2019-2020 Semester – П

End Semester Exam: 3 Hours

19MBP203BIOSTATISTICS AND RESEARCH METHODOLOGY4H-4C

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

COURSE OBJECTIVES

This course enables the students to learn

- About collection, interpretation and presentation of statistical data
- The analytics of data, probability, and hypothesis testing of samples
- The essential role of statistics in present, future use and applications of Biology.

COURSE OUTCOMES

On successful completion of this course the learners will be able to

- 1. Apply basic statistical concepts commonly used in health and medical sciences
- 2. Use basic analytical techniques to generate results
- 3. Interpret results of commonly used statistical analyses in written summaries
- 4. Demonstrate statistical reasoning skills correctly and contextually

UNIT I - Introduction of Biostatistics and Correlation

Introduction to Biostatistics, Basic Measures - Central Tendency and Dispersion, Variables in Bioscience, Correlation – Meaning and definition - Scatter diagram –Karl Pearson's Correlation Coefficient. Rank Correlation. Regression: Regression in two variables – Properties of Regression, uses of Regression

UNIT II - Test of Significance

Sampling parameters: Difference between sample and Population, Censoring, difference between parametric and non-parametric statistics. Sampling Distributions, Standard Error, Testing of Hypothesis, Level of Significance and Degree of Freedom, Confidence Interval;, Small sample test based on t-test, Large Sample Test based on Normal Distribution - Z- test and F test.

UNIT III Analysis of Variance

Distribution-free test - Chi-square test; Basic Introduction to Multivariate statistics, etc. Test of significance: Tests based on Means only-Both Large sample and Small sample tests - Chi square test - goodness of fit. Analysis of Variance – one way and two way classification, CRD, RBD Designs.

UNIT IV-Research

Research: Scope and significance – Types of Research – Research Process – Characteristics of good research – Problems in Research – Identifying research problems.

UNIT V - Sampling Design

Research Designs – Features of good research designs. Sampling Design: Meaning – Concepts – Steps in sampling – Criteria for good sample design. Scaling measurements - Types of scale, Types of sampling – random sampling and non- random sampling. Sampling Errors.

Suggested Readings

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

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



KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Mathematical Statistics	Semester VI	LTPC
Subject Code : 16MMU603A	Class : III B.Sc Mathematics	4204

Glossary of Statistical Terms

2 X 5 factorial	A factorial design with one variable having two levels and the other	
design	having five levels.	
Alpha	The probability of a Type I error.	
Abscissa	Horizontal axis.	
Additive law of	The rule giving the probability of the occurrence of one or more mutually	
probability	exclusive events.	
Adjacent values	Actual data points that are no more extreme than the inner fences.	
Alternative	The hypothesis that is adopted when H_0 is rejected. Usually the same as	
hypothesis (H ₁)	the research hypothesis.	
ß (Beta)	The probability of a Type II error.	
Categorical data	Data representing counts or number of observations in each category.	
Call	The combination of a particular row and column (the set of observations	
Cell	obtained under identical treatment conditions.	
Central limit	The theorem that specifies the nature of the sampling distribution of the	
theorem	mean.	
Chi-square test	A statistical test often used for analyzing categorical data.	
Conditional	The probability of one event <i>given</i> the occurrence of some other event	
probability	The probability of one event given the occurrence of some other event.	
Confidence	An interval, with limits at either end, with a specified probability of	
interval	including the parameter being estimated.	
Confidence	An interval, with limits at either end, with a specified probability of	
limits	including the parameter being estimated.	
Constant	A number that does not change in value in a given situation.	
Continuous	Variables that take on <i>any</i> value	
variables	variables that take on any value.	
Correlation	Relationship between variables.	
Correlation	Δ measure of the relationship between variables	
coefficient	A measure of the relationship between variables.	
Count data	Data representing counts or number of observations in each category.	
Covariance	A statistic representing the degree to which two variables vary together.	
Criterion	The veriable to be predicted	
variable		
Critical value	The value of a test statistic at or beyond which we will reject H0.	
Decision	A procedure for making logical decisions on the basis of sample data	
making	A procedure for making logical decisions on the basis of sample data.	
Degrees of	The number of independent pieces of information remaining after	

