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

Semester VI

16BCU602A

BIOSTATISTICS

Scope: On successful completion of this course the learner gains a clear knowledge about various aspects of Statistics, measures, hypothesis testing and application of them in their respective fields.

Objectives: To enable the students to understand the meaning , definition and functions of statistics through collection , representation , finding various measures such as mean, median , mode , correlation etc of statistics.

Unit 1

Definitions-Scope of Biostatistics- Variables in biology, collection, classification and tabulation of data- Graphical and diagrammatic representation.

Unit 2

Measures of central tendency – Arithmetic mean, median and mode. Measures of dispersion-Range, standard deviation, Coefficient of variation.

Unit 3

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

Unit 4

Regression: Regression in two variables – Regression coefficient problems – uses of regression.

Unit 5

Test of significance: Tests based on Means only-Both Large sample and Small sample tests - Chi square test - goodness of fit.

TEXT BOOK

Pillai R.S.N., and Bagavathi V., 2002.,Statistics , S. Chand & Company Ltd, New Delhi.

REFERENCES

Jerrold H.Z., (2003). Biostatistical Analysis, Fourth Edition, Pearson Education (Pte) .Ltd, New Delhi.

Arora, P.N., (1997). A foundation course statistics, S.Chand & Company Ltd, New Delhi.

Navnitham, P.A., (2004). Business Mathematics And Statistics, Jai Publications, Trichy,

Gupta S.P., (2001). Statistical methods, Sultan Chand & Sons, New Delhi

KARPAGAM ACADEMY OF HIGHER EDUCATION

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

Coimbatore – 641 021.

LECTURE PLAN
DEPARTMENT OF BIOCHEMISTRY

STAFF NAME: Dr.K.Poornima
SUBJECT NAME: Biostatistics SUB.CODE:16BCU602A
SEMESTER:VI CLASS: III B.Sc (BC)

Sl. No	Duration of Period	Topics to be Covered	Support material
		UNIT I	
1	1	Definitions-Scope of Biostatistics Variables in biology	T1:4-5 R3:1.13 R1:1-3 R2:1-2
2	1	Collection of primary data	T1:31-35 R3:3.4-3.14 R1:5-6 R2:27-31
3	1	Collection of Secondary data	T1:36-40 R3:3.2-3.4 R1:6-7 R2:31-33
4		Classification of data	T1:56-73 R3:5.2-5.9 R1:6-7 R2:44-51
5	1	Tabulation of data	T1:73-81 R3:5.18-5.35 R1:8-20 R2:52-57
6	1	Graphical representation of data	T1:101-118 R3:6.27-6.53 R2:60-78
7	1	Diagrammatic representation of data	T1:84-100 R3:6.2-6.27 R2:79-94
8	1	Revision and possible questions discussion of Unit I	
		Total no of hours planned for UNIT I = 07	

Unit II			
1	1	Measures of central tendency – Arithmetic mean problems	T1:121-146 R3:7.5-7.19 R1:21-28 R2:101-113
2	1	Median problems	T1:146-170 R3:7.19-7.32 R1:32-40 R2:113-117
3	1	Mode problem	T1:117-275 R3:7.33-7.43 R1:44-47 R2:117 -127
4	1	Measures of dispersion-Range	T1:234-239 R3:8.8 R1:64-65
5	1	Standard deviation	T1:249-304 R3:8.26-8.27 R1:75-85 R2:117-179
6	1	Practicing problems in standard deivation	T1:305-317 R3:8.27-8.32 R1:98-101
7	1	Coefficient of variation- Practicing problems	T1:270-275 R3:8.3-8.37 R1:85-88 R2:172-177
8	1	Revision and possible questions discussion of Unit II	
		Total no of hours planned for UNIT II = 08	
Unit III			
1	1	Correlation – Meaning and definition	T1:359-363 R3:10.2-10.3 R1:103-104
2	1	Scatter diagram	T1:363-365 R3:10.7-10.10
3	1	Karl Pearson’s correlation coefficient	T1:365-385 R3:10.25-10.28 R1:112-122 R2:260-263
4	1	Continuation of Karl Pearson’s correlation coefficient	T1:412-424 R3:10.44-10.61
5	1	Practicing problems in Karl Pearson’s correlation coefficient	T1:385-391 R3:10.37.10.43
6	1	Rank correlation	T1:412-424

			R3:10.44-10.61 R1:125-130 R2:269-271
7	1	Working out of problems in Rank correlation	R3:10.44-10.61 R1:125-130 R2:269-271
8	1	Revision and possible questions discussion of Unit III	
		Total no of hours planned for UNIT III = 08	
		Unit IV	
1	1	Regression - Introduction	T1:425-428 R3:11.2-11.4
2	1	Regression in two variables	T1:425-428 R3:11.2-11.4 R2:284-285
3	1	Regression equation of x on y	T1:431-445 R3:11.7 R1:149-170
4	1	Regression equation of y on x	T1:431-445 R3:11.9
5	1	Regression coefficient problems	T1:445-476 R3:11.10-11.46 R2:287-290
6	1	Uses of regression	T1:476 R3:11.47 R1:180
7	1	Revision and possible questions discussion of Unit IV	
		Total no of hours planned for UNIT IV = 07	
		UNIT V	
1	1	Test of significance – An Introduction	T1:765-770 R3:3.14
2	1	Tests based on Means only – Large sample	T1:779-785 R3:3.14-3.21 R1:318-324
3	1	Small samples – students t- test	T1:785-787 R3:3.30-3.32 R1:324-327
4	1	Working out of problems in test of significance	T1:787-789 R3:3.31-3.47
5	1	Chi square test - goodness of fit.	T1:790-792 R3:4.2-4.7 R1:392-400
6	1	Practicing problems in chi square test	T1:792-805 R3:4.20-4.48
7	1	Revision and possible question discussion of unit V	
		Total no of hours planned for UNIT V = 07	

Total number of hours planned to complete this course – 40

REFERENCE

TEXT BOOK

T1: Pillai R.S.N., and Bagavathi V., 2002.,Statistics , S. Chand & Company Ltd, New Delhi.

REFERENCES

R1: Arora, P.N., (1997). A foundation course statistics, S.Chand & Company Ltd, New Delhi.

R2: Navnitham, P.A., (2004). Business Mathematics And Statistics, Jai Publications, Trichy,

R3:Gupta S.P., (2001). Statistical methods, Sultan Chand & Sons, New Delhi

UNIT-I
SYLLABUS

Definitions-Scope of Biostatistics- Variables in biology, collection, classification and tabulation of data- Graphical and diagrammatic representation.

Introduction

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

Origin and Growth

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

Meaning

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

Definitions

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

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

Functions

The following are the important functions of statistics.

* Collection
* Numerical Presentation
* Diagrammatic Presentation
* Condensation
* Comparison
* Forecasting
* Policy Making
* Effect Measuring
* Estimation

* Tests of significance.

