frac{MS_B}{MS_W} \approx 42/4.5 \approx 9.3$$

The critical value is the number that the test statistic must exceed to reject the test. In this case, $F_{\text{crit}}(2,15) = 3.68$ at $\alpha = 0.05$. Since F = 9.3 > 3.68, the results are significant at the 5% significance level. One would reject the null hypothesis, concluding that there is strong evidence that the expected values in the three groups differ. The p-value for this test is 0.002.

After performing the F-test, it is common to carry out some "post-hoc" analysis of the group means. In this case, the first two group means differ by 4 units, the first and third group means differ by 5 units, and the second and third group means differ by only 1 unit. The standard error

of each of these differences is $\sqrt{4.5/6 + 4.5/6} = 1.2$. Thus the first group is strongly different from the other groups, as the mean difference is more times the standard error, so we can be highly confident that the population mean of the first group differs from the population means of the other groups. However there is no evidence that the second and third groups have different population means from each other, as their mean difference of one unit is comparable to the standard error.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019

Two-Way ANOVA:

The two-way analysis of variance is an extension to the one-way analysis of variance. There are two independent variables (hence the name two-way).

Assumptions:

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.
- The groups must have the same sample size.

Hypotheses:

There are three sets of hypothesis with the two-way ANOVA.

The null hypotheses for each of the sets are given below.

- 1. The population means of the first factor are equal. This is like the one-way ANOVA for the row factor.
- 2. The population means of the second factor are equal. This is like the one-way ANOVA for the column factor.
- 3. There is no interaction between the two factors. This is similar to performing a test for independence with contingency tables.

Factors:

The two independent variables in a two-way ANOVA are called factors. The idea is that there are two variables, factors, which affect the dependent variable. Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels.

Treatment Groups

Treatement Groups are formed by making all possible combinations of the two factors. For example, if the first factor has 3 levels and the second factor has 2 levels, then there will be 3x2=6 different treatment groups.

As an example, let's assume we're planting corn. The type of seed and type of fertilizer are the two factors we're considering in this example. This example has 15 treatment groups. There are 3-1=2 degrees of freedom for the type of seed, and 5-1=4 degrees of freedom for the type of fertilizer. There are 2*4 = 8 degrees of freedom for the interaction between the type of seed and type of fertilizer.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology

COURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

The data that actually appears in the table are samples. In this case, 2 samples from each treatment group were taken.

	Fert I	Fert II	Fert III	Fert IV	Fert V
Seed A-402	106, 110	95, 100	94, 107	103, 104	100, 102
Seed B-894	110, 112	98, 99	100, 101	108, 112	105, 107
Seed C-952	94, 97	86, 87	98, 99	99, 101	94, 98

Main Effect:

The main effect involves the independent variables one at a time. The interaction is ignored for this part. Just the rows or just the columns are used, not mixed. This is the part which is similar to the one-way analysis of variance. Each of the variances calculated to analyze the main effects are like the between variances

Interaction Effect:

The interaction effect is the effect that one factor has on the other factor. The degrees of freedom here is the product of the two degrees of freedom for each factor.

Within Variation

The Within variation is the sum of squares within each treatment group. You have one less than the sample size (remember all treatment groups must have the same sample size for a two-way ANOVA) for each treatment group. The total number of treatment groups is the product of the number of levels for each factor. The within variance is the within variation divided by its degrees of freedom.

The within group is also called the error.

F-Tests

There is an F-test for each of the hypotheses, and the F-test is the mean square for each main effect and the interaction effect divided by the within variance. The numerator degrees of freedom come from each effect, and the denominator degrees of freedom is the degrees of freedom for the within variance in each case.
KARPAGAM ACADEMY OF HIGHER EDUCATIONCLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019

Two-Way ANOVA Table

It is assumed that main effect A has a levels (and A = a-1 df), main effect B has b levels (and B = b-1 df), n is the sample size of each treatment, and N = abn is the total sample size. Notice the overall degrees of freedom is once again one less than the total sample size.

Source	SS	df	MS	F
Main Effect A	given	А, а-1	SS / df	MS(A) / MS(W)
Main Effect B	given	B, b-1	SS / df	MS(B) / MS(W)
Interaction Effect	given	A*B, (a-1)(b-1)	SS / df	MS(A*B) / MS(W)
Within	given	N - ab, ab(n-1)	SS / df	
Total	sum of others	N - 1, abn - 1		

Summary

The following results are calculated using the Quattro Pro spreadsheet. It provides the p-value and the critical values are for alpha = 0.05.

Source of Variation	SS	df	MS	F	P-value	F-crit
Seed	512.8667	2	256.4333	28.283	0.00008	3.682
Fertilizer	449.4667	4	112.3667	12.393	0.000119	3.056
Interaction	143.1333	8	17.8917	1.973	0.122090	2.641

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IIIBATCH-2017-2019

Within	136.0000	15	9.0667		
Total	1241.4667	29			

From the above results, we can see that the main effects are both significant, but the interaction between them isn't. That is, the types of seed aren't all equal, and the types of fertilizer aren't all equal, but the type of seed doesn't interact with the type of fertilizer.

Here is the correct table:

Source of Variation	SS	df	MS	F
Sample	3.920	1	3.920	4.752
Column	9.680	1	9.680	11.733
Interaction	54.080	1	54.080	65.552
Within	3.300	4	0.825	
Total	70.980	7		

Blocking:

The analysis of experiment is used to eliminate one source of variability is discussed in completely randomized block design. The analysis of experiment is used to eliminate two or three source of variability is discussed in randomized block design.

Completely Randomised Block Design:

Consider the statistical analysis of the the completely randomized design or oneway classification.Suppose the experiment has available the results of k independent random samples from k different populatios. Testing the hypothesis that the mean of these k populations are all equal. We complete the completely randomized design by One-Way Analysis of Variance.

KARPAGAM ACADEMY OF HIGHER EDUCATION CLASS: II M.Sc Microbiology COURSE NAME: Biostatistics and Research Methodology COURSE CODE: 17MBP304 UNIT: III BATCH-2017-2019

Randomised Block Design:

The analysis of experiment in a two way classification. This kind of arrangement is also called a randomized block design, provided the treatment are allocated at random within each block.

CLASS: II M.Sc Microbiology COURSE CODE: 17MBP304

COURSE NAME: Biostatistics and Research Methodology UNIT: III BATCH-2017-2019

POSSIBLE QUESTIONS :

Part-A(6 mark)

1. Two types of batteries are tested for their length of life and the following data are obtained.

	Type I	Type II
Sample size	9	8
Mean	600 hrs	640 hrs
Variance	121	144

Use the 5% level of significance to test whether the mean values differ significantly? (Table value is t $_{0.025, 15} = 2.131$).

2. A survey of 320 families with 5 children each revealed the following distributions. No. of Boys : No. of Girls No. of Families: Is the result consistent with the hypothesis that male and female births are equally probable? (Table value is $\chi^2_{0.05, 5} = 11.0705$). 3. The systolic pressure of 10 persons in the age of 45 - 50 is given below. 148, 128, 147, 127, 150, 145, 124, 140, 142, 149. From the data at 5% level discuss the suggestion that the average systolic pressure of the population is 150. (Table value is t $_{0.025.9} = 3.17$). 4. Write the procedure of classification of two – way ANOVA 5. Write the steps in testing the hypothesis. 6. Below are given the gain in weights in Kg. of cows fed on 2 diets X and Y. Diet X: 25 Diet Y: 24 Test at 5% level whether the two diets differ in their effect on mean weights. (Table value is t $_{0.025, 15} = 2.131$). 7. 300 digits were chosen at random from a set of table, the frequency of the digits was as follows. Digits: Frequency: 28 Using χ^2 test assess the hypothesis that the digits were distributed uniformly in the table. (Table value is $\chi^2_{0.05, 9} = 16.9190$). 8. Random samples are drawn from 2 populations and their results were obtained. Sample X: 16 Sample Y: 19 22 Find variance and test whether the 2 samples have same variance. (Table value is F $_{0.05, 9, 11} = 3.105$). 9. Samples of two types of electric bulbs were tested for length of life and the following data were obtained.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IIIBATCH-2017-2019

	Type I	Type II
Sample size	8	7
Mean	1134 hrs	1024 hrs
SD	35	40

Use the 5% level of significance to test whether the mean values differ significantly? (Table value is t $_{0.025, 13} = 2.160$).

10. The theory predicts that proportion of beans in the four groups A, B, C and D should be 9,3,3,1. In an experiment with 1600 beans, the numbers in the 4 types were 882, 313, 287 and 118. Does the experiment result support the theory? (Table value is $\chi^2_{0.05, 3}$ =7.181).

Part-C(10 mark)

1. The height of 10 children selected at random from a given locality had a mean 63.2 cm and variance 6.25 cm. Test at 5% level of significance, the hypothesis that the children of the given locality are on the average less than 65 cm. (Table value is t $_{0.05, 9} = 2.16$).

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

UNIT-IV

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.

RESEARCH METHODOLOGY

Research in common refers to a search for knowledge. We can define research as a scientific and systematic search for pertinent information on a specific topic. Research is thus an original contribution to the existing stock of knowledge making for its advancement. It is the persuit of truth with the help of study, observation, comparison and experiment.

Meaning of research:

Research is an academic activity and as such the term should be used in a technical sense. According to "Clifford Wooody" research comprises defining and redefining problem ,formulating hypothesis or suggested solutions; collecting, organizing and evaluating data, making deductions and reaching conclusion and atlast carefully testing the conclusion to determine whether they fit the formulating hypothesis.