M.Jeganathan, Department of Mathematics, KAHE

freedom (df)	estimating one or more parameters.	
Danaity	Height of the curve for a given value of X- closely related to the	
Density	probability of an observation in an interval around X.	
Dependent		
variables	The variable being measured. The data or score.	
Discrete		
variables	variables that take on a small set of possible values.	
D' '	The degree to which individual data points are distributed around the	
Dispersion	mean.	
Distribution free	Statistical tests that do not rely on parameter estimation or precise	
tests	distributional assumptions.	
T 22	The difference between two population means divided by the standard	
Effect size	deviation of either population.	
	The degree to which repeated values for a statistic cluster around the	
Efficiency	narameter	
Event	The outcome of a trial	
Event	Δ set of events that represents all possible outcomes	
LAndustive	The average value calculated for a statistic over an infinite number of	
Expected value	samples	
Exposted	samples.	
fraguencies	The expected value for the number of observations in a cell if H ₀ is true.	
Europeine entel		
Experimental	Another name for the research hypothesis.	
Translamate and data	A set of to shall made have been the Tabase for a more atting that is assessed by	
Exploratory data	A set of techniques developed by Tukey for presenting data in visually	
analysis (EDA)	meaningful ways.	
	The ability to generalize the results from this experiment to a larger	
Frequency	A distribution in which the values of the dependent variable are tabled or	
distribution	plotted against their frequency of occurrence.	
Frequency data	Data representing counts or number of observations in each category.	
Goodness of fit	A test for comparing observed frequencies with theoretically predicted	
test	frequencies.	
Grand total	The sum of all of the observations.	
(∑X)		
Hypothesis	A process by which decisions are made concerning the values of	
testing	parameters.	
Independent	Those variables controlled by the experimenter	
variables	Those variables controlled by the experimenter.	
Independent	Events are independent when the occurrence of one has no effect on the	
events	probability of the occurrence of the other.	
Interaction	A situation in a factorial design in which the effects of one independent	
Interaction	variable depend upon the level of another independent variable.	
Intercept	The value of Y when X is 0.	
Internal apple Scale on which equal intervals between objects represent equal d		
interval scale	«differences are meaningful.	
Interval estimate	A range of values estimated to include the parameter.	
Joint probability	The probability of the co-occurrence of two or more events.	
Kurtosis	A measure of the peakedness of a distribution.	
Leptokurtic	A distribution that has relatively more scores in the center and in the tails.	

Linear relationship	A situation in which the best-fitting regression line is a straight line.		
Linear regression	Regression in which the relationship is linear.		
Marginal totals	Totals for the levels of one variable summed across the levels of the other variable.		
Matched	An experimental design in which the same subject is observed under more		
samples			
Mean absolute			
deviation	Mean of the absolute deviations about the mean.		
(m.a.d.)			
Mean	The sum of the scores divided by the number of scores.		
Measurement	The assignment of numbers to objects.		
Measurement data	Data obtained by measuring objects or events.		
Measures of central tendency	Numerical values referring to the center of the distribution.		
Median location	The location of the median in an ordered series.		
Median (Med)	The score corresponding to the point having 50% of the observations below it when observations are arranged in numerical order.		
Mesokurtic	A distribution with a neutral degree of kurtosis.		
Midpoints	Center of interval average of upper and lower limits.		
Mode (Mo)	The most commonly occurring score.		
Monotonic	A relationship represented by a regression line that is continually		
relationship	increasing (or decreasing), but perhaps not in a straight line.		
Multiplicative			
law of	The rule giving the probability of the joint occurrence of independent		
probability	events.		
Mutually	Two events are mutually exclusive when the occurrence of one precludes		
exclusive	the occurrence of the other.		
Negative	A relationship in which increases in one variable are associated with		
relationship	decreases in the other.		
Negatively	A distribution that trails off to the left		
skewed	A distribution that trans off to the left.		
Nominal scale	Numbers used only to distinguish among objects.		
normal	A specific distribution begins a characteristic hall shaped form		
distribution	A specific distribution having a characteristic ben-shaped form.		
Ordinal scale	Numbers used only to place objects in order.		
Ordinate	Vertical axis.		
Outlier	An extreme point that stands out from the rest of the distribution.		
<i>p</i> level	The probability that a particular result would occur by chance if H_0 is true. The exact probability of a Type I error.		
Parameters	Numerical values summarizing population data.		
Parametric tests	Statistical tests that involve assumptions about, or estimation of, population parameters.		
Pearson product-moment correlation coefficient (<i>r</i>)	The most common correlation coefficient.		

Percentile	The point below which a specified percentage of the observations fall.	
Phi	The correlation coefficient when both of the variables are measured as dichotomies.	
Platykurtic	A distribution that is relatively thick in the "shoulders."	
Pooled variance	A weighted average of the separate sample variances.	
Population variance	Variance of the population (usually estimated, rarely computed.	
Population	Complete set of events in which you are interested.	
Positively skewed	A distribution that trails off to the right.	
Power	The probability of correctly rejecting a false H ₀ .	
Predictor variable	The variable from which a prediction is made.	
Protected t	A technique in which we run <i>t</i> tests between pairs of means only if the analysis of variance was significant.	
Quantitative data	Data obtained by measuring objects or events.	
Random sample	A sample in which each member of the population has an equal chance of inclusion.	
Random Assignment	Assigning participants to groups or cells on a random basis.	
Range	The distance from the lowest to the highest score.	
Range restrictions	Refers to cases in which the range over which X or Y varies is artificially limited.	
Ranked data	Data for which the observations have been replaced by their numerical ranks from lowest to highest.	
Rank - randomization tests	A class of nonparametric tests based on the theoretical distribution of randomly assigned ranks.	
Ratio scale	A scale with a true zero point ratios are meaningful.	
Real lower limit	The points halfway between the top of one interval and the bottom of the next.	
Real upper limit	The points halfway between the top of one interval and the bottom of the next.	
Rectangular distribution	A distribution in which all outcomes are equally likely.	
Regression	The prediction of one variable from knowledge of one or more other variables.	
Regression equation	The equation that predicts Y from X.	
Regression coefficients	The general name given to the slope and the intercept (most often refers just to the slope.	
Rejection region	The set of outcomes of an experiment that will lead to rejection of H_0 .	
Rejection level	The probability with which we are willing to reject H0 when it is in fact correct.	
Related samples	An experimental design in which the same subject is observed under more than one treatment.	
Relative frequency view	Definition of probability in terms of past performance.	