Characteristics

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

Collection of data

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

Primary data

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

Secondary Data

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

Methods of collection of primary Data

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

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

Sources of secondary data

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

Classification

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

Basis

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

i) Geographical Classification or Spatial Classification

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

ii)Chronological or Temporal or Historical Classification

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

iii) Qualitative Classification

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

iv)Quantitative Classification

Some data can be classified in terms of magnitudes.

Tabulation

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

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

i) Statistical Tables

ii) Diagrams or Graphs

Parts of a table

A good table has the following parts or components:

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

Frequency Distribution

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

The following are the *steps* of constructing a frequency distribution:
1. Specify the number of class intervals. A class is a group (category) of interest. No totally

accepted rule tells us how many intervals are to be used. Between 5 and 15 class intervals are generally recommended. Note that the classes must be both *mutually exclusive and all-inclusive.* Mutually exclusive means that classes must be selected such that an item can't fall into two classes, and all-inclusive classes are classes that together contain all the data.

2. When all intervals are to be the same width, the following rule may be used to find the required class interval width:

W = (L - S) / K

where:

W= class width, **L=** the largest data, **S=** the smallest data, **K=** number of classes

Example

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

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

Cumulative Frequency Distribution

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

Diagrams and Graphs

Diagrams

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

Rules for construction

* Title

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

Types of Diagram

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

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

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

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

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

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

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

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

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

POSSIBLE QUESTIONS
UNIT I

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

PART B (5 x 2= 20Marks)

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

PART C (5 X 6 = 30 Marks)

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

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

vi)	Central plan	15
	Total	100

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

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

Draw a Pie Diagram for the above data.

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

Brand	**July 2005 Sales (in Rs. '000)**
Maruti	81
Hyundai Santro	39
Honda city	14
Matiz	20
Opel Astra	10
Fait Uno	12
Ford	37
Mitsubishi Lancer	11
Mercedes	3

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics 16BCU602-A
UNIT I

1	1	The most suitable form of presentation for publicity and Propaganda is	**diagram**	numerals	graph	map	**diagram**
2	1	Mailed enquiry method can be adopted if the respondents are	illiterate	**literate**	blind	physically challenged	**literate**
3	1	Pie-chart represents the components of a factor by	percentages	angles	**sectors**	deciles	**sectors**
4	1	The number of questions of a questionnaire should be	5	20	25	**as small as possible keeping in view the purpose of the survey**	**as small as possible keeping in view the purpose of the survey**
5	1	Primary data are	**always more reliable compared to secondary data**	less reliable compared to secondary data	depends on the care with which data have been selected	depends on the agency collecting the data	**always more reliable compared to secondary data**
6	1	The census data published for state wise population in India will be known as	Quantitative classification	Two-way classification	**Geographical classification**	Quanlitative classification	**Geographical classification**
7	1	Statistics implies	both data and science	data only	**data, science and measure in samples**	samples	**data, science and measure in samples**
8	1	Data taken from the publication, 'Agricultural Situation in India' will be considered as	primary data	**secondary data**	primary and secondary data	published data	**secondary data**
9	1	Year wise recording of data of food production will be called as:	Geographical classification	**Chronological classification**	Quantitative classification	Quanlitative classification	**Chronological classification**
10	1	Who is the father of Biostatistics?	R.A.Fisher	W.Gosset	**Sir Francis Galton**	S.C.Gupta	**Sir Francis Galton**
11	1	Number of source of data is	**2**	3	4	1	**2**
12	1	Compared with primary data secondary data are	more reliable	**less reliable**	equally reliable	not reliable	**less reliable**
13	1	In quantitative classification data are classified on the basis of	attributes	time	location	**magnitudes**	**magnitudes**
14	1	Classification according to class-intervals would yield	raw data	discrete data	qualitative data	**grouped data**	**grouped data**
15	1	In qualitative classification data are classified on the basis of	**attributes**	time	location	class intervals	**attributes**
16	1	In geographical classification data are classified on the basis of	**area**	attributes	time	location	**area**
17	1	Which source is one that itself collects the data?	**primary data**	secondary data	published data	non published data	**primary data**
18	1	Individual observations are called	**raw data**	grouped data	ungrouped data	simple data	**raw data**
19	1	Which one is Geographical classification?	1990-91	**North**	Male	442	**North**
20	1	In discrete frequency distribution values are given as	class intervals	grouped data	**ungrouped data**	equal frequency data	**ungrouped data**
21	1	In continuous frequency distribution values are given as	**class intervals**	grouped data	ungrouped data	equal frequency data	**class intervals**
22	1	Which of the following is the one dimensional diagram?	**square diagram**	multiple bar diagram	rectangular diagram	Pie-Chart	**square diagram**
23	1	In a bar diagram, the base line is	Horizontal	**Vertical**	False baseline	true baseline	**Vertical**
24	1	Pictograms are shown by	Dots	Lines	Circles	**Pictures**	**Pictures**
25	1	Histogram is suitable for the data presented as	**continuous grouped frequency distribution**	discrete grouped frequency distribution	individual series	discrete series	**continuous grouped frequency distribution**
26	1	Numerical data presented in descriptive form are called	classified presentation	tabular presentation	graphical presentation	**textual presentation**	**textual presentation**
27	1	A simple table represents	**only one factor or variable**	always two factors or variables	two or more number of factors or variables	only four factors or variables	**only one factor or variable**
28	1	The row headings of a table are known as	sub-titles	**stubs**	reference notes	captions	**stubs**
29	1	The origin of the word statistics has been traced from the Latin word	statista	**status**	statistic	statistique	**status**
30	1	Relative error is always	Positive	Negative	**Positive or Negative**	Zero	**Positive or Negative**
31	1	The column headings of a table are known as	sub-titles	stubs	reference notes	**captions**	**captions**
32	1	Statistical data are collected for,	Collecting data without any purpose	**a given purpose**	any purpose	for all purpose	**a given purpose**
33	1	In grouped data, the number of classes preferred are	Minimum possible	**adequate**	Maximum possible	any arbitrarily chosen number	**adequate**
34	1	Class interval is measured as	The sum of the upper and lower limit	Half of the sum of the upper and lower limit	**The difference between the upper and lower limit**	Upper limit + lower limit	**The difference between the upper and lower limit**
35	1	The shape of the trilinear charts is that of a	Cone	Cube	**Equilateral triangle**	Pyramid	**Equilateral triangle**

UNIT-II
SYLLABUS

Measures of central tendency – Arithmetic mean, median and mode.Measures of dispersion-Range, standard deviation, Coefficient of variation.

INTRODUCTION

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

Measures of Central Tendency

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

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

- Arithmetic Mean or simply Mean
- Median
- Mode

Arithmetic Mean or simply Mean

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

Individual series

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

Example:

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

30 70 10 75 500 8 42 250 40 36

Solution

Here N=10

$\Sigma X = 30 + 70 + 10 + 75 + 500 + 8 + 42 + 250 + 40 + 36 = 1061$

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

Discrete series

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

Example

Calculate the mean number of person per house.

No. of person : 2 3 4 5 6

No. of house :10 25 30 25 10

Solution

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

$\Sigma f = 100$ $\Sigma f X = 400$

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

Continuous series

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

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

Example

Calculate the mean for the following.