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 not been 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:

1. 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);

2. To portray accurately the characteristics of a particular individual, situation or a group

(studies with this object in view are known as descriptive research studies);

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IVBATCH-2017-2019

3. To determine the frequency with which something occurs or with which it is associated with something else (studies with this object in view are known as **diagnostic research**

studies);

4. To test a hypothesis of a causal relationship between variables (such studies are known as

hypothesis-testing research studies).

Types of research:

i)Descriptive (Vs) Analytical:

Descriptive research includes surveys and fact-finding enquiries of different kinds. The major purpose of descriptive research is description of the state of affairs as it exist at present. The main characteristics of this method is that the researcher of this method is that the researcher has no control over the variables.

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, where as fundamental research is mainly concerned with generalization and with the formulation of a theory.

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. Qualitative research is specially important in the behavioural science where the aim to discover the underlying motives of human behaviour.

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 a new concepts to reinterpret existing one. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

v) Some other types of research :

All other types of research are variations of one or more of the above stated approach, 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 factors.

Research Process:

Research process consist of series of actions or steps necessary to effectively carry out research and the desired sequencing of these steps.

- Formulating the research problem.
- Extensive literature survey.
- Developig the hypothesis.
- Preparing the research design.
- Determining the sample design.
- Collecting the data.
- Execution of the project.
- Analysis of data.
- Hypothesis testing.
- Generalisation and interpretation.
- Preparation of the report and the presentation of the result.

1. Formulating the research problem:

There are two types of research problems, viz., those which relate to states 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

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

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 view.

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. In 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 proposed. The basic outcome of this review will be the knowledge 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:

(a) Discussions with colleagues and experts about the problem, its origin and the objectives in

seeking a solution;

(b) Examination of data and records, if available, concerning the problem for possible trends,

peculiarities and other clues;

(c) Review of similar studies in the area or of the studies on similar problems; and

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

(d) 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, afteronly 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

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IVBATCH-2017-2019

usually the consideration of the following:

(i) the means of obtaining the information;

(ii) the availability and skills of the researcher and his staff (if any);

(iii) explanation of the way in which selected means of obtaining information will be organised

and the reasoning leading to the selection;

(iv) the time available for research; and

(v) 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'. A 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 sample.

The researcher must decide the way of selecting a sample or what is popularly known as the sample design. In 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 way constitutes a sample design. Samples can be either probability samples or non-probability samples. With probability samples each element has a known probability of being included in the sample but the 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.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

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 collect 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.

7. 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 this 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 analyzing 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

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IVBATCH-2017-2019

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 come from different universes and if the difference is due to chance, the conclusion would be that the two samples belong to the same universe. Similarly, the technique of analysis of variance can help us in analysing whether three or more varieties of seeds grown on certain fields yield significantly different results or not. In brief, the researcher can analyse the collected data with the help of various statistical measures.

9. Hypothesis-testing:

After analysing the data as stated above, the researcher is in a position to test the hypotheses, if any, he had formulated earlier. Do the facts support the hypotheses or they happen to be contrary? This is the usual question which should be answered while testing hypotheses. Various tests, such as Chi square test, t-test, F-test, have been developed by statisticians for the

purpose. The hypotheses may be tested through the use of one or more of such tests, depending upon the nature and object of research inquiry. Hypothesis-testing will result in either accepting the hypothesis or in rejecting it. If the researcher had no hypotheses to start with, generalisations established on the basis of data may be stated as hypotheses to be tested by subsequent researches in times to come.

KARPAGAM ACADEMY OF HIGHER EDUCATIONCLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

10. Generalisations and interpretation:

If a hypothesis is tested and upheld several times, it may be possible for the researcher to arrive at generalisation, i.e., to build a theory. As a matter of fact, the real value of research lies in its ability to arrive at certain generalisations. If the researcher had no hypothesis to start with, he might seek to explain his findings on the basis of some theory. It is known as interpretation. The process of interpretation may quite often trigger off new questions which in turn may lead to further researches.

11. Preparation of the report or the thesis:

Finally, the researcher has to prepare the report of what has been done by him. Writing of report must be done with great care keeping in view the following:

1. The layout of the report should be as follows: (i) the preliminary pages; (ii) the main text,

and (iii) the end matter.

In its preliminary pages the report should carry title and date followed by acknowledgements and foreword. Then there should be a table of contents followed by a list of tables and list of graphs and charts, if any, given in the report.

The main text of the report should have the following parts:

(a) Introduction: It should contain a clear statement of the objective of the research and an explanation of the methodology adopted in accomplishing the research. The scope of the study along with various limitations should as well be stated in this part.

(b) Summary of findings: After introduction there would appear a statement of findings and recommendations in non-technical language. If the findings are extensive, they should be summarised.

(c) Main report: The main body of the report should be presented in logical sequence and broken-down into readily identifiable sections.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

(d) Conclusion: Towards the end of the main text, researcher should again put down the results of his research clearly and precisely. In fact, it is the final summing up.

At the end of the report, appendices should be enlisted in respect of all technical data. Bibliography, i.e., list of books, journals, reports, etc., consulted, should also be given in the end. Index should also be given specially in a published research report.

2. Report should be written in a concise and objective style in simple language avoiding vague

expressions such as 'it seems,' 'there may be', and the like.

3. Charts and illustrations in the main report should be used only if they present the information

more clearly and forcibly.

4. Calculated 'confidence limits' must be mentioned and the various constraints experienced

in conducting research operations may as well be stated.

Criteria of good research :

- 1. The purpose of the research should be clearly defined and common concepts be used.
- 2. The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement.
- 3. The procedural design of the research should be carefully planned to yield results.
- 4. The researcher should report with complete frankness.
- 5. The analysis of data should be sufficiently adequate to reveal its significance and the method of analysis used should be appropriate.
- 6. Greater confidence in research is warranted.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

Qualities of good research :

- Good Research is systematic
- Good Research is logical
- Good Research is Empirical
- Good Research is Replicable

Problems Encountered by Researchers in India :

- > The lack of a scientific training in the methodology of research.
- > There is insufficient interaction between the university department.
- Most of the business units in our country do not have the confidence that the material supplied by them to the researcher will not be misused.
- Research studies overlapping one another are undertaken quite often for want of adequate information.
- > There does not exist a code of conduct.
- Many researcher in our country also face the difficulty of adequate and timely secretarial assistance.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

- > There is also the difficulty of timely availability of published data.
- Library management and functioning is not satisfactory.
- There is also the difficulty of timely availability of published data from various government and other agencies.
- > There may be at a time, take place the problem of conceptualization.

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 situation for the same. Usually we say that a research problem exist if one researcher find out the best solution for the research problem.

Selecting the problem:

- 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.
- > Controversial subject should not become the choice of averager.
- Too narrow or two vague problem should be avoided.
- > The subject selected for research should be familiar and feasible.
- The importance of the subject , the qualifications and the training of a researcher, the costs involved.
- > The selection of a problem must be proceeded by a preliminary study.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IVBATCH-2017-2019

Technique involved in defining a problem:

The techniques for the purpose involves the undertaking of the following step one another:

- 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 discussion.
- v) Rephrasing the research problem.

Research Design:

A research design is the arrangement of condition for collection and analysis of data in a manner that aims to combine relevance to the research process with economy in procedure .In fact, the research design is the conceptual structure within which research is conducted; it constitutes the blue print for the collection, measurement and analysis of the data.

Different research design:

- Research design in case of exploratory research studies.
- > Research design in case of descriptive and diagnostic research studies.
- > Research design in case of hypothesis testing research studies.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IVBATCH-2017-2019

Important Experimental Designs

Experimental design refers to the framework or structure of an experiment and as such there are

several experimental designs. We can classify experimental designs into two broad categories, viz., informal experimental designs and formal experimental designs. Informal experimental designs are those designs that normally use a less sophisticated form of analysis based on differences in magnitudes, whereas formal experimental designs offer relatively more control and use precise statistical procedures for analysis.

Important experiment designs are as follows:

- (a) Informal experimental designs:
- (i) Before-and-after without control design.
- (ii) After-only with control design.
- (iii) Before-and-after with control design.
- (b) Formal experimental designs:
- (i) Completely randomized design (C.R. Design).
- (ii) Randomized block design (R.B. Design).

We may briefly deal with each of the above stated informal as well as formal experimental designs.

1. Before-and-after without control design: In such a design a single test group or area is selected and the dependent variable is measured before the introduction of the treatment. The treatment is then introduced and the dependent variable is measured again after the treatment has been introduced. The effect of the treatment would be equal to the level of the phenomenon after the treatment minus the level of the phenomenon before the treatment.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

2. After-only with control design: In this design two groups or areas (test area and control area) are selected and the treatment is introduced into the test area only. The dependent variable is then

measured in both the areas at the same time. Treatment impact is assessed by subtracting the value

of the dependent variable in the control area from its value in the test area.

3. Before-and-after with control design: In this design two areas are selected and the dependent

variable is measured in both the areas for an identical time-period before the treatment. The treatment is then introduced into the test area only, and the dependent variable is measured in both for an identical time-period after the introduction of the treatment. The treatment effect is determined by subtracting the change in the dependent variable in the control area from the change in the dependent variable in test area.