Research	The hypothesis that the experiment was designed to investigate		
hypothesis	The hypothesis that the experiment was designed to investigate.		
Sample	Set of actual observations. Subset of the population.		
Sample statistics	Statistics calculated from a sample and used primarily to describe the sample.		
Sample variance (s ²)	Sum of the squared deviations about the mean divided by N - 1.		
Sample with	Sampling in which the item drawn on trial N is replaced before the		
replacement	drawing on trial $N + 1$.		
Sampling			
distribution of	The distribution of the differences between means over repeated sampling		
differences	from the same population(s).		
between means			
Sampling distribution of the mean	The distribution of sample means over repeated sampling from one population.		
Sampling	The distribution of a statistic over repeated sampling from a specified		
distributions	population.		
Sampling error	Variability of a statistic from sample to sample due to chance.		
Scales of			
measurement	Characteristics of relations among numbers assigned to objects.		
Scatter plot	A figure in which the individual data points are plotted in two-dimensional space.		
Scatter diagram	A figure in which the individual data points are plotted in two-dimensional space.		
Scattergram	A figure in which the individual data points are plotted in two-dimensional space.		
Sigma	Symbol indicating summation.		
Significance	The probability with which we are willing to reject H_0 when it is in fact		
level	correct.		
Simple effect	The effect of one independent variable at one level of another independent variable.		
Skewness	A measure of the degree to which a distribution is asymmetrical.		
Slope	The amount of change in Y for a one unit change in X.		
Spearman's correlation coefficient for ranked data (r_s)	A correlation coefficient on ranked data.		
Standard deviation	Square root of the variance.		
Standard error	The standard deviation of a sampling distribution.		
Standard error	The standard deviation of the sampling distribution of the differences		
of differences	hetween means		
between means			
Standard error	The average of the squared deviations about the regression line		
of estimate	The average of the squared deviations about the regression fille.		
Standard scores	Scores with a predetermined mean and standard deviation.		
Standard normal	A normal distribution with a mean equal to 0 and variance equal to 1.		

distribution	Denoted <i>N</i> (0, 1).
Statistics	Numerical values summarizing sample data.
Student's <i>t</i> distribution	The sampling distribution of the <i>t</i> statistic.
Subjective probability	Definition of probability in terms of personal subjective belief in the likelihood of an outcome.
Sufficient statistic	A statistic that uses all of the information in a sample.
Sums of squares	The sum of the squared deviations around some point (usually a mean or predicted value).
Symmetric	Having the same shape on both sides of the center.
T scores	A set of scores with a mean of 50 and a standard deviation of 10.
Test statistics	The results of a statistical test.
Type I error	The error of rejecting H_0 when it is true.
Type II error	The error of not rejecting H_0 when it is false.
Unconditional probability	The probability of one event <i>ignoring</i> the occurrence or nonoccurrence of some other event.
Unimodal	A distribution having one distinct peak.
Variables	Properties of objects that can take on different values.
Weighted average	The mean of the form: $(a_1X_1 + a_2X_2)/(a_1 + a_2)$ where a_1 and a_2 are weighting factors and X_1 and X_2 are the values to be average.



KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics		
Subject : Biostatistics and Research Metl	hodology Semester II	LTPC
Subject Code : 19MBP203	Class : I M.Sc Microbiology	4004

UNIT I

Introduction to Biostatistics, Basic Measures - Central Tendency and Dispersion, Variables in Bioscience, Correlation – Meaning and definition - Scatter diagram –Karl Pearson's Correlation Coefficient. Rank Correlation. Regression: Regression in two variables – Properties of Regression, uses of Regression

Suggested Readings

- 1. Jerrold H. Zar. (2003). *Biostatistical Analysis*. (4th ed.). Pearson Education (P) Ltd, New Delhi.
- 2. Kothari. C.R. (2004). *Research Methodology Methods and Techniques*. (2nd ed.). New Age International Pvt. Ltd, New Delhi.

Unit – I

Introduction

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

Origin and Growth

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

Meaning

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

Definitions

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

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

Functions

The following are the important functions of statistics.