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

Solution:

C.I	**f**	**m**	**f m**
20-30	5	25	125
30-40	8	35	280
40-50	12	45	540
50-60	15	55	825

60-70	6	65	390
70-80	4	75	300

$\Sigma f = 50$ $\Sigma f m = 2460$

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

Median

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

Individual series

For odd number of item

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

For even number of item

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

Example

Calculate median for the following.

22 10 6 7 12 8 5

Solution

Here N =7

Arrange in ascending order or descending order.

5 6 7 8 10 12 22

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

= 4 th item = 8

Discrete series

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

Example

Find the median for the following.

X : 10 15 17 18 21

F: 4 16 12 5 3

Solution

X	f	c.f
10	4	4

15	16	20
17	12	32
18	5	37
21	3	40

N= 40

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

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

= 16.

Continuous series

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

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

Example

Calculate the median height given below.

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

Solution :

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

$\Sigma f = 30$

Position of the median = N/2 th item = 30 / 2 =15.

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

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

Mode :

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

Individual series

The value which occur more times are identified as mode.

Example

Determine the mode

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

Solution:

Unimode = 32.

Discrete series

Determine the mode

Size of dress	**No. of set**
18	55
$\boxed{20 \quad 120}$	
22	108
24	45

here mode represents highest frequency .

Mode = 20

Continuous series

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

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

Example

Determine the mode

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

Solution

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

$Z = L + [\, i(f_1 - f_0) / (2f_1 - f_0 - f_2)]$
$= 20+[10(35-20)/(2(35)-20-20)] = 20+5$
$= 25.$

Empirical relation

- Mode = 3 median -2 mean.

Measures of Dispersion

Measure of dispersion deals mainly with the following three measures

- Range
- Standard deviation
- Coefficient of variation

Range

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

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

Individual series

Example

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

8 ,10, 5, 9,12,11

Solution

Range = L – S
= 12- 5 =7

Coefficient of range = (L-S) /(L+S)
= (12-5) / (12+5)
= 7 /17 = 0.4118

Continuous series

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

Calculate the range.

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

Solution

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

Here L=75 & S=25

Range = L – S = 75-25 = 50

Standard deviation

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

Individual series

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

Example

Calculate standard deviation for the following data.

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

Solution

X	X^2
40	1600
41	1681
45	2025
49	2401
50	2500
51	2601
55	3025
59	3481

60 3600
603600
510 $\Sigma x^2 = 26504$

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

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

$$= 7.09$$

Discrete series

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

Example

Calculate standard deviation for the following data.

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

Solution

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

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

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

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

$$= 1.44$$

Continuous series

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

Example

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

Solution

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

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

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

$$= 9.79$$

Coefficient of variation

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

Example

Calculate the coefficient of variation.

Mean= 51, standard deviation = 7.09

Solution

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

= (7.09 /51) x100

= 13.9

POSSIBLE QUESTIONS
UNIT I

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

PART B (5 x 2= 20Marks)

15. Calculate the Mean for the following.

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

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

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

18. What is mean by Bimodal?
19. Calculate the Median for the following data.
80 100 50 90 120 110
20. Write the relation between Standard Deviation and Variance.
21. Calculate the Average number of students per class for the following data.
26 46 33 25 36 27 34 29
22. Find Median and Mode for the following data.
13 16 17 15 18 14 19 15 12
23. Define Range.
24. Find the Arithmetic Mean for the following data.
70 60 75 50 42 95 46
25. Calculate the Range and its Coefficient for the following data.
17 10 56 19 12 11 18 14
26. Find the median for 57, 58, 61, 42, 38, 65, 72, and 66
27. Write the empirical relation for Mode.

PART C (5 X 6 = 30 Marks)

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

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

2. Draw Percentage Bar Diagram for the following data.

Food	Rs.200
Clothing	Rs.60

Education Rs.70
Rent Rs.130
Miscellaneous Rs.40

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

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

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

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

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

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

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

Draw a Frequency Polygon to the above Frequency Distribution.

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

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

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

10	14	20	12	21	16	19	17	14	25

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

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

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

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

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

X	6	9	12	15	18
f	7	12	13	10	8

15. Calculate the Median for the following.

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

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

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

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

17. Calculate the Arithmetic Mean for the following data.

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

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

77 73 75 70 72 76 75 72 74 76

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics 16BCU602-A
UNIT II

1	2	Mean is a measure of	**central value**	dispersion	correlation	regression	**central value**
2	2	Mode is that value in a frequency distribution which possesses	minimum frequency	frequency one	**maximum frequency**	last frequency	**maximum frequency**
3	2	The most stable measure of central tendency is	**the mean**	the median	the mode	range	**the mean**
4	2	Sum of the deviations about mean is:	minimum	**zero**	maximum	one	**zero**
5	2	The formula used to calculate arithmetic mean for individual series by direct method is	**∑X/N**	∑Fx/N	∑Fm/N	N/ ∑Fx	**∑X/N**
6	2	The formula used to calculate arithmetic mean for individual series by short cut method is	**A+ ∑d/N**	A+ ∑Fd/N	A+∑Fm/N	A+ N/ ∑d	**A+ ∑d/N**
7	2	In continuous series the formula for A.M is	X/N	N/∑X	∑X/N	**∑fm/N**	**∑fm/N**
8	2	A.M = 8. N = 12 then ∑X =	76	80	86	**96**	**96**
9	2	12. 34. 56. 34. 45. 11 in this series the mode is	12	56	**34**	11	**34**
10	2	The data given as 5. 12. 16. 24. 35. 44 will be called as	a continuous series	a discrete series	**an individual series**	time series	**an individual series**
11	2	Which of the following divides the series into two equal parts	Mean	Mode	**Median**	Range	**Median**
12	2	Mode of the following data 3. 6. 5. 7. 8. 4. 9	3	7	**no mode**	5	**no mode**
13	2	Which of the following is not a measure of dispersion?	Range	Quartile deviation	Standard deviation	**Median**	**Median**
14	2	Range of the given values is given by	**L-S**	L+S	S-L	LS	**L-S**
15	2	Which one of the following is relative measure of dispersion?	range	O.D	S.D	**Coefficient of variation**	**Coefficient of variation**
16	2	Range of a set of values is 65 and maximum value in the series is 83. The minimum value of the series is	74	9	**18**	65	**18**
17	2	If standard deviation is 5. then the variance is	5	**25**	625	2.23068	**25**
18	2	If the value of mode and mean is 60 and 66 then the value of median is	**64**	46	54	44	**64**
19	2	Standard deviation is also called	**Root mean square deviation**	Mean square deviation	Root deviation	Root median square deviation	**Root mean square deviation**
20	2	X: 10-15 15-20 20-25 25-30 30-35 35-40	8	**30**	33	12	**30**
21	2	f: 12 4 10 6 11 8 . range is					
22	2	If the S.D and the C.V of a series are 5 and 25. then the mean	**500**	200	250	100	**500**
23	2	Which one of the following is a measure of central tendency	**median**	range	variation	correlation	**median**
24	2	Mean of the following values is 5 15 20 10 20	5	**14**	41	20	**14**
25	2	The position of the median for an individual series is taken as	**(n+1)/2**	(n+2)/2	n/2	n/4	**(n+1)/2**