4. Completely randomized design (C.R. design): Involves only two principles viz., the principle

of replication and the principle of randomization of experimental designs. It is the simplest possible design and its procedure of analysis is also easier. The essential characteristic of the design is that subjects are randomly assigned to experimental treatments (or vice-versa). For instance, if we have 10 subjects and if we wish to test 5 under treatment A and 5 under treatment B, the randomization process gives every possible group of 5 subjects selected from a set of 10 an equal opportunity of being assigned to treatment A and treatment B. One-way analysis of variance (or one-way ANOVA)* is used to analyse such a design. Even unequal replications can also work in this design. It provides maximum number of degrees of freedom to the error. Such a design is generally used when experimental areas happen to be homogeneous.

1. **Two-group simple randomized design:** In a two-group simple randomized design, first of all the population is defined and then from the population a sample is selected randomly. Further, requirement of this design is that items, after being selected randomly from the population, be randomly assigned to the experimental and control groups (Such random assignment of items to two groups is technically described as principle of randomization). Thus, this design yields two groups as representatives of the population.

(ii) Random replications design: The limitation of the two-group randomized design is usually eliminated within the random replications design. In the illustration just cited above, the teacher

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: IVBATCH-2017-2019

differences on the dependent variable were ignored, i.e., the extraneous variable was not controlled. But in a random replications design, the effect of such differences are minimised (or reduced) by providing a number of repetitions for each treatment. Each repetition is technically called a 'replication'. Random replication design serves two purposes viz., it provides controls for the differential effects of the extraneous independent variables and secondly, it randomizes any individual differences among those conducting the treatments.

5. Randomized block design (R.B. design) is an improvement over the C.R. design. In the R.B

design the principle of local control can be applied along with the other two principles of experimental designs. In the R.B. design, subjects are first divided into groups, known as blocks, such that within each group the subjects are relatively homogeneous in respect to some selected variable. The variable selected for grouping the subjects is one that is believed to be related to the measures to be obtained in respect of the dependent variable. The number of subjects in a given block would be equal to the number of treatments and one subject in each block would be randomly assigned to each treatment. In general, blocks are the levels at which we hold the extraneous factor fixed, so that its contribution to the total variability of data can be measured. The main feature of the R.B. design is that in this each treatment appears the same number of times in each block. The R.B. design is analysed by the two-way analysis of variance (two-way ANOVA)* technique.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
UNIT: IVBATCH-2017-2019

POSSIBLE QUESTIONS

Part-A(6 mark)

- 1. Briefly describe the different steps involved in a research process..
- 2. Explain the research process in detail.
- 3. What is the objective of doing a research?
- 4. Write the important concept relating research design
- 5. Explain about identifying the research problems.
- 6. Explain the meaning and significance of a research design.
- 7. What are the problems occurring in doing a research?
- 8. Explain the data processing and analysis.
- 9. Write about the types of research.
- 10. What are the features of a good design?

Part- B(10 mark)

1. Write about the types of research.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

UNIT-V

SYLLABUS

Sampling Design: Meaning – Concepts – Steps in sampling – Criteria for good sample design. Scaling measurements – Techniques – Types of scale.

SAMPLING DESIGN

The researcher should select a representative of the total population as possible in order to produce a miniature cross-section. The selected respondents constitute what is technically called a sample and the selection process is called sampling techniques. The survey so conducted is called a sample survey.

Sample design:

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items from the sample. Sample design is determined before the data collected.

STEPS IN SAMPLE DESIGN

While developing a sampling design, the researcher must pay attention to the following points:

(i) **Type of universe:** The first step in developing any sample design is to clearly define the set of objects, technically called the Universe, to be studied. The universe can be finite or infinite. In finite universe the number of items is certain, but in case of an infinite universe the number of items is infinite, i.e., we cannot have any idea about the total number of items. The population of a city, the number of workers in a factory and the like are examples of finite universes, whereas the number of stars in the sky, listeners of a specific radio programme, throwing of a dice etc. are examples of infinite universes.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

(ii) **Sampling unit:** A decision has to be taken concerning a sampling unit before selecting sample. Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual. The researcher will have to decide one or more of such units that he has to select for his study.

(iii) **Source list:** It is also known as 'sampling frame' from which sample is to be drawn. It contains the names of all items of a universe (in case of finite universe only). If source list is not available, researcher has to prepare it. Such a list should be comprehensive, correct, reliable and appropriate. It is extremely important for the source list to be as representative of the population as possible.

(iv) **Size of sample:** This refers to the number of items to be selected from the universe to constitute a sample. This a major problem before a researcher. The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfills the requirements of efficiency, representativeness, reliability and flexibility. While deciding the size of sample, researcher must determine the desired precision as also an acceptable confidence level for the estimate. The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed. The size of population must be kept in view for this also limits the sample size. The parameters of interest in a research study must be kept in view, while deciding the size of the sample. Costs too dictate the size of sample that we can draw. As such, budgetary constraint must invariably be taken into consideration when we decide the sample size.

(v) **Parameters of interest:** In determining the sample design, one must consider the question of the specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population. There may also be important sub-groups in the population about whom we would like to make estimates. All this has a strong impact upon the sample design we would accept.

(vi) **Budgetary constraint:** Cost considerations, from practical point of view, have a major impact upon decisions relating to not only the size of the sample but also to the type of sample. This fact can even lead to the use of a non-probability sample.

(vii) **Sampling procedure:** Finally, the researcher must decide the type of sample he will use i.e., he must decide about the technique to be used in selecting the items for the sample. In fact, this technique or

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

procedure stands for the sample design itself. There are several sample designs (explained in the pages that follow) out of which the researcher must choose one for his study. Obviously, he must select that design which, for a given sample size and for a given cost, has a smaller sampling error.

Criteria of Selecting A Sampling Procedure:

- Inappropriate Sampling Frame
- Defective Measuring Device
- Non-respondents
- Indeterminacy Principle
- Natural basis in the Reporting of Data.

Characterestics Of a Good Sample Design:

- Sample Design must result in truly representative sample.
- Sample design must be such which results in a small sampling error.
- Sample Design must be viable in the context of funds available for the research study.
- Sample design must be such so that systematic bias can be controlled in a better way.
- Sample should be such that the results of the sample study can be applied.

Complex Random Sampling Designs:

i) Systematic Sampling:

The most practical way of sampling is to select every ith item on a list.Sampling of this type is known as systematic sampling.Systematic sampling has a certain plus point.It can be taken as an

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

improvement over asimple random sample in as much as the systematic sample is spread more evenly over the entire populations.it is an Easier and less costlier method of sampling and can be conveniently used in even in case of large population.But there are certain dangers too in using this type of sampling.

ii) Stratified sampling:

If a population from which a sample is to be drawn does not constitute a homogenous group, stratified sampling is generally applied in order to obatin a representative sample.Under stratified sampling the population isdivided into several sub-populations that are individually more homogeneous than the population and then we select item from each stratum to constitute a sample.

iii) Cluster Sampling:

A sample can be taken into a number of smaller non-overlapping areas andthen to randomly select a number of these smaller areas with the ultimate sample consisting of all units in these small areas or clusters. Thus in cluter sampling the total population is divide into the number of relatively small subdivisions which are them clusters of still smaller units and then some of these clusters are randomly selected for inclusion in the overall sample.

iv) Area Sampling:

If cluster happen to be some geographic subdivisions, in that case cluster sampling is better known as area sampling. In other words, cluster designs, where the primary sampling unit represents acluster of units based on geographic area, are distinguished as area sampling.

v) Multi-Stage Sampling:

Multi-stage sampling is a further development of the principle of cluter sampling.Ordinarily multi-stage sampling is applied in big inquiries extending to a considerable large geographical area.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

vi) Sampling with probability proportion to size :

In case the cluster sampling units do not have the same number or approximately of elements, it is considered appropriate to use a random selection process wherethe probability of each cluster being included in the sample is proportional to the size of the clusters.

vii) Sequential Sampling:

This sampling design is some what complex sample design. The ultimate size of the sample under this technique is not fixed in advance, but is determined according to mathematical decision rules on the basis of information yielded as survey progresses. This is usually adopted in case of statistical quality control.

TECHNIQUE OF DEVELOPING MEASUREMENT TOOLS

The technique of developing measurement tools involves a four-stage process, consisting of the

following:

- (a) Concept development;
- (b) Specification of concept dimensions;
- (c) Selection of indicators; and
- (d) Formation of index.

The first and foremost step is that of concept development which means that the researcher should arrive at an understanding of the major concepts pertaining to his study. This step of concept development is more apparent in theoretical studies than in the more pragmatic research, where the fundamental concepts are often already established.

The second step requires the researcher to specify the dimensions of the concepts that he developed in the first stage. This task may either be accomplished by deduction i.e., by adopting a more or less intuitive approach or by empirical correlation of the individual dimensions with the total concept and/or the other concepts. For instance, one may think of several dimensions such as product reputation,

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

customer treatment, corporate leadership, concern for individuals, sense of social responsibility and so forth when one is thinking about the image of a certain company.

Once the dimensions of a concept have been specified, the researcher must develop indicators for measuring each concept element. Indicators are specific questions, scales, or other devices by which respondent's knowledge, opinion, expectation, etc., are measured. As there is seldom a perfect measure of a concept, the researcher should consider several alternatives for the purpose. The use of more than one indicator gives stability to the scores and it also improves their validity.

The last step is that of combining the various indicators into an index, i.e., formation of an index. When we have several dimensions of a concept or different measurements of a dimension,

we may need to combine them into a single index. One simple way for getting an overall index is to provide scale values to the responses and then sum up the corresponding scores. Such an overall index would provide a better measurement tool than a single indicator because of the fact that an "individual indicator has only a probability relation to what we really want to know. This way we must obtain an overall index for the various concepts concerning the research study.