- σ Collection
- σ Numerical Presentation
- σ Diagrammatic Presentation
- σ Condensation
- σ Comparison
- σ Forecasting
- σ Policy Making
- σ Effect Measuring
- σ Estimation
- σ 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.

Samples vs. Populations

Population: A complete set of observations or measurements about which conclusions are to be drawn.

Sample: A subset or part of a population.

Not necessarily random

Statistics vs. Parameters

Parameter: A summary characteristic of a population.

Summary of Central tendency, variability, shape, correlation

E.g., Population mean, Population Standard Deviation, Population Median, Proportion of population of registered voters voting for Bush, Population correlation between Systolic & Diastolic BP

Statistic: A summary characteristic of a sample. Any of the above computed from a sample taken from the population.

E.g., Sample mean, Sample Standard Deviation, median, correlation coefficient

MEASURES OF CENTRAL TENDENCY

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

- Arithmetic Mean or simply Mean
- Median
- Mode

Arithmetic Mean or 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 \overline{X}

Individual series

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

Example 1

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

30, 70, 10, 75, 500, 8, 42, 250, 40, 36

Solution

Here N=10

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

X = 1061 / 10 = 106.1

Discrete series

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

Example 2

Calculate the mean number of person per house.

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

Solution

Х	f	fХ	
2	10	20	
3	25	75	
4	30	120	
5	25		125
6	<u>10</u>	60	
	Σ f =100	Σ f X= 400	
_			

X = 400 / 100 = 4.

Continuous series

_

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

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

Example 3

Calculate the mean for the following.

Marks :	20-30 30-40 40-50	50-60 60-70 70-80
	20 30 30 40 40 30	50 00 00 70 70 00

No. of student : 5 8 12 15 6 4

Solution

C.I	f	m	fm
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 <u>4</u> 75 <u>300</u> $\Sigma f = 50$ $\Sigma f m = 2460$

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 items Median Position = (N+1) / 2

For even number of item

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

Example 1

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 2

Find the median for the following.

X:10 15 17 18 21

F: 4 16 12 5 3

Solution

Х	f	c.f
10	4	4
15	16	20
17	12	32
18	5	37
21	<u>3</u>	40
	N= 40	

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

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

= 16.

Continuous series

 $M = L + [((N/2) - c.f) \times i]$ f.

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

1

Example 3

Calculate the median height (Ht) for the No .of Students (NoS) given below.

Ht: 145-150 150-155 155-160 160-165 165-170 170-175

NoS: 2	5	10	8	4
Solution				
Height	No. of st	udent	c.f	
145-150	2		2	
150-155	5		7	
155-160	10		17	
160-165	8		25	
165-170	4		29	
170-175	<u>1</u>		30	
	Σ f = 30			

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

$$M = L + [((N/2) - c.f) \times i]$$

f.
= 155+ [(15-7)x5] = 155+(40/10) = 159.
10

Mode

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

Individual series

In a set of observations the value which occur more number of time is known as Mode. In other way the most frequented value in a set of value is Mode.

Example 1

Determine the mode for the set of Individual observations given as follows 32, 35, 42, 32, 42, 32.

Solution

Mode = 32 Uni-model

Discrete series

Determine the Mode

Size of dress No. of set

18	55
20	120
22	108
24	45

Here the mode represents highest frequency ie. I 120.

So, 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

Marks : 0-10 10-20 20-30 30-40 40-

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)

Continuous series

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

Calculate the range for the following continuous frequency distribution

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

Unit I

Here L=75 & S=25

Range = L - S = 75-25 = 50

Quartile Deviation

Quartile Deviation is half of the difference between the first and the third quartiles. Hence it is called Semi Inter Quartile Range.

Coefficient of Quartile Deviation

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as

$$=\frac{\frac{Q_{3}-Q_{1}}{2}}{\frac{Q_{3}+Q_{1}}{2}}=\frac{Q_{3}-Q_{1}}{Q_{3}+Q_{1}}$$

It is pure number free of any units of measurement. It can be used for comparing the dispersion in two or more than two sets of data.

Example

The wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, and 1885. Find the quartile deviation and coefficient of quartile deviation.

Solution

After arranging the observations in ascending order, we get

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right) th \text{ item}$$

= Value of $\left(\frac{20+1}{4}\right) th$ item

= Value of
$$(5.25)$$
th item

= 5th item + 0.25(6th item - 5th item) = 1240 + 0.25(1320 - 1240)

$$Q_1 = 1240 + 20 = 1260$$
$$Q_3 = \text{Value of } \frac{3(n+1)}{4}th \text{ item}$$
$$= \text{Value of } \frac{3(20+1)}{4}th \text{ item}$$

= Value of (15.75)th item

= 15th item + 0.75(16th item - 15th item) = 1750 + 0.75(1755 - 1750)

 $Q_3 = 1750 + 3.75 = 1753.75$

Quartile Deviation (Q.D)

$$=\frac{Q_3-Q_1}{2}=\frac{1753.75-1260}{2}=\frac{492.75}{2}=246.875$$

Coefficient of Quartile Deviation

$$=\frac{Q_3-Q_1}{Q_3+Q_1}=\frac{1753.75-1260}{1753.75+1260}=0.164$$

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 1

Calculate standard deviation for the following data.