UNIT-III
SYLLABUS

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

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

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

Example :

X	50	60	70	95	100	105
Y	23	32	37	41	46	50

Example :

X	34	25	18	10	7
Y	51	49	42	33	19

Simple or Partial or Multiple

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

Methods

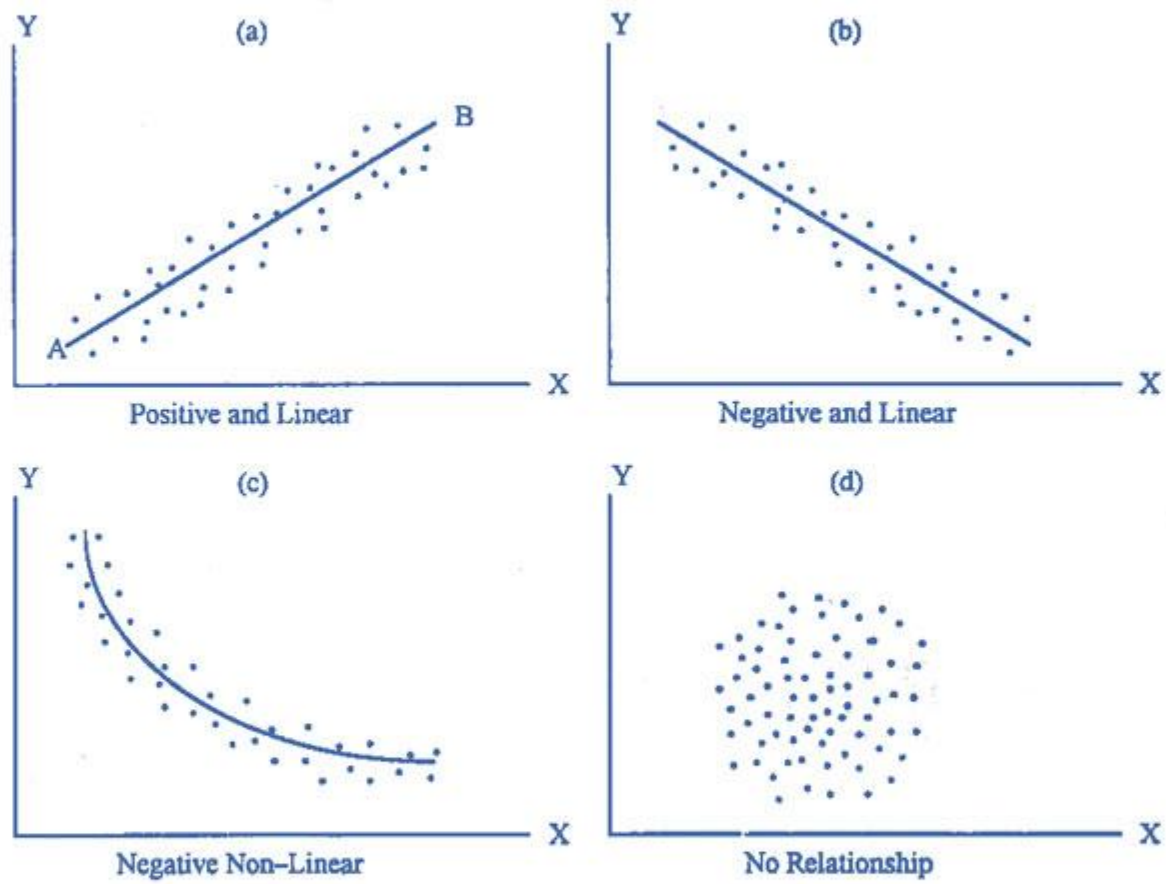
The following four methods are available under simple linear correlation and among them; product moment method is the best one.

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

- Correlation coefficient by concurrent deviation method (r_c).

Scatter Diagram

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



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

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

Karl Pearson's Coefficient

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

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

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

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

Pearson's Coefficient computational formula

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

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

Subject	Age x	Weight Level y
1	43	99
2	21	65

3	25	79
4	42	75
5	57	87
6	59	81

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

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

.

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

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

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

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

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

Subject	Age x	Weight Level y	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569

6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Step 6: Use the following formula to work out the correlation coefficient. The answer is: 1.3787 × 10-4 the range of the correlation coefficient is from -1 to 1. Since our result is 1.3787 × 10-4, a tiny positive amount, we can't draw any conclusions one way or another.

Spearman's Rank Correlation Coefficient

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

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

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

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 : 21 36 42 37 25

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

Solution

		RANK			
X	Y	X	Y	d	D^2
21	47	5	1	4	16
36	40	3	4	-1	1
42	37	1	5	-4	16
37	42	2	3	-1	1
25	43	4	2	2	4

Total $\sum d = 0$ $\sum d^2 = 38$

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

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

$= 1 - 1.9 = -0.9$

Tied Ranks

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

Example

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

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

Solution

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

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

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

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

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

$= 1- 1.3571 = -0.3571$

POSSIBLE QUESTIONS
UNIT III

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

PART B (5 x 2= 20Marks)

1) What are the types of Correlation?
2) Write any two properties of Correlation.
3) What is the range of Correlation Coefficient?
4) Define Positive Correlation.
5) What is meant by Regression?
6) What are the formulae for Regression co-efficients?
7) Distinguish between Correlation and Regression.
8) Write the formula for Rank Correlation, when more than one rank is repeated.
9) If b_{xy} = -0.2337 and b_{yx}= - 0.6643 then find the Correlation Coefficient.
10) What is Negative Correlation? Give an example?
11) Write down the formula for Karl Pearson's Coefficient of Correlation.
12) Define Scatter Diagram.
13) What is Simple Correlation?

PART C (5 X 6 = 30 Marks)

1) Calculate the Correlation Coefficient from the following variables.

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

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

X	25	20	28	22	40	60	20

Y	40	30	50	30	20	10	30

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

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

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

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

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

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

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

Karpagam Academy of Higher Education
Department of Biochemistry
Biostatistics 16BCU602-A
UNIT III