Scaling

In research we quite often face measurement problem (since we want a valid measurement but may not obtain it), specially when the concepts to be measured are complex and abstract and we do not possess the standardised measurement tools. Alternatively, we can say that while measuring attitudes and opinions, we face the problem of their valid measurement. Similar problem may be faced by a researcher, of course in a lesser degree, while measuring physical or institutional concepts. As such we should study some procedures which may enable us to measure abstract concepts more accurately. This brings us to the study of scaling techniques.

Meaning of Scaling

Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude and other concepts. This can be done in two ways viz., (i) making a judgement about some characteristic of an individual and then placing him directly on a scale that has been defined in terms of that characteristic and (ii) constructing questionnaires in such a way that the score of individual's responses assigns him a place on a scale. It may be stated here that a scale is a continuum, consisting of the highest point (in terms of some characteristic e.g., preference, favourableness, etc.) and the lowest point along

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

with several intermediate points between these two extreme points. These scale-point positions are so related to each other that when the first point happens to be the highest point, the second point indicates a higher degree in terms of a given characteristic as compared to the third point and the third point indicates a higher degree as compared to the fourth and so on. Numbers for measuring the distinctions of degree in the attitudes/opinions are, thus, assigned to individuals corresponding to their scale-positions. All this is better understood when we talk about scaling technique(s). Hence the term 'scaling' is applied to the procedures for attempting to determine quantitative measures of subjective abstract concepts. Scaling has been defined as a "procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some of the characteristics of numbers to the properties in question.

Scale Classification Bases

The number assigning procedures or the scaling procedures may be broadly classified on one or more of the following bases: (a) subject orientation; (b) response form; (c) degree of subjectivity; (d) scale properties; (e) number of dimensions and (f) scale construction techniques. We take up each of these separately.

(a) Subject orientation: Under it a scale may be designed to measure characteristics of the respondent who completes it or to judge the stimulus object which is presented to the respondent. In respect of the former, we presume that the stimuli presented are sufficiently homogeneous so that the between stimuli variation is small as compared to the variation among respondents. In the latter approach, we ask the respondent to judge some specific object in terms of one or more dimensions and we presume that the between-respondent variation will be small as compared to the variation among the different stimuli presented to respondents for judging.

(b) Response form: Under this we may classify the scales as categorical and comparative. Categorical scales are also known as rating scales. These scales are used when a respondent scores some object without direct reference to other objects. Under comparative scales, which are also known as ranking scales, the respondent is asked to compare two or more objects. In this sense the respondent may state that one object is superior to the other or that three models of pen rank in order 1, 2 and 3. The essence of ranking is, in fact, a relative comparison of a certain property of two or more objects.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

(c) Degree of subjectivity: With this basis the scale data may be based on whether we measure subjective personal preferences or simply make non-preference judgements. In the former case, the respondent is asked to choose which person he favours or which solution he would like to see

employed, whereas in the latter case he is simply asked to judge which person is more effective in some aspect or which solution will take fewer resources without reflecting any personal preference.

(d) Scale properties: Considering scale properties, one may classify the scales as nominal, ordinal, interval and ratio scales. Nominal scales merely classify without indicating order, distance or unique origin. Ordinal scales indicate magnitude relationships of 'more than' or 'less than', but indicate no distance or unique origin. Interval scales have both order and distance values, but no unique origin. Ratio scales possess all these features.

(e) Number of dimensions: In respect of this basis, scales can be classified as 'unidimensional'

and 'multidimensional' scales. Under the former we measure only one attribute of the respondent or object, whereas multidimensional scaling recognizes that an object might be described better by using the concept of an attribute space of 'n' dimensions, rather than a single-dimension continuum

(f) Scale construction techniques: Following are the five main techniques by which scales can be developed.

(i) Arbitrary approach: It is an approach where scale is developed on ad hoc basis. This is the most widely used approach. It is presumed that such scales measure the concepts for which they have been designed, although there is little evidence to support such an assumption.

(ii) Consensus approach: Here a panel of judges evaluate the items chosen for inclusion in the instrument in terms of whether they are relevant to the topic area and unambiguous in implication.

(iii) Item analysis approach: Under it a number of individual items are developed into a test which is given to a group of respondents. After administering the test, the total scores are calculated for every one. Individual items are then analysed to determine which items discriminate between persons or objects with high total scores and those with low scores.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

(iv) Cumulative scales are chosen on the basis of their conforming to some ranking of items with ascending and descending discriminating power. For instance, in such a scale the endorsement of an item representing an extreme position should also result in the endorsement of all items indicating a less extreme position.

(v) Factor scales may be constructed on the basis of intercorrelations of items which indicate that a common factor accounts for the relationship between items. This relationship is typically measured through factor analysis method.

Important Scaling Techniques

We now take up some of the important scaling techniques often used in the context of research specially in context of social or business research.

Rating scales: The rating scale involves qualitative description of a limited number of aspects of a thing or of traits of a person. When we use rating scales (or categorical scales), we judge an object in absolute terms against some specified criteria i.e., we judge properties of objects without reference to other similar objects.

These ratings may be in such forms as "like-dislike", "above average, average, below average", or other classifications with more categories such as "like very much—like some what—neutral—dislike somewhat—dislike very much"; "excellent—good—average—below average—poor", "always—often—occasionally—rarely—never", and so on. There is no specific rule whether to use a two-points scale, three-points scale or scale with still more points. In practice, three to seven points scales are generally used for the simple reason that more points on a scale provide an opportunity for greater sensitivity of measurement. Rating scale may be either a graphic rating scale or an itemized rating scale.

(i) The graphic rating scale is quite simple and is commonly used in practice. Under it the various points are usually put along the line to form a continuum and the rater indicates his rating by simply making a mark (such as ü) at the appropriate point on a line that runs from one extreme to the other. Scale-points with brief descriptions may be indicated along the line, their function being to assist the rater in performing his job. The following is an example of five-points graphic rating scale when we wish to ascertain people's liking or disliking any product: position along the line which fact may increase the difficulty of analysis. The meanings of the terms like "very much" and "some what" may depend upon

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

respondent's frame of reference so much so that the statement might be challenged in terms of its equivalency. Several other rating scale variants (e.g., boxes replacing line) may also be used.

(ii) The itemized rating scale (also known as numerical scale) presents a series of statements from which a respondent selects one as best reflecting his evaluation. These statements are ordered progressively in terms of more or less of some property. An example of itemized scale can be given to illustrate it. Suppose we wish to inquire as to how well does a worker get along with his fellow workers? In such a situation we may ask the respondent to select one, to express his opinion, from the following:

- He is almost always involved in some friction with a fellow worker.
- He is often at odds with one or more of his fellow workers.
- He sometimes gets involved in friction.
- He infrequently becomes involved in friction with others.
- He almost never gets involved in friction with fellow workers.

The chief merit of this type of scale is that it provides more information and meaning to the rater, and thereby increases reliability. This form is relatively difficult to develop and the statements may not say exactly what the respondent would like to express. Rating scales have certain good points. The results obtained from their use compare favourably with alternative methods. They require less time, are interesting to use and have a wide range of applications. Besides, they may also be used with a large number of properties or variables. But their value for measurement purposes depends upon the assumption that the respondents can and do make good judgements. If the respondents are not very careful while rating, errors may occur. Three types of errors are common viz., the error of leniency, the error of central tendency and the error of hallo effect. The error of leniency occurs when certain respondents are either easy raters or hard raters. When raters are reluctant to give extreme judgements, the result is the error of central tendency. The error of hallo effect or the systematic bias occurs when the rater carries over a generalised impression of the subject from one rating to another. This sort of error takes place when we conclude for example, that a particular report is good because we like its form or that someone is intelligent because he agrees with us or has a pleasing personality. In other words, hallo effect is

likely to appear when the rater is asked to rate many factors, on a number of which he has no evidence for judgement.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

Measurement and scaling:

Measurement is a process of mapping aspects of a domain onto the other aspects of a range according to some rule of correspondence. In measuring, we devise some form of the scale in the range and then transfer or map the properties of object from the domain on to this scale.

Measurement Scale:

The most widely used classification of measurement scales are

a)Nominal Scale:

Nominal scale is simply a system of assigning number symbols to events in order to label them. The usual example of this is the assignment of numbers of basket ball players in order to identify them such numbers cannot be considered to be associated with an ordered scale for their order is of no consequence. The numbers are just convenient labels for the particular class of events and as such have no quantitative value. Nominal scales provide convenient ways of keeping track of people, object and events. Nominal scale is the least powerful scale of measurement.

b)Ordinal scale:

The lowest level of the ordered scale that is commonly used is the ordinal scale. The ordinal scales places event in order, but there is no attempt to make the intervals of the scale equal in terms of some rule.

c) Interval scale:

In the case of interval scale, the scales are adjusted in terms of some rule that has been established as a bazsis for making the units equal. The units are equal only in so far as one accepts the assumption on which the rule is based. Interval scales can have an arbitrary zero, but is not possible to determine for them what may be called an absolute zero or the unique origin.

CLASS: II M.Sc MicrobiologyCOURSE NAME: Biostatistics and Research MethodologyCOURSE CODE: 17MBP304UNIT: VBATCH-2017-2019

d) Ratio scale:

Ratio scales have an absolute true zero of measurement. The absolute zero is not as precise as it was once believed to be ratio scazles represents the actual amounts of variables.