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

Solution

Х	X ²
40	1600
41	1681
45	2025
49	2401
50	2500
51	2601

55	3025
59	3481
60	3600
<u>60</u>	<u>3600</u>
510	$\Sigma x^2 = 26504$

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

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

= 7.09

Discrete series

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

Example 2

Calculate standard deviation for the following data.

X: 0	1	2	3	4	5	
F: 1	2	4	3	0	2	
Solutio	n					
x	f		fx		x²	fx²
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

2

5

10

25

<u>50</u>

	Σf = 12	Σ fx= 29	$\Sigma fx^2 = 95$
σ = √	($\Sigma fx^2 / \Sigma$	f) –(Σ fx / Σ f) ²	
= √	(95/12)-	- (29 /12) ²	= 1.44

Continuous series

$\sigma = \sqrt{(\Sigma \text{ fm}^2 / \Sigma \text{ f}) - (\Sigma \text{ fm} / \Sigma \text{ f})^2}$								
Examp	le 3							
C.I :	0-10	10-20	20-30	30-40	40-50			
F :	2	5	9	3	1			
Solutio	on							
C.I	f	m	fm	m²	fm²			
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	<u>1</u>	45	<u>5</u>	2025	2025			
	20		460		12500			

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

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

= 9.79

Coefficient of variation

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

Example 1

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

CORRELATION AND REGRESSION ANALYSIS

3.1 Simple Linear Correlation

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

3.2 Type of Correlation

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	34	25	18	10	7
Y:	23	32	37	41	46	50	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.

3.3 Methods of Finding Correlation Coefficient

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).

3.3.1 Scatter Diagram

Scatter diagram is a graphic picture of the sample data. Suppose a random sample of n pairs of observations has the values $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$. These points are plotted on a rectangular co-ordinate system taking independent variable on X-axis and the dependent variable on Y-axis. Whatever be the name of the independent variable, it is to be taken on X-axis. Suppose the plotted points are as shown in figure (a). Such a diagram is called scatter 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.

3.3.2 Karl Pearson's Correlation 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

$$r = \frac{\sum xY - \frac{\sum x\sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

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	ху	\mathbf{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 $43 \times 99 =$
4.257				

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

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

Step 6: Use the following formula to work out the correlation coefficient.

The answer is: $1.3787 \times 10-4$ the range of the correlation coefficient is from -1 to 1. Since our result is $1.3787 \times 10-4$, a tiny positive amount, we can't draw any conclusions one way or another.

3.3.3 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

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

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

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

Example 1

X :	21	36	42	37	25
Y :	47	40	37	42	43

For the data given above, calculate the rank correlation coefficient.

Solution

		Rank of	X and Y					
Х	Y	R(X)	R(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 2

Find the rank correlation coefficient for the percentage of marks secured by a group of 8 students

in Economics and Statistics.

Marks in Eco:	50	60	65	70	75	40	70	80
Marks in Stat:	80	71	60	75	90	82	70	50

Solution

		Rank of X	K and Y		
Х	Y	Х	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	$\Sigma d = 0$	$\sum d^2$
			1 Otal	<u>_</u> u = 0	=113.5

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

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

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

Therefore $\rho = 1 \text{-} \begin{pmatrix} 6\{113.5 \text{+} 0.5\} / 8(8^2 \text{-} 1) \end{pmatrix}$

3.4 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.

3.4.1 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 the coefficients 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,

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

3.4.2 Methods of forming the regression equations

- Regression equations based on normal equations.
- Regression equations based on X and Y and b_{YX} and b_{XY} .

Example 1

From the following data, obtain the two regression equations.

7

X 6 2 10 4 8

Y 9 11 5 8

use normal equations

Solution

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

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

The normal equations are,

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

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

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 2

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

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

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

iv) Regression equation of Y on X, (Y - Y)

 $= b_{yx} (X-X) (Y-38) = -0.6643(X-32) \Longrightarrow$

Y = 59.26 - 0.6643X

(ii) Regression equation of X on Y, ($X - X^{()}$)

 $= b_{xy} (Y-Y) (X-32) = -0.2337Y + 8.88 \Longrightarrow$

X = 40.88 - 0.233 Y

- (iii) $r = +\sqrt{b_{yx}b_{xy}} = -0.3940$
- (iv) $Y = 59.26 0.6643 \times 30 = 39$

3.4.3 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.