1	3	Scatter diagram method is a	**Graphic**	Mathematical	Numerical	Algebraical	**Graphic**
2	3	If every change in X produces a corresponding decrease in Y then X and Y are said to be	uncorrelated	independent	positively correlated	**negatively correlated**	**negatively correlated**
3	3	Rank correlation was discovered by	R.A.Fisher	Sir Francis Galton	Karl Pearson	**Spearman**	**Spearman**
4	3	Correlation is used to measure	**closeness of relationship between variables**	one variable from another	nature of the distribution	central value	**closeness of relationship between variables**
5	3	Coefficient of correlation lies between	**1 and –1**	0 and 1	0 and ∞	0 and –1	**1 and –1**
6	3	While drawing a scatter diagram if all points appear to form a straight linegetting downward from left to right, then it is inferred that there is	a perfect positive correlation	simple positive correlation	**a perfect negative correlation**	no correlation	**a perfect negative correlation**
7	3	The range of the rank correlation coefficient is	0 to 1	**-1 to 1**	0 to ∞	-∞ to +∞	**-1 to 1**
8	3	The technique used in measuring the closeness of the relationship between theVariables is referred to as	Range	standard deviation	median	**correlation analysis**	**correlation analysis**
9	3	Correlation is the	Relationship between two values	**Relationship between two variables**	Significance between two variables	Significant level between two values	**Relationship between two variables**
10	3	If r =+1, the given two variables are	**perfectly positive**	perfectly negative	no correlation	Negative	**perfectly positive**
11	3	If r =-1, the given two variables are	perfectly positive	**perfectly negative**	positive	negative	**perfectly negative**
12	3	If r =0, the given two variables are having	perfect positive correlation	perfect negative correlation	**no correlation**	slight correlation	**no correlation**
13	3	Coefficient of correlation lies between	**1 and –1**	0 and 1	0 and ∞	0 and –1	**1 and –1**
14	3	The range of the rank correlation coefficient is	0 to 1	**-1 to 1**	0 to ∞	-∞ to +∞	**-1 to 1**
15	3	Formula for rank correlation is					
	3	The formula for computing Pearsonian r is	**$\Sigma xy / N\sigma_x \sigma_y$**	$\Sigma x / N\sigma_x \sigma_y$	$\Sigma y / N\sigma_x \sigma_y$	$\Sigma xy / \sigma_x \sigma_y$	**$\Sigma xy / N\sigma_x \sigma_y$**
16	3	In rank correlation the sum of the differences of ranks between two variables shall be	**zero**	more than 1	less than 1	10	**zero**
17	3	Estimation of the value of one variable from the given value of another variable is done by	correlation analysis	**regression analysis**	chi square test	student's t test	**regression analysis**
18	3	If advertising and sales are correlated the expected amount of sales for a give Advertising expenditure is calculated by	correlation analysis	**regression analysis**	chi square test	student's t test	**regression analysis**
19	3	If advertising and sales are correlated the required amount of expenditure for attaininga given amount of sales is calculated by	correlation analysis	**regression analysis**	chi square test	student's t test	**regression analysis**
20	3	If three ranks are equal at 5^{th} place, the rank given to them is	5	**6**	7	5.5	**6**
	3	When equal ranks are assigned the adjustments made by adding	**1/12(m3-m)**	1/6(m2-m)	1/5(m2+m)	1/12(m2-m)	**1/12(m3-m)**
21	3	The only method used with ranks not the actual value is	Karl pearson's coefficient of correlation	**rank correlation**	scatter diagram method	concurrent deviation method	**rank correlation**
22	3	If the data are of a qualitative nature like honesty,efficiency and intelligence etc the method used is	Karl pearson's coefficient of correlation	**rank correlation**	scatter diagram method	concurrent deviation method	**rank correlation**
23	3	If the dots in scatter diagram is too scattered we can say that	r = +1	r = -1	**r = 0**	r = 5	**r = 0**

UNIT V

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

SYLLABUS

UNIT IV

Regression: Regression in two variables – Regression coefficient problems – uses of regression.

Simple Linear Regression

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

Uses

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

Method of Least Squares

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

- the objective is to create a BEST FIT line to the data concerned
- the criterion is the called the method of least squares
- i.e. the <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.

○

Calculating a and b:

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

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

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

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

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

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

Methods of forming the regression equations

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

Problem

From the following data, obtain the two regression equations.

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

Solution

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

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

The normal equations are ,

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

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

By substituting the values from the table, we get
5a+30b = 40 -------1
30a + 180b = 214 --------2
Solving these two equations we get,
a=11.90 and b= -0.65
Therefore the regression Y on X is Y = 11.90-0.65X.

Let the regression equation X on Y is X = a' + b'Y
The normal equations are,
$\sum x = Na + b\sum y$
$\sum xy = a\sum y + b\sum y^2$

By substituting the values from the table, we get
5a'+40'b = 30 -------3
40a' + 340b' = 214 --------4
Solving these two equations we get,
a' = 16.40 and b= -1.30
Therefore the regression equation X on Y is X = 16.40-1.30Y

Example From the data given below, find

(i) the two regression equations
(ii) The correlation coefficient between the variables X and y
(iii) The value of Y when X= 30

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

Solution

X	Y	x= X- X`	Y= Y-Y`	xy	x^2	y^2
25	43	-7	5	-35	49	25
28	46	-4	8	-32	16	64
35	49	3	11	33	9	121
32	41	0	3	0	0	9
31	36	-1	-2	2	1	4

36	32	4	-6	-24	16	36
29	31	-3	-7	21	9	49
38	30	6	-8	-48	36	64
34	33	2	-5	-10	4	25
32	39	0	1	0	0	1
320	380	0	0	-93	140	398

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

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

(Y – 38) = -0.6643(X-32) $\Rightarrow$ 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 = 59.26-0.6643x30= 39

Properties of Regression coefficients

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

POSSIBLE QUESTIONS
UNIT IV

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

PART B (5 x 2= 20Marks)

1) What is meant by Regression?
2) What are the formulae for Regression co-efficients?
3) Distinguish between Correlation and Regression.
4) What is Simple Correlation?
5) Define Regression Equation.
6) When X = 40, Y = 60, σ_x= 10, σ_Y = 15 and r = 0.7 find the Regression Equation of Y on X.

PART C (5 X 6 = 30 Marks)

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

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

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

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

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

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

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

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

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

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

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

Annual Income (Rs. in 000's)	62	78	41	53	85	34
Amount of Life Insurance (Rs. in 00's)	25	30	10	15	50	7

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

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

Calculate Spearman's Rank Correlation Coefficient.

7) You are given the following data:

	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8

Correlation coefficient between X and Y = 0.66

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

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

UNIT-V
SYLLABUS

Test of significance: Tests based on Means only-Both Large sample and Small sample tests – Student's t test, F-test, Chi square test - goodness of fit

Hypothesis:

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

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

Hypothesis testing:

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

Example:

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

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

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

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

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

Alternative hypothesis:

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

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

The null hypothesis would be

$$H_0 : \mu = 100$$

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

$$H_1 : \mu \neq 100$$

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

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

Errors in testing of hypothesis:

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

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

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

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

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

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

Level of significance:

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

α = P (Committing Type-I error)
= P (H_0 is rejected when it is true)

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

Power of the test:

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

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

$= 1 - \beta$

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

Critical region:

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

One tailed and two tailed tests:

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

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

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

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

Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^N c_n$ possible samples. If we calculate some particular statistic from each of the ${}^N c_n$ samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

Standard error:

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

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

Utility of standard error:

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

Test statistic:

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

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

Procedure for testing of hypothesis:

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

Large sample tests:

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

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

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

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

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

For this test

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

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

where μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots, x_n$ be a random sample from a normal population with mean μ and variance σ^2

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

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

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

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

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

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

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

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

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 respectively

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

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

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

For this test

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

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

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

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

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

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2}}} \sim N(0,1)$$ [since $\mu_1 - \mu_2$ =0 from H_0]

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

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

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

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

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

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

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

For this test

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

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

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

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

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

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

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

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

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively

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

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

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

For this test

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

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

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

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

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

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

Now calculate $|Z|$

Find out the tabulated value of Z at α% l.o.s i.e. Z_α

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

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

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

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

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

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

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

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

For this test

The null hypothesis is $H_0 : p = p_0$

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

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

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

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

Now calculate $|Z|$

Find out the tabulated value of Z at α% l.o.s i.e. Z_α

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

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

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

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

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

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}} \text{ and } S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

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

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

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

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

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

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

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

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

Now calculate $|Z|$

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

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

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

- **As σ is unknown,**

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

Step 2: If μ_0 falls into the above confidence intervals, then do *not* reject H_0. Otherwise, reject H_0.