Sources of error in measurement:

Measurement should be precise and unambiguous in an ideal research study. The type of error in measurements are

i) Respondent:

At times the respondent may be reluctant to express strong or negative feelings or it is just possible that he may have very little knowledge but may not admit his ignorance.

ii) Situation:

Situational factors may also care in the way of correct measurement. Any condition which places a strain or interview can have serious effects on the interview-respondent rapport.

iii) Measurer:

The interviewer can distort responses by rewording or reordering questions. His behaviour, style and looks may encourage or discourage certain replies from respondents.

iv) Instrument:

Error may arise because of the defective measuring instrument.

Test of Sound Measurement:

i) Test of validity:

Validity is the most critical criterion and indicates the degree to which an instrument measure what it is suppose to measure. Validity can also be thought as utility.

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

ii) Test of Reliability:

The test of reliability is another important tests of sound measurement. A measuring instrument is reliable if it provides consistent results.

iii) Test of practicality:

The practicality characterestics of a measuring instrument can be judged in terms of economy, convenience and interpretability.

Meaning of scaling:

Scaling describes the procedures of assigning numbers to various degree of opinion, attitude and other concepts. This can be done in two ways

- i) Making a judgement about some characteristics of an individual and then placing him directly on a scale that has been defined in terms of that characteristics.
- ii) Constructing Questionnaires in such a way that the score of individuals responses assigns him a place on a scale.

Scale classification Bases:

- Subject orientation
- Response form
- Degree of subjectivity
KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II M.Sc Microbiology
COURSE CODE: 17MBP304COURSE NAME: Biostatistics and Research Methodology
BATCH-2017-2019

- Scale properties
- Number of Dimension
- Scale Construction Techniques

Important Scaling techniques:

- Rating scales
- Method of paired comparison
- Method of rank order.

Prepared by Y.Sangeetha, Asst Prof, Department of Mathematics, KAHE

KARPAGAM ACADEMY OF HIGHER EDUCATION

CLASS: II M.Sc Microbiology C COURSE CODE: 17MBP304

COURSE NAME: Biostatistics and Research Methodology UNIT: V BATCH-2017-2019

POSSIBLE QUESTIONS :

Part-B(6 mark)

- 1. Explain the criteria of good research.
- 2. Explain about the types of scaling and rating.
- 3. What are the features of good research design?
- 4. Explain the steps in sampling.
- 5. Describe the criteria for good sample design.
- 6. Explain about the types of scaling
- 7. Write about the measurements in research.
- 8. Describe the criteria for good sample design.
- 9. Explain errors in measurement.
- 10. Explain about scaling techniques.
- 11. Describe the criteria for good sample design.

Part-C(10 mark)

1. Describe the criteria for good sample design.

Prepared by Y.Sangeetha, Asst Prof, Department of Mathematics, KAHE

KARPAGAM ACADEMY OF HIGHER ED DEPARTMENT OF MATHEMATI(BIOSTATISTICS AND RESEARCH METHODOLO

Questions	choice 1	choice 2
	UNI	T - I
The word statistics is used as	singular word	a plural word
Classification is a process of arranging data in	grouping of related facts in different classes	different rows
To represent two or more sets of interrelated data, we use	bar diagram	pie diagram
Histogram is a graph of	Time series	frequency distribution
Univariate data consists of	one variable	two variables
Data are generally obtained from	Primary sources	Secondary sources
In geographical classification data are classified on the basis of	area	attributes
In qualitative classification data are classified on the basis of	area	attributes
In quantitative classification data are classified on the basis of	- area	attributes
Number of source of data is	2	3
Squares and rectangles are	Two dimensional diagram	One dimensional diagram
Data originally collected for an investigation is known as	- Tabulation	Primary data
The heading of a row in a statistical table is known as	stub	caption
Statistics can	prove anything	disprove anything
Statistics is also a science of Statistics is	estimates quantitative science	both a and b a qualitative

Statistics considers	a single item	a set of item
Statistics can be considered as	an art	a science
The other name of cumulative frequency curve is	Ogive	Bars
Number of methods of collection of primary data is	2	3
Number of questions in a questionnaire should be	- 5	10
Sources of secondary data are	Published sources	Unpublished sources
Compared with primary data, secondary data are	more reliable	less reliable
are column headings	stub	heading
Mid value=	lower boundary/2	upper boundary/2
The origin of the word statistics has been traced to the Latin word	• statista	status
Graphs of frequency distribution are	histogram	pie diagram
cubes are	Two dimensional diagram	One dimensional diagram
is the difference between the value of the smallest item and the value of the largest item.	class interval	frequency
is one which is used by the individual or agency which	primary data	secondary data
Exclusive class intervals suit	discrete variables	continuous
A table is a systematic arrangement of statistical data in	- columns	rows
- The collected data in any statistical investigation are known as	raw data	arranged data
The emitting form of a frequency polygon is called	- histogram	ogive
In chronological classification data are classified on the basis of	time	attributes
Bar diagrams are dimensional diagrams Diagrams and graphs are tools of In a two dimensional diagram	two collection of data only height is considered	three presentation only width is considered
Which one of the following is a measure of central tendency?	Median	range
The total of the values of the items divided by their number of items is known as	Median	Arithmetic mean
In the short-cut method of arithmetic mean, the deviation is taken as	x – A	A - x
The sum of the deviations of the values from their arithmetic mean is	- 1	one

The formula for the weighted arithmetic mean is Find the Mean of the following values, 5, 15, 20, 10, 40	$\sum wx / \sum w$	$\sum w / \sum wx$
Which of the followings represents median?	First quartile	Third quartile
Which of the measure of central tendency is not affected by extreme values?	Mode	Median
Sum of square of the deviations about mean is	Maximum	one
Median is the value of item when all the items are in order of magnitude.	First	second
Find the Median of the following data 160, 180, 175, 179, 164, 178, 171, 164, 176.	160	175
The position of the median for an individual series is taken as	(N + 1) / 2	(N + 2) / 2
Mode is the value, which has	Average frequency density	less frequency density
A frequency distribution having two modes is said to be Mode has stable than mean.	unimodal less	bimodal more
Which of the following is not a measure of dispersion?	Range	quartile deviation
Range of the given values is given by	L- S	L+S
Which one of the following is relative measure of dispersion?	Range	Q.D
Coefficient of variation is defined as	(AM * 100)/S.D	(S.D* 100)/A.M
If the values of median and mean are 72 and 78 respectively, then find the mode.	16	60
Find Mean for the following 3, 4, 5.	4.25	2.25
The coefficient of range	L-S/L+S	L+S /L-S
Second quartile is also called as	Mode	mean
If A.M = 8, N=12, then find $\sum X$.	76	80
If the value of mode and mean is 60 and 66 then, find the value of	64	46
median. The formula for median for continuous series is	M = (N+1) / 2	M = L + [(N/2 + cf)/f] * i
Median is	Average point	Midpoint
Mode is the value which	Is a mid point	Occur the most
Is known as positional average	Median	Mean
The median of marks 55, 60, 50, 40, 57, 45, 58, 65, 57, 48 of 10 students is	55	57
The middle most value of a frequency distribution table is known as	Mean	Median
The middle most value of a frequency distribution table is known as	Mean	Median
Measures of central tendency is also known as	Dispersion	averages
From the given data 35,40,43,32,27 the coefficient of range is	23	0.23
If $S.D = 6$, then find variance.	6	36
Which one of the following shows the relation between variance and standard deviation?	var = square root of S.D	S.D = square root of variance
If variance is 64, then find S.D.	8	13
Which of the following measures of averages divide the observation into two parts	Mean	Median

Which of the following measures of averages divide the observation into four equal parts	Mean	Median
Arithmetic mean of the series 1, 3, 5, 7, 9 is	5	6
Arithmetic mean of the series 3, 4, 5, 6, 7 is	5.5	6
The Arithmetic mean for the series 3, 5, 5, 2, 6, 2, 9, 5, 8, 6, is	5	6
The median value for the series 3, 5, 5, 2, 6, 2, 9, 5, 8, 6 is	6	5
The mode for the series 3, 5, 6, 2, 6, 2, 9, 5, 8, 6 is	5	6
The Arithmetic mean for the series 51.6, 50.3, 48.9, 48.7, 48.5 is	49.8	50
The Median for the series 51.6, 50.3, 48.9, 48.7, 49.5, is	49.8	50
The Arithmetic mean for the series 51.6, 50.3, 48.9, 48.7, 49.5 is	49.8	50
The Mode for the series 51.6, 50.3, 48.9, 48.7, 49.5 is	49.8	50
If standard deviation is 5, then the variance is	5	625
Standard deviation is also called as	Root mean square deviation	mean square deviation