Questions

1)	Calculate th	ne Mean	for the	followir	ıg.					
	Х	20	30	35	15	10				
	f	2	3	4	3	2				
2)	Define Med	lian and	give Exa	mple.						
3)	Calculate th	ne Range	and its	Coeffici	ent for	the follo	wing data.			
	X :	12	14	16	18	20				
	f :	1	3 5	53	1					
4)	What do yo	u mean	by Bimc	dal?						
5)	Calculate th	ne Media	an for th	e follow	ing data	э.				
	80 100	50	90	120	110					
6)	Write the re	elation b	etween	Standa	rd Devia	ition and	Variance.			
7)	Calculate th	ie Avera	ge numl	per of st	udents	per class	for the fol	llowing data.		
2	26 46	33	25	36	27	34	29			
8)	Find Media	n and M	ode for	the follo	owing da	ata.	4.2			
2	13 16 1/	15	18	14	19	15	12			
9) 10)	Define the 1	term Rai	nge.							
10)	Find the Ari	thmetic	iviean to	or the to	or	data.				
11)	0 60 Calculate th	75 0 Dongo	50 and ita	4Z	95 ont for i	40 the fello	uing data			
11)		E Kange		12		10	1 A Udid.			
12)	1/ 10 Define the t	50 term Ou	19 artila De	IZ	11	10	14			
12)	Find the me	dian for	· 57 58	61 /2	38 65 -	72 and 6	6			
14)	Write the e	mnirical	relation	01, 42, for Mo	de 05, 1	, 2, and 0	0			
15	What are t	he types	s of Cor	relatio	19					
16)	Write any	two pro	perties	of Corr	r. elation					
17)	What is the	e range	of Corr	elation	Coeffic	cient?				
18)	Define Pos	sitive Co	orrelatio	on.	000111					
19)	What is me	eant by	Regress	sion?						
20)	What are th	ne form	ulae for	Regree	ssion co	o-efficie	nts?			
$(21)^{-3}$	Distinguish	betwee	n Corre	lation a	nd Reg	ression.				
(22)	Write the f	ormula	for Ran	k Corre	elation.	when m	ore than o	one rank is i	repeated.	
23)	If $\mathbf{b}_{xy} = -0$.	2337 ar	$d b_{w} =$	- 0.664	3 then	find the	Correlatio	on Coefficie	ent.	
$(24)^{-2}$	What is Ne	gative (Correlat	ion? G	ive an e	example	?			
(25)	Write down	the for	rmula fo	or Karl	Pearso	n's Coef	ficient of	Correlation	1.	
26)	Define Scat	tter Dia	gram.						-	
$(27)^{-0}$	What is Sir	nple Co	orrelatio	n?						
(28)	Define Reg	ression	Equatio	on.						
29)	When X –	40 Y -	60 g	= 10 o	v = 15	and $\mathbf{r} = 0$) 7 find th	ne Regressio	on Equation of	Y on X
47)	$\Lambda - \Lambda$	то, 1 —	$00, 0_X$	- 10, 0	$\gamma - 13$	and 1 - 0	J. / IIIU UI	ie regiessie	m Equation of	$1 \text{ OII } \Lambda$.

Exercise

1)	Calculate	the Mod	e for th	e follov	ving Co	ontinuo	ıs Frequ	iency l	Distributi	on.		
Sal	ary (in Rs.	1000s)	0 -19	20-39	40 - 59	960-79	80-99					
No	. of Employ	yees	5	20	35	20	12					
2)	Find the M	Iean and	l the Sta	andard]	Deviati	on for t	he giver	n belov	w data set			
	10	14	20	12	21	16	19	17	14	25		
3)	Calculate	the Stan	dard De	eviation	and Co	oefficie	nt of Va	riance	(CV) for	the fo	ollowing	g data.
	Х	0 – 10	10 - 20	20 - 30	30 - 4	0	40 - 50)				-
	f	2	5	9	3	1						
4)	Calculate	the Med	ian for	the foll	owing (Continu	ous Free	quency	y Distribu	ition.		
,	Wages (in	Rs.)	0 - 19	20 - 39	40 - 59	960-7	9	80 - 9	, 99			
	No. of Wo	orkers	5	20	35	20	12					
5)	Calculate	the Coef	ficient	of Varia	ation fo	or the fo	llowing	data.				
	X 6	9	12	15	18		-					
	f 7	12	13	10	8							
6)	Calculate	the Med	ian for	the foll	owing.							
	Hourly Wa	ages (in	Rs.)	40 - 50	50 - 6	50 60 -	- 70 70) - 80	80 - 90	90-	- 100	
	Number of	f Emplo	yees	10	20	15	5	30	15	1(0	
7)	The follo	wing da	ata givo	e the d	letails	about	salaries	(in the	housands	of r	upees)	of seven
	employees	s randon	nly sele	cted fro	m a Ph	armace	utical Co	ompan	ıy.			
	Serial No.			1	2	3	4	5	6	7		
	Salary per	Annum	('000)	89	57	104	73	26	121	81		
	Calcul	ate the S	Standard	d Devia	tion an	d Coeff	icient of	f varia	nce of the	e giver	n data.	
8)	Calculate	the Arith	nmetic l	Mean fo	or the fo	ollowin	g data.			•		
	Height (cn	ns) : 1	60 161	1 162	163 1	64 165	5 166					
	No. of Per	sons :	27 36	43	78 (65 48	28					
9)	Calculate	the Coef	ficient	of Varia	ance fo	r the fo	llowing	data.	77	73	75	70
	72	76	75	72	74	7						
10)	Calculate	the Corr	elation	Coeffic	ient fro	om the f	ollowin	g varia	ables.			