Example 1:

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

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

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

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

[solutions:]

(a)

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

Then,

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

Therefore, reject H_0.

(b)

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

Therefore, ***not*** reject H_0.

(c)

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

A 95% confidence interval is

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

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

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let

$\alpha = 0.05$. Please test the hypothesis

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

based on
(a) classical hypothesis test
(b) p-value
(c) confidence interval.
[solution:]

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

(a)

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

we reject H_0.
(b)

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

we reject H_0.
(c)
$100 \times (1 - \alpha)\% = 95\%$ confidence interval is

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

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

Hypothesis Testing for the Mean (Small Samples)

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

Using the t-Test for a Mean μ when the sample is small		
Procedure	**Equations**	**Example 4**
State the claim mathematically and verbally. Identify the null and alternative hypotheses	State H_0 and H_a	$H_0 : \mu \geq 16500$ $H_a : \mu < 16500$ $n = 14, \bar{x} = 15700, s = 1250$
Specify the level of	Specify α	$\alpha = 0.05$

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

Chi-Square Tests and the F-Distribution

Goodness of Fit

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

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

H_0: The distribution fits the proposed proportions

H_1: The distribution differs from the claimed distribution.

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

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

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

$$E_i = np_i$$

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

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

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

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

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be ≥ 5.

Performing the Chi-Square Goodness-of-Fit Test (p 496)		
Procedure	Equations	Example **(p 497)**
Identify the claim. State the null and alternative hypothesis.	State H_0 and H_1	H_0: Classical 4% Country 36% Gospel 11% Oldies 2% Pop 18% Rock 29%
Specify the significance level	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	d.f. = #categories - 1	$d.f. = 6-1 = 5$
Find the critical value	χ^2_α: Obtain from Table 6 Appendix B	$\varphi^2_{0.01}(d.f = 5) = 15.086$

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

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

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

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

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

Music Type	Observed	Distribution	Expected
Classical	8	0.04	20
Country	210	0.36	180
Gospel	72	0.11	55
Oldies	10	0.02	10

Pop	75	0.18	90
Rock	125	0.29	145

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: Click **Calc->Calculator**. **Store the results in** C5 and calculate the **Expression** (C2-C4)**2/C4. Click on **OK** and C5 should contain the calculated values.

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

Next add up the values in C5 and the sum is the test statistic as follows: Click on **Calc->Column Statistics**. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

Sum of C5

```
Sum of C5 = 22.7132
```

Next calculate the P-value: Click on **Calc->Probability Distributions->Chi-square**. Select **Cumulative Probability** and enter 5 **Degrees of Freedom** Enter the value of the test statistic 22.7132 for the **Input Constant**. Click **OK**.

The following is displayed on the Session Window.

Cumulative Distribution Function

```
Chi-Square with 5 DF

      x  P( X <= x )
22.7132     0.999617
```

$P(X \leq 22.7132) = 0.999617$ So the P-value = 1 – 0.999617 = 0.000383. This is less that α = 0.01 so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

Chi-Square with M&M's

H_0: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: α = 0.05
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^2{}_{0.05}(d.f.=5)=11.071$
Rejection Region: $\chi^2 \geq 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where O is the actual number of M&M's of each color in the bag and E is the proportions specified under H_0 times the total number.
Reject H_0 if the test statistic is greater than the critical value (1.145)

Section 10.2 Independence

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c* contingency table** shows the observed frequencies for the two variables. The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where f_{ij} is the frequency that A equals A_i and B equals B_j.

	A₁	**A₂**	**A₃**	**A₄**	**A**
B₁	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B₂	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B₃	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
B	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum\,of\ row\ i) * (sum\,of\ column\ j)}{sample\ size}$$ (

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	$\frac{300*47}{550} \approx 25.64$	$\frac{300*87}{550} \approx 47.45$	$\frac{300*193}{550} \approx 105.27$	$\frac{300*180}{550} \approx 98.18$	$\frac{300*43}{550} \approx 23.45$	300
Large	$\frac{250*47}{550} \approx 21.36$	$\frac{250*87}{550} \approx 39.55$	$\frac{250*193}{550} \approx 87.73$	$\frac{250*180}{550} \approx 81.82$	$\frac{250*43}{550} \approx 19.55$	250
Total	47	87	193	180	43	550

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be ≥ 5

The sampling distribution for the test is a chi-square distribution with

$(r-1)(c-1)$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

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

where O represents the observed frequencies and E represents the expected frequencies. To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

Performing a Chi-Square Test for Independence (p 507)		
Procedure	Equations	Example2 **(p 507)**
Identify the claim. State the null and alternative hypotheses.	State H_0 and H_1	H_0: CEO's ages are independent of company size H_1: CEO's ages are dependent on company size.
Specify the level of significance	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	$d.f. = (r-1)(c-1)$	$d.f. = (2-1)(5-1) = 4$
Find the critical value.	χ^2_α :Obtain from Table 6, Appendix B	$\chi^2_\alpha \geq 13.277$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 13.277$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$\sum \frac{(O-E)^2}{E} \approx 77.9$ Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above

Make a decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 77.9 > 13.277 we reject the null hypothesis Equivalently $P(X \geq 77.0) < \alpha$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		CEO's ages and company size are dependent.

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

   Solution:

   There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion p = 5/20 = 0.25. Using the PROB-VALUE method the steps in this test are:

   1) H_0: π = 0.5 and H_A: π ¹ 0.5
   2) We will use the Z-distribution
   3) We will use the 5%-level, thus α = 0.05
   4) The test statistic is $z = (0.25 - 0.5)/\sqrt{0.25/20} = -2.24$
   5) Table A-4 shows that P(|Z| > 2.24) » 0.025.
   6) Because PROB-VALUE <α, we reject H_0. We conclude π is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

   Solution:

   Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

   $$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$
   $$P(X \geq 8) = 0.1208 \quad + 0.0537 \quad + 0.0161 \quad + 0.0029 \quad + 0.002 \quad = 0.1937$$

   Adopting the 5% uncertainty level, we see that PROB-VALUE >α. Thus we fail to reject H_0. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

   Solution:

   (a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

(b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference. We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

High Density	Low Density	Sparsely Settled
1.84	2.04	1.07
3.06	2.28	2.31
3.62	4.01	0.91
4.91	1.86	3.28
3.49	1.42	1.31

Solution:

We will use the multi-sample Kruskal-Wallis test with an uncertainly level α = 0.1. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5} + \frac{39^2}{5} + \frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the $\chi 2$ distribution with $3-1=2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

	Distance (km)			Distance (km)	
Person	1996	2006	Person	1996	2006
1	8.6	8.8	7	7.7	6.5
2	7.7	7.1	8	9.1	9
3	7.7	7.6	9	8	7.1
4	6.8	6.4	10	8.1	8.8
5	9.6	9.1	11	8.7	7.2
6	7.2	7.2	12	7.3	6.4

Has the length of the journey to work changed over the decade?