UCATION CS)GY (I7MBP304)

choice 3	choice 4	Answer
both singular and plural words	neither singular nor plural word	both singular and plural words
different columns and rows	different columns grouping of related facts in different classes	grouping of related facts in different classes
histogram cumulative frequency distribution three variable	multiple bar diagram normal distribution four	multiple bar diagram frequency distribution one variable
Both primary and secondary sources	neither primary nor secondary sources	Both primary and secondary sources
time	location	area
time	location	attributes
time	magnitude	magnitude
4 Three dimensional	1 Multi dimensional diagram	2 Two dimensional diagram
diagram Secondary data	Published data	Primary data
title	heading	stub
neither prove nor disprove anything but it is	none of these	neither prove nor disprove anything but it is just a tool
probabilities both quantitative and qualitative science	neither a nor b neither quantitative nor qualitative	both a and b both quantitative and qualitative science

either a single item or a set of	neither a single item or a set of	a set of item
item	item	
both an art and	neither an art nor	both an art and
science	a science	science
Histogram	Pie diagram	Ogive
4	5	5
maximum	minimum	minimum
Either Published	primary source	Either Published
sources or	1 2	sources or
Unpublished		Unpublished
sources		sources
equally reliable	uniformly reliable	less reliable
bar	captions	captions
lower	lower boundary+	lower boundary+
boundary+	upper boundary	upper boundary)/2
boundary)/2		
statistik	statistique	status
bar chart	circle	histogram
Three	Multi dimensional	Three dimensional
dimensional	diagram	diagram
diagram	e	6
number of items	range	range
both		primary data
both	neither	continuous
		variables
both columns	stubs	both columns and
and rows		rows
classified data	tabulated data	raw data
bar diagram	frequency curve	frequency curve
class intervals	location	time
one	multi	one
analysis	summarization	presentation
height,width and	Both height and	only height is
thickness are	width are	considered
considered	considered	
variation	correlation	Median
mode	range	Arithmetic mean
(x - A) / c	(A - x) / c	$\mathbf{x} - \mathbf{A}$
two	zero	zero

∑x / n 41	$\sum x / \sum f$ 20	$\sum wx / \sum w$ 18
Second quartile	Q.D	Second quartile
sixth deciles	Mean	Median
zero	Minimum	Minimum
Middle most	last	Middle most
176	180	175
N/2	N/4	(N + 1) / 2
greatest frequency density	graetest frequency	greatest frequency density
trimodal	modal	bimodal
same	most	less
standard deviation	median	median
S+L	LS	L-S
S.D	coefficient of	coefficient of
S.D/A.M	(1/S.D)*100	(S.D* 100)/A.M
70	76	60
3	2.28	3
L-S	L+S	L-S /L+S
median	G.M	median
86	96	96
54	44	64
M =L - (N/2+cf)/f* i	M = L + [(N/2 - cf) / f] * i	M = L + [(N/2 - cf) / f] * i
Most likely	Most remote	Midpoint
point	point	Mapolin
Average of all	Likely	Occur the most
Mode	Range	Median
52.5	56	56
Mode	Range.	Median
Mode	Range	Median
correlation	tendency	correlation
13	0.13	13
42	12	36
variance = S.D	variance / $S.D = 1$	S.D = square root of variance
14	11	8
Mode	Range	Median

Mode	Quartile	Quartile
5.5	6.5	5
5	6.5	5
5.5	6.5	5
5.5	6.5	5
5.5	6.5	6
48.9	49.6	49.8
48.9	49.6	49.6
48.9	49.6	48.9
48.9	49.6	48.9
25	2.23068	25
Root deviation	Root median square deviation	Root mean square deviation

KARPAGAM ACADEMY OF HIGHER EDUCATIC DEPARTMENT OF MATHEMATICS BIOSTATISTICS AND RESEARCH METHODOLOGY (17)

Questions	choice 1	choice 2	choice 3
	UN	II - II	
The regression line cut each other at the point of	Average of X only	Average of Y only	Average of X and Y
Given the coefficient of correlation being 0.8, the coefficient of determination will be	0.98	0.64	0.66
Given the coefficient of correlation being 0.9, the coefficient of determination will be	0.98	0.81	0.66
If the coefficient of determination being 0.49, what is the coefficient of correlation	0.7	0.8	0.9
Given the coefficient of determination being 0.36, the coefficient of correlation will be Which one of the following refers the term Correlation?	0.3	0.4	0.6
	Relationship between two values	Relationship between two variables	Average relationship between two variables
If $r = +1$, then the relationship between the given two variables is	perfectly positive	perfectly negative	no correlation
If $r = -1$, then the relationship between the given two variables is	perfectly positive	perfectly negative	no correlation
variables is	Perfectly positive	perfectly negative	no correlation
Coefficient of correlation value lies between	1 and -1	0 and 1	0 and ∞
While drawing a scatter diagram if all points appear to form a straight line getting Downward from left to right, then it is inferred that there is	Perfect positive correlation	simple positive correlation	Perfect negative correlation
The range of the rank correlation coefficient is	0 to 1	-1 to 1	0 to ∞
If $r = 1$, then the angle between two lines of regression is	Zero degree	sixty degree	ninety degree
Regression coefficient is independent of	Origin	scale	scale
is negative, then the Regression coefficient of Y on X is	Positive	negative	not certain
If the correlation coefficient between two variables X and Y is positive, then the Regression coefficient of X on Y is	Positive	negative	not certain
There will be only one regression line in case of two variables if	r =0	r = +1	r = -1
The regression line cut each other at the point of	Average of X only	Average of Y only	Average of X and Y
It b_{xy} and b_{yx} represent regression coefficients and if $b_{yx} > 1$ then b_{xy} is	Less than one	greater than one	equal to one

Rank correlation was discovered by	R.A.Fisher	Sir Francis Galton	Karl Pearson
Formula for Rank correlation is	1- ($6\Sigma d^2 / (n(n2-1)))$	1- ($6\Sigma d^2$ /(n(n2+1)))	$\begin{array}{c} 1+ \ (\ 6\Sigma d^2 \ /(\\ n(n2+1))) \end{array}$
With $b_{xy}=0.5$, $r = 0.8$ and the variance of Y=16, the standard deviation of X=	6.4	2.5	10
The coefficient of correlation $r =$	$(b_{xy}, b_{yx})^{1/4}$	$(b_{xy} b_{yx})^{-1/2}$	$(b_{xy}, b_{yx})^{1/3}$
If two regression coefficients are positive then the coefficient of correlation must be	Zero	negative	positive
If two-regression coefficients are negative then the coefficient of correlation must be	Positive	negative	zero
The regression equation of X on Y is	X = a + bY	X = a + bX	X = a - bY
The regression equation of Y on X is	X = a + bY	X = a + bX	X = a - bY
The given two variables are perfectly positive, if	r = +1	r = -1	$\mathbf{r} = 0$
The relationship between two variables by plotting the values on a chart, known as-	coefficient of correlation	Scatter diagram	Correlogram
If x and y are independent variables then,	$cov(x,y) \neq 0$	cov(x,y)=1	cov(x,y)=0
Correlation coefficient is the of the two regression	Mode	Geometric	Arithmetic
$b_{rrr} = 0.4, b_{rrr} = 0.9$ then r =	0.6	0.3	0.1
b = 1/5 r = 8/15 s = 5 then s =	40/13	13/40	40/3
The geometric mean of the two regression coefficients.	Correlation	regression	coefficient of range
If two variables are uncorrelated, then the lines of regression	Do not exist	coincide	Parallel to each other
If the given two variables are correlated perfectly negative, then	r = +1	r = -1	$\mathbf{r} = 0$
If the given two variables have no correlation, then	r = +1	r = -1	$\mathbf{r} = 0$
If the correlation coefficient between two variables X and Y is, the Regression coefficient of Y on X is positive	Negative	positive	not certain
If the correlation coefficient between two variables X and Y is, the Regression coefficient of Y on X is negative	Negative	positive	not certain
is independent of origin and scale.	Correlation coefficient	regression coefficients	coefficient of range
The angle between two lines of regression is ninety degree, if	r = 2	$\mathbf{r} = 0$	$\mathbf{r} = 1$
is used to measure closeness of relationship between variables.	Regression	mean	Rank correlation
If r is either $+1$ or -1 , then there will be only one line in case of two variables	Correlation	regression	rank correlation
When $b_{xy}=0.85$ and $b_{yx}=0.89$, then correlation coefficient r =	0.98	0.5	0.68
If b_{xy} and b_{yx} represent regression coefficients and if $b_{xy} < 1$, then b_{yx} is While drawing a scatter diagram if all points appear to form	less than 1	greater than one	equal to one
a straight line getting			
Downward from left to right, then it is inferred that there is	Perfect positive correlation	simple positive correlation	Perfect negative correlation
If r =1, the angle between two lines of regression is	Zero degree	sixty degree	ninety degree

Regression coefficient is independent of			both origin and
5 1	Origin	scale	scale
There will be only one regression line in case of two			
variables if	r =0	r = +1	r = -1

)N

MBP304)

choice 4	Answer
d) the median of X on Y	Average of X and Y
0.54	0.64
0.54	0.81
0.6	0.7
0.5	0.6
Relationship between two things	Relationship between two variables
high positive	perfectly positive
low Positive	perfectly negative
both positive and negative	no correlation
0 and -1.	1 and -1
no correlation	Perfect negative correlation
$-\infty$ to ∞	-1 to 1
thirty degree	ninety degree
neither origin nor scale.	Origin
zero	negative
zero	Positive
r is either $+1$ or -1	r =0
the median of X on Y	Average of X and Y
equal to zero	Less than one