			0					
Sales in ('0000)	57	58	59	59	60	61	62	64
Advertisement Expenditure ('000)	17	16	15	18	12	14	19	11

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

Х	25	20	28	22	40	60	20
V	40	30	50	30	20	10	30

 Y
 40
 30
 50
 30
 20
 10
 30

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

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

13) 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

14) From the data given find two Regression Equations.

Χ	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.
- 15) 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.

16) 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

17) 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.

18) You are given the following data:

	Х	Y
Arithmetic Mean	36	85
Standard Deviation	11	8
C 1.4 CC · 41	· X 1X	0.00

Correlation coefficient between X and Y = 0.66

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

Introduction to Statistics

Learning Objectives

- 1. Define statistics
- 2. Describe the uses of statistics
- 3. Distinguish descriptive & inferential statistics
- 4. Define population, sample, variable, parameter, & statistic
- 5. Explain the types of data
- 6. Compare the sources of data

What Is Statistics?

- 1. The science of data
- 2. Involves:
 - Collecting
 - Classifying
 - Summarizing
 - Organizing
 - Analyzing
 - Interpreting

What Is Statistics?

- **1.** The science of data
- 2. Involves:
 - Collecting
 - Classifying
 - Summarizing
 - Organizing
 - Analyzing
 - Interpreting

Understanding

Purposes



What Is Statistics?

1. The science of data Why? 2. **Involves:** Understanding Collecting Classifying © 1984-1994 T/Maker Co. **Purposes** Summarizing **Decision-**Organizing Making Analyzing Manna Interpreting





Accounting Auditing Costing

ManagementDescribe employeesQuality improvement



Accounting Auditing Costing 	Management Describe employees Quality improvement
Finance Financial trends Forecasting 	Marketing Consumer preferences Marketing mix effects

Statistical Methods





Descriptive Statistics

Involves
 Collecting data
 Presenting data
 Characterizing data

2. PurposeDescribe data

Descriptive Statistics

- Involves
 Collecting data
 Presenting data
 Characterizing data
- 2. PurposeDescribe data



Inferential Statistics

 Involves
 Estimation
 Hypothesis testing

Inferential Statistics

 Involves
 Estimation
 Hypothesis testing



Inferential Statistics

Population?

Involves 1. Estimation Hypothesis testing 2. Purpose Make decisions about population characteristics

1. Population: All items of interest

1. Population: All items of interest



- 1. Population: All items of interest
- 2. Variable: Population characteristic



- 1. Population: All items of interest
- 2. Variable: Population characteristic
- 3. Sample: Portion of population



- 1. Population: All items of interest
- 2. Variable: Population characteristic
- 3. Sample: Portion of population



- 1. Population: All items of interest
- 2. Variable: Population characteristic
- 3. Sample: Portion of population
- 4. Parameter: Summary measure about population



- 1. Population: All items of interest
- 2. Variable: Population characteristic
- 3. Sample: Portion of population
- 4. Parameter: Summary measure about population
- 5. Statistic: Summary measure about sample























Data Type Examples

Quantitative Interval • What is the temperature? ____ (° Celsius) Ratio How tall are you? (Inches) Qualitative Nominal Do you own savings bonds? ____ Yes ___ No Ordinal • How do you rate this (excellent, good, bad)?

Thinking Challenge

Are they quantitative or qualitative?

Gender

 Male, female

 Weight

 123, 140.2 etc.

 Auto speed

 78, 64, 45 etc.

- 4. Temperature
 - 78, 64, 85 etc.
- 5. # Siblings
 - **0-2, 3-5, 6**+
- 6. Letter gradeA, B, C etc.
















Conclusion

- 1. Defined statistics
- 2. Described the uses of statistics
- Distinguished descriptive & inferential statistics
- 4. Defined population, sample, variable, parameter, & statistic
- 5. Explained the types of data
- 6. Compared the sources of data