Solution:

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $H_0 : \eta = 0$ and $H_A : \eta \neq 0$. We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

S = {-,+,+,+,+,0,+,-,+,-,+,+}

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with $\pi = 0.5$. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the $\alpha = 10\%$ level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

8. One hundred randomly sampled residents of a city subject to periodic flooding are classified according to whether they are on the floodplain of the major river bisecting the

city or off the floodplain. These households are then surveyed to determine whether they currently have flood insurance of any kind. The survey results are as follows:

	On the Floodplain	Off the Floodplain
Insured	50	10
No Insurance	15	25

Test a relevant hypothesis.

Solution:

We will do a χ^2 test for a relationship between insurance and house location. The null hypothesis is no relationship (independence). Augmenting the data with expected frequencies, we have:

	On the Floodplain	Off the Floodplain
Insured	50 (39)	10 (21)
No Insurance	15 (26)	25 (14)

The corresponding χ^2 value is 22.16. Table A-8 shows that with 1 degree of freedom, $P(\chi^2 > 20)$ is zero to 3 decimal places. Thus for any reasonable level of uncertainty (any $\alpha < 0.0005$), we can reject the null hypothesis.

9. The occurrence of sunshine over a 30-day period was calculated as the percentage of time the sun was visible (i.e., not obscured by clouds). The daily percentages were:

Day	Percentage of sunshine	Day	Percentage of sunshine	Day	Percentage of sunshine
1	75	11	21	21	77
2	95	12	96	22	100
3	89	13	90	23	90

4	80	14	10	24	98
5	7	15	100	25	60
6	84	16	90	26	90
7	90	17	6	27	100
8	18	18	0	28	90
9	90	19	22	29	58
10	100	20	44	30	0

If we define a sunny day as one with over 50% sunshine, determine whether the pattern of occurrence of sunny days is random.

Solution:

For this we can use the number-of-runs test. Rather than calculate runs across two samples, here we will simply note if a day has 50% or more sunshine. The sample becomes

S={+,+,+,+,-,+,+,-,+,+,-,+,+,-,+,+,-,-,-,-,+,+,+,+,+,+,+,+,-}

We see that the sample consists of 12 runs. There are n_x = 21 sunny days, and n_y = 9 cloudy days. Because n_x< 20, we cannot use the normal approximation given in Table 10-5. Instead the probability is computed using combinatorial rules, and is approximately 0.4. This is far too large for rejection of the randomness hypothesis. We cannot conclude the pattern is non-random.

10. Test the normality of the DO data (a) using the Kolmogorov–Smirnov test with the ungrouped data of Table 2-4 and (b) using the χ^2 test with k = 6 classes of Table 2-6.

Solution:

(a) We will take the mean and standard deviation as known rather than estimated from the sample. Doing so results in calculated PROB-VALUES that are smaller than the true values (i.e., we are more likely to reject the null hypothesis). For the DO data the mean and standard deviation are 5.58 and 0.39 respectively. We sort the data, and then find the differences between the observed and expected cumulative distributions. The table below shows the results for a few of the 50 observations:

x_i	$S(x_i)$	$F(x_i)$	$\lvert S(x_i)-F(x_i)\rvert$
4.2	0.020	0.015	0.005
4.3	0.040	0.023	0.017
4.4	0.060	0.032	0.028
…	…	…	…
5.9	0.780	0.692	0.088
…	…	...	…
6.7	0.960	0.960	0.000
6.8	0.980	0.972	0.008
6.9	1.000	0.981	0.019

The maximum difference is 0.088. Table A-9 shows that with 50 degrees of freedom, the corresponding PROB-VALUE is about 0.6. We obviously cannot reject the hypothesis of normality.

(b) Here we will take the mean and standard deviation as unknown, to be estimated from the sample. In other words, we estimate two parameters from the sample. In building the χ^2 table, we combine the first two and the last two categories in Table 2-6 to ensure at least 2 expected frequencies per cell. This reduces the number of categories 4, as seen in the table below:

Group	Minimum	Maximum	O_j	E_j	$(O_j-E_j)^2/E_j$
1	4.000	4.990	9	3.3	10.13
2	5.000	5.490	10	17.0	2.89
3	5.500	5.990	20	21.7	0.14
4	6.000	6.990	11	7.0	2.24

The observed Chi-square value is 15.4. With $k-p-1=4-2-1=1$ degrees of freedom, Table A-8 shows that the PROB-VALUE is less than 0.0005. We therefore reject the null hypothesis.

Note that with only 4 classes, we can obtain only a rough idea of the distribution of DO. The 4 classes given in Table 2-6 do not yield a distribution that is at all similar to the normal distribution. In practice one would need many classes (and observations) for the χ^2 test to be reliable.