Spearman	Spearman
1 /(n(n2-1))	1- (6Σd ² /(n(n2-1)))
25.6	2.5
$(b_{xy.} b_{yx})^{1/2}$	$(b_{xy.} b_{yx})^{1/2}$
one	positive
one	Positive
Y = a + bX	X = a + bY
Y = a + bX	Y=a+bX
$r \neq +1$	r = +1
rank correlation	Scatter diagram
cov(x,y) > 1	cov(x,y)=0
median	Geometric mean
-0.6	0.6
3	40/3
coefficient of	Correlation
variation	coefficient
perpendicular to	perpendicular to
each other	each other
$r \neq +1$	r = -1
$r \neq +1$	$\mathbf{r} = 0$
r≠+1 zero	r = 0 positive
r≠+1 zero	r = 0 positive
r ≠ +1 zero zero	r = 0 positive Negative
r ≠ +1 zero zero coefficient of	r = 0 positive Negative Correlation
r ≠ +1 zero zero coefficient of variation	r = 0 positive Negative Correlation coefficient
$r \neq +1$ zero zero coefficient of variation r = -1	 r = 0 positive Negative Correlation coefficient r = 1
$r \neq +1$ zero zero coefficient of variation r = -1 correlation	 r = 0 positive Negative Correlation coefficient r = 1 correlation
$r \neq +1$ zero zero coefficient of variation r = -1 correlation mean	 r = 0 positive Negative Correlation coefficient r = 1 correlation regression
$r \neq +1$ zero zero coefficient of variation r = -1 correlation mean 0.87	<pre>r = 0 positive Negative Correlation coefficient r = 1 correlation regression 0.87</pre>
$r \neq +1$ zero zero coefficient of variation r = -1 correlation mean 0.87 equal to zero	<pre>r = 0 positive Negative Correlation coefficient r = 1 correlation regression 0.87 greater than one</pre>
$r \neq +1$ zero zero coefficient of variation r = -1 correlation mean 0.87 equal to zero	r = 0 positive Negative Correlation coefficient r = 1 correlation regression 0.87 greater than one
r ≠ +1 zero zero coefficient of variation r = -1 correlation mean 0.87 equal to zero no correlation	r = 0 positive Negative Correlation coefficient r = 1 correlation regression 0.87 greater than one Perfect negative correlation

neither origin nor	
scale.	Origin
r is either +1 or –1	r =0

KARPAGAM ACADEMY OF HIGHER EDUCAT DEPARTMENT OF MATHEMATICS BIOSTATISTICS AND RESEARCH METHODOLOGY (

Questions	choice 1	choice 2 UNIT - III	choice 3
Completely randomized design is similar			
to	three way	one way	two way
Randomized block design is similar to			
	two way	three way	one way
ANOVA is the technique of analysis of	standard		
-	deviation	variance	mean
Under one way classification, the			
influence of only attribute or factor is			
considered	two	three	one
Under two way classification, the			
influence of only attribute or factor is	2		
considered	four	two	three
The word is used to indicate			
various statistical measures like mean,			
standard deviation, correlation etc, in the			1 .1 .
universe.	Statistic	parameter	hypothesis
The term STATISTIC refers to the			
statistical measures relating to the		1 (1)	1
·	Population	nypotnesis	sample
A hypothesis may be classified as	Simula	Commencie	
Level of significance is the methodility of	Simple	Composite	null Not committing
Level of significance is the probability of	Tuno Lorror	Tuno II orror	Not committing
Degrees of freedom are related to	Type Tentor	Type II ellor	No. of
	No of		independent
	observations	hypothesis	observations in
	in a set	under test	a set
A critical function provides the basis for	Accenting	under test	no decision
	H	rejecting H _o	about H _o
Student's t test is applicable in ease of	110	for sample of	uoout 110
student s t-test is applicable in case of	Small	size between	
	samples	5 and 30	I arge samples
Student's t-test is applicable only when	The variate	the variable is	Large samples
	values are	distributed	The sample is
·	independent	normally	not large
If the calculated value is less than the table	maepenaem	nonnuny	nov im ge
value then we accept the			
hypothesis.			
<i>y</i> ₁	Alternative	null	both

Small sample test is also known as			
	Exact test	t – test	normal test
The formula for c ² is	å(O–E) ² /E	å(E+O) ² /E	å(О-Е) /Е
If a statistic 't' follows student's t distribution with n degrees of freedom then t^2 follows	c ² distributio n with (n-1) degrees of freadom	c ² distribution with n degrees	c^{2} distribution with n^{2} degrees of freedom
The distribution used to test goodness of	E	or needoni	needom
fit is	r distribution	c ² distribution	t distribution
Degree of freedom for statistic chi-square incase of contingency table of order 2x2 is-			
	3	4	2
Larger group from which the sample is			
drawn is called	Sample	sampling	universe
Any hypothesis concerning a population is	G 1	1	statistical
called a	Sample	population	measure
Rejecting Ho when it is true leads	Type I error	Type II error	correct decision
Accept Ho when it is true leads	• •	• •	
-	Type I error	Type II error	correct decision
Type II error occurs only if	Reject Ho	Accept Ho	
	when it is	when it is	Accept Ho
	true	false	when it is true
The correct decision is	Reject Ho	Accept Ho	
	when it is	when it is false	Reject Ho when it is false
The maximum probability of committing		Tuibe	
type I error, which we specified in a test is			
known as	Null	alternative	
	hypothesis	hypothesis	DOF
If the computed value is less than the	Null	Null	Alternative
critical value, then	hypothesis is	hypothesis is	hypothesis is
	accepted	rejected	accepted
If the computed value is greater than the	Null	Null	Alternative
critical value, then	nypotnesis is	nypotnesis is	nypotnesis is
In sampling distribution the standard arror	accepted	Tejecteu	accepted
in sampling distribution the standard error	nn	na	nna
If the sample size is greater than 30 then	I arge	pq	прч
the sample is called	sample	small sample	nonulation
If the sample size is less than 30 then the	Large	sman sample	ropaiation
sample is called	sample	small sample	population
Z – test is applicable only when the			
sample size is	zero	one	small

The degrees of freedom for two samples in			
t – test is	$n_1 + n_2 + 1$	$n_1 + n_2 - 2$	$n_1 + n_2 + 2$
An assumption of $t - test$ is population of			
the sample is	Binomial	Poisson	normal
The degrees of freedom of chi – square			
test is	(r-1)(c-1)	(r+1)(c+1)	(r+1)(c-1)
In chi – square test, if the values of			
expected frequency are less than 5, then			
they are			
combined together with the neighbouring	Goodness of		
frequencies. This is known as	fit	DOF	LOS
The expected frequency of chi - square	(RT + CT) /	(RT - CT) /	
test can be calculated as	GT	GT	(RT * CT) / GT
In F – test, the variance of population from			
which samples are drawn are			
1	equal	not equal	small
If the data is given in the form of a series	1	1	
of variables, then the DOF is	n	n-1	n+1
The characteristic of the chi–square test is			
	DOF	LOS	ANOVA
If $S^2 > S^2$ then the E statistic is	201		
If $S_1 > S_2$, then the F – statistic is	C / C	C / C	$\mathbf{S}^2 / \mathbf{S}^2$
	s_1 / s_2	s_2 / s_1	s_1 / s_2
The value of Z test at 5% level of			
significance is	3.96	2.96	1.96
In, the variance of population from		Chi-Square	
which samples are drawn are equal	t-test	test	Z-test
F – statistics is	Variance	Variance	Variance
	between the	within the	between the
	samples /	samples /	rows /
	variance	variance	variance
	within the	between the	between the
	samples	samples	columns
Analysis of variance utilizes:		Chi-Square	
	t-test	test	Z-test
F – test whish is also known as	Chi-Square		varience ratio
	test	Z-test	test
The technique of analysis of variance			
refered to as	ANOVA	F-test	Z-test
The two variations, variation within the			
samples and variations between the			
samples	Chi- square		
are tested for their significance by	test	F – test	t-test
0			

Under classification , the influence			
of only one attribute or factor is			
considered.	two way	three way	one way
Under classification , the influence			
of two attribute or factors is considered			
	two way	three way	one way

ION

I7MBP304) choice 4	Answer
t test	one way
many	two way
range	variance
many	one
one	two
none of these	parameter
universe	sample
all the above	all the above
any of the above	above
No. of dependent	independent
observations in a	observations
set	in a set
all the above	all the above
	Small
none of the above	samples
all the above	all the above
sample	null

F-test	t – test
å(O-E) ² /O	å(O–E) ² /E
c ² distribution	c ² distributio
with (n+1)	n with (n-1)
degrees of	degrees of
freedom	freedom
	c ² distributio
Z distribution	n
1	1
noromotor	universe
statistical	statistical
hypothesis	hypothesis
nypomesis	nypotnesis
either (a) or (b)	Type I error
	correct
either (a) or (b)	decision
	Accept Ho
reject Ho when it	when it is
is false	false
	Reject Ho
	when it is
none of these	false
level of	level of
significance	significance
	Null
	hypothesis is
population	accepted
	Null
	hypothesis is
small sample	rejected
sart(nna)	sart(npa)
יזי יוייץי	Large
Null hypothesis	sample
alternative	1
hypothesis	small sample
-	-
large	large