UNIT V

1	5	The hypothesis under test is	simple hypothesis	alternative hypothesis	**null hypothesis**	complex hypothesis	**null hypothesis**
2	5	A test based on a test statistic is classified as	**randomized test**	non-randomized test	sequential test	Bayes test	**randomized test**
3	5	Student's t test is applicable in case of	**small samples**	for sample of size between 5 and 30	large samples	sample size more than 100	**small samples**
4	5	Reject H_0 when it is false is known as	Type I error	Type II error	**Correct decision**	wrong decision	**Correct decision**
5	5	Accept H_0 when it is true is known as	Type I error	Type II error	**Correct decision**	wrong decision	**Correct decision**
6	5	Accept H_0 when it is false is known as	Type I error	**Type II error**	Correct decision	wrong decision	**Type II error**
7	5	In a sample of 10 items the degree of freedom for student's t test is	10	**9**	12	11	**9**
8	5	If the sample size is less than 30 then those samples may be regarded as	large samples	**small samples**	parameter	attitude	**small samples**
9	5	The range of statistic-t is	-1 to +1	**$-\infty$ to $+\infty$**	0 to ∞	0 to 1	**$-\infty$ to $+\infty$**
10	5	The distribution used to test goodness of fit is	F distribution	**χ^2 distribution**	t distribution	Z distribution	**χ^2 distribution**
11	5	Large sample theory is applicable when	**n>30**	n<30	n=30	n=10	**n>30**
12	5	95% of fiducial limits of population mean are	**A.M ± 1.96 S.E**	A.M ± 3.96 S.E	A.M ± 2.58 S.E	A.M ± 3.58 S.E	**A.M ± 1.96 S.E**
13	5	99% of fiducial limits of population mean are	A.M ± 1.96 S.E	A.M ± 3.96 S.E	**A.M ± 2.58 S.E**	A.M ± 3.58 S.E	**A.M ± 2.58 S.E**
14	5	A part of the population selected for study is called as	statistic	**sample**	parameter	event	**sample**
15	5	Test statistic Z =	**(X -µ)/S.E(X)**	(X +µ)/S.E(X)	Xµ/S.E(X)	(X* µ)/S.E(X)	**(X -µ)/S.E(X)**
16	5	In students t test standard error of single mean =	**S/√n-1**	S/√n-2	S/√n+1	S/√n+2	**S/√n-1**
17	5	Null hypothesis is denoted by	H_1	**H_0**	H_2	H	**H_0**
18	5	Alternative hypothesis is denoted by	H_0	**H_1**	H_2	H	**H_1**
19	5	The value of 5% level of significance is	2.58	**1.96**	1.64	2.33	**1.96**
20	5	The value of 1% level of significance is	**2.58**	1.64	1.96	2.33	**2.58**
21	5	If the sample size n=2, the students t-distribution reduces to	Normal distribution	F- distribution	**Cauchy distribution**	z – distribution	**Cauchy distribution**
22	5	If n, the sample size is larger than 30, the students t-distribution reduces to	**Normal distribution**	F- distribution	Cauchy distribution	Z- distribution	**Normal distribution**
23	5	The d.f. for students t based on a random sample of size n is	**n-1**	n	n-2	(n-2)/2	**n-1**
24	5	Degrees of freedom for statistic-chi-square in case of contingency table of order(3x3)is	3	**4**	2	1	4
25	5	Students 't' test is defined by the statistic	t = x-µ/√^{s}/n	√^{t} = x-µ/√^{s}/n	**t = x-µ/s/√^{n}**	t = x-µ/s/n	**t = x-µ/s/√^{n}**
26	5	Chi-square variate has degree of freedom	2	**1**	5	7	**1**
27	5	Degree of freedom is denoted by	Φ	**ν**	γ^2	fi	**ν**
28	5	Students t- distribution, the name was given by	Arthur Henshalme	Beniamin	**William S.Cosset**	Karl Pearson	**William S.Cosset**
29	5	Chi-Square test was first used in testing statistical hypothesis by	**Karl Pearson**	Robert mayer	A.Fischer	Spearman	**Karl Pearson**
30	5	While testing significance of the difference of two sample means in case of small samples, the degree of freedom is calculated by	$V=n_1+n_2$	$V=n_1+n_2-1$	**$V=n_1+n_2-2$**	$V=n_1+n_2+2$	**$V=n_1+n_2-2$**
31	5	Analysis of variance utilizes	**F- test**	Chi-square test	Z-test	t-test	**F- test**
32	5	Analysis of variance technique originated in	**Agricultural research**	Industrial research	Biological research	Business research	**Agricultural research**
33	5	Analysis of variance technique was developed by	Gosset	**R.A.Fisher**	Laplace	Karl Pearson	**R.A.Fisher**
34	5	Design of experiment was introduced by	**R.A.Fisher**	Spearman	Karl Pearson	Lorenz	**R.A.Fisher**
35	5	The variation within samples is known as	error	**block**	degrees of freedom	local control	**block**

60 5

Reg.No. ______________

[16BCU602 A]

KARPAGAM ACADEMY OF HIGHER EDUCATION

COIMBATORE – 21

DEPARTMENT OF BIOCHEMISTRY

III B.SC., BIOCHEMISTRY - SIX SEMESTER

16BCU602 A – BIOSTATISTICS

FIRST INTERNAL EXAMINATION -DECEMBER 2018

DATE: 17.12.18 AN

TIME: 2 HOURS MAXIMUM: 50 MARKS

PART- A (20 x 1 = 20 marks)

Answer ALL the following:

1. Data taken from the publication, 'Agricultural Situation in India' will be considered as
(a) primary data (b) secondary data
© both primary and secondary data (d) tertiary data

2. Statistics deals with
(a)qualitative information (b)quantitative information
c. both qualitative and quantitative information (d)only numerals

3. Year wise recording of data of food production will be called as:
(a) Geographical classification (b)Chronological classification
(c) Quantitative classification (d)qualitative classification

4. Who is the father of Biostatistics?
(a) R.A.Fisher (b) W.Gosset
(c) Sir Francis Galton (d) S.C.Gupta

5. Statistics can be considered as
a. an art b. a science c. both an art and a science
d. neither an art nor a science

6. The most suitable form of presentation for publicity and Propaganda is
(a) diagram (b) graph
(c) map (d) numerals

7. Histogram is suitable for the data presented as
(a) continuous grouped frequency distribution
(b) discrete grouped frequency distribution
(c) individual series
(d). discontinuous series

8. Numerical data presented in descriptive form are called
(a) classified presentation (b) tabular presentation
(c) graphical presentation (d) textual presentation

9. Mean is a measure of
(a) central value (b) dispersion
(c) correlation (d) significance

10. Mode is that value in a frequency distribution which possesses
(a). minimum frequency (b). frequency one
(c). maximum frequency (d). only two frequency

11. The most stable measure of central tendency is
(a) the mean (b) the median
(c) the mode (d) percentile

12. Sum of the deviations about mean is:
(a) minimum (b) zero (c) maximum (d) two

13. Mode of the following data 3, 6, 5, 7, 8, 4, 9
(a) 3 (b) 7 (c) no mode (d) 5

14 . Which of the following is a measure of central tendency?
(a) Range (b) Quartile deviation
(c) Standard deviation (d) Median

15. Median is also called as
(a) first quartile (b) second quartile
(c) third quartile (d) fourth quartile

16. The census data published for state wise population in India will be known as
(a) Quantitative classification (b) Two-way classification
(c) Geographical classification (d) one way classification

17. Classification according to class-intervals would yield
(a) raw data (b) discrete data
(c) qualitative data (d) grouped data

18. In qualitative classification data are classified on the basis of
(a) attributes (b) time (c) location (d) class intervals

19 In geographical classification data are classified on the basis of
(a) area (b) attributes (c) time (d) location

20. Which source is one that itself collects the data?
(a) primary data (b) secondary data
(c) published data (d) tertiary data

PART B (3x2=6 marks)
Answer ALL the questions

21. List out the parts of table.
22. What are the various methods of collecting primary data?
23. Define Median and give Example

PART C (3x8=24 marks)
Answer ALL the questions

24.a. Explain about the Classification of data.

OR

b. Draw a suitable Pie Diagram to represent the following submitted as a part of the budget proposal of the govt. of India for the year 1995 – 96.

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

25.a. . Calculate the Arithmetic Mean for the following data.

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

OR

b. Explain in detail the various methods of collecting primary data.

26. a. . Calculate the Median for the following Continuous Frequency Distribution.

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

OR

b. . Differentiate diagrams and graph. In what way the graphic presentation is advantageous than any other method?

Class : III Bsc Biochemistry.
Sub : Biostatistics
Sub code : 16BCU602A
No of copies : 30