$n_1 + n_2 - 1$	$n_1 + n_2 - 2$
exponential	normal
(r-1)(c+1)	(r-1)(c-1)

pooling (RT*CT)	pooling (RT * CT) / GT		
large	equal		
(r – 1)(c – 1)	n-1		
independence of	e of		
attributes	attributes		
S_1^{3} / S_2^{3}	S_1^2 / S_2^2		
0.96	1.96		
F-test	F-test Variance		
Variance within	between the		
the rows /	samples /		
variance within	variance		
the columns	samples		
F-test	F-test		
	varience		
t-test	ratio test		
Chi- square test	ANOVA		

many

one way

many

two way

KARPAGAM ACADEMY OF HIGHER EDUCATION DEPARTMENT OF MATHEMATICS

BIOSTATISTICS AND RESEARCH METHODOLOGY (I7MBP304) choice 4 choice 2

Questions choice 1

UNIT - IV

choice 3

Scales have an absolute
or true zero of measurement

	Ordinal	Nominal	interval	ratio
The section of				
constitutes the main body of the				
report where in the results of the				
study are presented in clear.	Appendix	results	methods	Ordinal
Study to portray accurately				
characteristics of a particular				
individual, situation or a group is				Hypothesis
called research	Exploratory	Diagnostic	Descriptive	testing
Critical evaluation made by the				
researcher with the facts and				
information already available is				Hypothesis
called research.	Analytical	Exploratory	Diagnostic	testing
Research to find reason, why				
people think or do certain things	Quantitative	Applied	Qualitative	Fundamental
is an example of	Research	Research	research	research
Which one is considered a major				
component of the research study	Interpretatio	research		
	n	report	finding	draft
Research task remains incomplete				
till the has been				objective and
presented.	Report	objective	finding	finding
What is the last step in a research	Writing	writing	writing	research
study	report	finding	limitations	report
Which is the final step in report	Writing	writing	writing	writing
writing	report	finding	drafts	limitations
What is usually appended to the				
research work				research
	Editing	bibliography	coding	report
The is one which				
gives emphasis on simplicity and	popular	research	article	writing
attractiveness	report	report	report	limitations
should slow				
originality and should necessarily				
be on attempt to solve some	Interpretatio	research		
		4	1. 1.	1 0

The researcher must remain				
caution about the that				
can possibly arise in the process				
of interpreting results	Analysis	conclusions	findings	error
Which one should be considered	-		both	
while interpreting a given data			validity and	technical
	Validity	reliability	reliability	jargon
is asking questions	Indirect	mailed	through	personal
face to face	method	questionnaire	post	interview
Journals, books, magazines etc		1	both	
are useful sources of collecting			primary and	
		secondary	secondary	
	Primary data	data	data	objective
The collected raw data to detect	-			-
errors and are called,				
			classificatio	
	Editing	coding	n	all the above
The formal, systematic and				
intensive process of carrying on a				
scientific method of	Research		interpretatio	research
analysis is	Design	research	n	analysis
Refers to the process of				
assigning numerals or symbols to			classificatio	
answers of response	Coding	editing	n	all the above
The research study, which is				
based on describing the				
characteristic of a particular	Experience			
individual or group	survey	Descriptive	Diagnostic	Exploratory
Research is a				
	Finding	assumption	statement	all the above
The research, which has the				
purpose of improving a product				
or a process testing theoretical				
concepts in actual problem				
situations isresearch.	Statistical	Applied	Domestic	Biological
The chart of research process	Closely		Closely	moderately
indicates that the process consists	related	unrelated	unrelated	related
of a number of	activities	activities	activities	activities

The objective of a good design is -	- Maximize			
	the bias		Minimize	
	andmaximiz	Minimize the	the bias and	Maximize
	e	bias and	maximize	the bias and
	the	minimize	the	maximize
	reliability of	the reliability	reliability of	the reliability
	data	of data	data	of data
A is used whenever a				
full written report of the study is	Popular	Technical		
required.	report	report	article	monograph
The is one which gives				
emphasis on simplicity and	Popular	Technical		
attractiveness.	report	report	article	monograph
Which of the following are				
measurements of scale?	Nominal	ordinal	interval	all the above
Scale is a system of				
assigning numbers, symbols to				
events in order to label them.	Interval	ordinal	Nominal	ratio
The qualitative phenomena are				
considered in the				
scale.	Ordinal	Nominal	interval	ratio
Scales can have an				
arbitrary zero, but it is not				
possible to determine the				
absolute zero.				
	Ordinal	Nominal	interval	ratio

Answer			
ratio			
results			
Descriptive			
Analytical			
Qualitative research			
research report			
Report Writing report writing drafts			
bibliograph y			
popular report			
research report			

error

reliability personal interview

secondary data

Editing

research

Coding

Descriptive all the above

Applied Closely related activities

Maximize the bias and maximize the reliability of data	
Technical report	
Popular report all the above	
Nominal	
Ordinal	

interval

KARPAGAM ACADEMY OF HIGHER EDUCATION DEPARTMENT OF MATHEMATICS

BIOSTATISTICS AND RESEARCH METHODOLOGY (I7MBP304)

Questions	choice 1	choice 2 UNIT - V	choice 3	choice 4
Psychometric Methods book is written by	J.P.Guilford	Likert	L.L.Thursto ne	Louis Guttman
Respondents are asked to rank	Comparative	arbitrary	rating scale	differential
their choices in	scaling	scaling		scale
is developed on ad-hoc basis	Differential scale	arbitrary scale	rating scale	ranking scale
scale is	Comparative	likert	differential	rating scale
developed by utilizing item analysis approach	scale	scale	scale	
Scalogram analysis is developed	J.P.Guilford	Likert		Louis
by			L.L.Thursto ne	Guttman
A complete enumeration of all	sampling	sample	census	all the above
items in the population is known	unit	design	inquiry	
as				
The selected respondents constitute	population	sample	sample size	population size
The selection process of	survey	sampling	sample	census
respondents is		technique	survey	inquiry
called				
The survey conducted to select		sample	census	population
the respondents is	sampling	survey	inquiry	size
	technique	1	1	1
A sample design is a definite plan	universe	sample	population	sample
given		design		survey
The number of items in universe	finite	infinite	both	zero
The population of a city number	infinite	finite	both	zero
of workers in a company	minite	mine	ootii	2010
is				
Source list is also known	sampling	sampling	sampling	population
as	size	size	frame	size
The size of the sample should be	large	optimum	small	all the above
Inappropriateness in sampling frame will result in	systematic bias	optimum	problems	sampling error

Sampling error with increase in size of sample	decrease	increase	both	optimum
Sampling error can be measured from	sample design	sample size	population	sample design and sample size
On the representation basis samples may	probability sampling	non- probability	both	restricted
On element selection basis the samples may be	restricted	unrestricte d	both	probability sampling
Non-probability sampling is also known as	quota sampling	purposive sampling	deliberate sampling	all the three
Quota sampling is an example of	probability sampling	non- probability sampling	both	purposive sampling
Probability sampling is also known as	random sampling	choice sampling	random and choice sampling	multistage sampling
Lottery method of selecting data is an example of	random sampling	choice sampling	purposive sampling	quota sampling
Systematic sampling is an improved version of	quota sampling	simple random sampling	choice sampling	purposive sampling
If population is not drawn from homogeneous group	simple random sampling	quota sampling	choice sampling	stratified sampling
In total population is divided into number of relatively small sub divisions	cluster sampling	choice sampling	stratified sampling	quota sampling
When a particular lot is to be accepted or rejected on the basis of single sampling it is known as	double sampling	single sampling	area sampling	purposive sampling

Survey designed to determine attitude of students toward new teaching plan is known as	cross stratification sampling	stratificatio n sampling	cluster sampling	multi stage sampling
 Sample design is determined	before	after	both	based on the survey
datas are collected Indeterminary principle step comes in	step in sample design	criteria to select sample procedure	both	step doesnot occur
The measurement of sampling error is called as	precision of sampling plan	sampling survey	sampling plan	representati on basis
The different sub populations divided to constitute a sample is known as	stratified sampling	survey	population	strata
Every nth item is selected in	stratified sampling	systematic sampling	judgement sampling	all the above
is conducted for determining a more appropriate and efficient stratification plan	survey	sample	pilot study	sample plan
is considered more appropriate when universe happens to be small	purposive sampling	area sampling	cluster sampling	simple random sampling
When we use rating scales we judge an object interms against some specified criteria.	real	absolute	imaginary	perfect
Rating scale is also known as The graphical scale is and is commonly used	Categorical scale Problematic	arbitrary scale critical	cumulative scales simple	all the above real
in practice. is also known as numerical scale	Itemized rating scale	graphical rating scale	cumulative scale	likert scale

The chief merit of itemized rating scale is it provides information	more	deep	critical	all the above
occurs when	error of	error of	error of	cumulative
the respondents are either easy raters or hard raters	hallo effect	leniency	central tendency	scales
occurs when	error of	error of	error of	graphical
the rater carries a generalized impression of the subject from one rating to another.	hallo effect	leniency	central tendency	rating scale
When the raters are reluctant to give extreme judgments, the result is	error of hallo effect	error of leniency	error of central tendency	cluster sampling
Systematic bias is also known	Error of	error of	error of	cumulative
as	hallo effect	leniency	central tendency	scales
occurs when the	error of	error of	error of	cluster
rater is asked to rate more factors, which has no evidence for	hallo effect	leniency	central tendency	sampling
judgment.			1.1 . 1	
as ranking scale	rating scale	comparativ e scale	likert scale	graphical rating scale
We make relative judgments against similar objects in	comparative scale	likert scale	differential scale	rating scale
Paired comparisions	nominal	ordinal	ratios	interval
Ordinal data can be converted to data through Law	nominal	ordinal	ratio	interval
of comparative judgment.				
Law of comparative judgment is developed by	J.P.Guilford	Likert	L.L.Thursto	all the three

Answer

J.P.Guilf ord Comparat ive scaling arbitrary scale likert scale Louis Guttman census inquiry sample sampling technique sample survey populatio n both finite sampling frame optimum systemati c bias

decrease

sample design and sample size both both all the three nonprobabilit у sampling random and choice sampling random sampling

simple random sampling stratified sampling

cluster sampling

single sampling
cross stratificat ion sampling

before

criteria to select sample procedure precision of sampling plan strata

systemati c sampling pilot study

purposive sampling

absolute

Categoric al scale simple

Itemized rating scale

error of leniency error of hallo effect error of central tendency Error of hallo effect error of hallo effect comparat ive scale comparat ive scale ordinal interval

more

L.L.Thur stone