Question Opt		Opt 2	Opt 3	Opt 4	Answer	
The word statistics is used as	singular word	a plural word	both singular	neither singular	both singular and	
			and plural words	nor plural word	plural words	
Classification is a process of arranging data in	grouping of	different rows	different	different	grouping of	
	related facts in		columns and	columns	related facts in	
	different		rows	grouping of	different classes	
	classes			related facts in		
				different classes		
To represent two or more sets of interrelated data,	bar diagram	pie diagram	histogram	multiple bar	multiple bar	
we use				diagram	diagram	
Histogram is a graph of	Time series	frequency	cumulative	normal	frequency	
		distribution	frequency	distribution	distribution	
			distribution			
Univariate data consists of	one variable	two variables	three variable	four	one variable	
Data are generally obtained from	Primary	Secondary	Both primary	neither primary	Both primary and	
	sources	sources	and secondary	nor secondary	secondary	
			sources	sources	sources	
In geographical classification data are classified on	area	attributes	time	location	area	
the basis of						
In qualitative classification data are classified on the	area	attributes	time	location	attributes	
basis of						
In quantitative classification data are classified on	area	attributes	time	magnitude	magnitude	
the basis of						
Number of source of data is	2	3	4	1	2	
Squares and rectangles are	Two	One	Three	Multi	Two dimensional	
	dimensional	dimensional	dimensional	dimensional	diagram	
	diagram	diagram	diagram	diagram		
Data originally collected for an investigation is	Tabulation	Primary data	Secondary data	Published data	Primary data	
known as						
The heading of a row in a statistical table is known	stub	caption	title	heading	stub	
as						

Statistics can	prove anything	disprove	neither prove	none of these	neither prove nor
		anything	nor disprove		disprove anything
			anything but it is		but it is just a
			just a tool		tool
Statistics is also a science of	estimates	both a and b	probabilities	neither a nor b	both a and b
Statistics is	quantitative	a qualitative	both quantitative	neither	both quantitative
	science	science	and qualitative	quantitative nor	and qualitative
			science	qualitative	science
Statistics considers	a single item	a set of item	either a single	neither a single	a set of item
			item or a set of	item or a set of	
			item	item	
Statistics can be considered as	an art	a science	both an art and	neither an art nor	both an art and
			science	a science	science
The other name of cumulative frequency curve is	Ogive	Bars	Histogram	Pie diagram	Ogive
Number of methods of collection of primary data is	2	3	4	5	5
Number of questions in a questionnaire should be	.5	10	maximum	minimum	minimum
Sources of secondary data are	Published	Unpublished	Either Published	primary source	Either Published
	sources	sources	sources or		sources or
			Unpublished		Unpublished
			sources		sources
Compared with primary data, secondary data are	more reliable	less reliable	equally reliable	uniformly reliable	less reliable
are column headings	stub	heading	bar	captions	captions
Mid value=	lower	upper	lower	lower boundary+	lower boundary+
	boundary/2	boundary/2	boundary+	upper boundary	upper
			upper		boundary)/2
			boundary)/2		
The origin of the word statistics has been traced to the Latin word	statista	status	statistik	statistique	status
Graphs of frequency distribution are	histogram	pie diagram	bar chart	circle	histogram

cubes are	Two	One	Three	Multi	Three	
	dimensional	dimensional	dimensional	dimensional	dimensional	
	diagram	diagram	diagram	diagram	diagram	
is the difference between the value	class interval	frequency	number of items	range	range	
of the smallest item and the valueof the largest item.						
is one which is used by the individual or	primary data	secondary data	both		primary data	
agency which collect it.						
Exclusive class intervals suit	discrete	continuous	both	neither	continuous	
	variables	variables			variables	
A table is a systematic arrangement of statistical data	columns	rows	both columns	stubs	both columns and	
in			and rows		rows	
The collected data in any statistical investigation are	raw data	arranged data	classified data	tabulated data	raw data	
known as						
The emitting form of a frequency polygon is called -	histogram	ogive	bar diagram	frequency curve	frequency curve	
In chronological classification data are classified on	time	attributes	class intervals	location	time	
the basis of						
Bar diagrams are dimensional diagrams	two	three	one	multi	one	
Diagrams and graphs are tools of	collection of	presentation	analysis	summarization	presentation	
	data					
In a two dimensional diagram	only height is	only width is	height,width and	Both height and	only height is	
	considered	considered	thickness are	width are	considered	
			considered	considered		
Which one of the following is a measure of central	Median	range	variation	correlation	Median	
The total of the values of the items divided by their	Madian	Arithmetic	mode		A mithematic macon	
number of items is known as	Median	mean	mode	range	Anumetic mean	
In the short-cut method of arithmetic mean, the	y A	A v	$(\mathbf{x}, \mathbf{A})/\mathbf{a}$	$(\Lambda x)/2$	x A	
deviation is taken as	$\mathbf{X} - \mathbf{A}$	A - X	(X - A) / C	$(\mathbf{A} - \mathbf{X}) / \mathbf{C}$	X - A	
The sum of the deviations of the values from their	1	0.000	two	7070	7070	
arithmetic mean is	- 1	one	two	Zero	Zero	
The formula for the weighted arithmetic mean is	$\sum wx / \sum w$	$\sum w / \sum wx$	$\sum x / n$	$\sum \mathbf{x} / \sum \mathbf{f}$	$\sum wx / \sum w$	
Find the Mean of the following values. 5, 15, 20, 10,	5	18	41	20	18	
Which of the followings represents median?	First quartile	Third quartile	Second quartile	Q.D	Second quartile	

Which of the measure of central tendency is not affected by extreme values?	Mode	Median	sixth deciles	Mean	Median	
Sum of square of the deviations about mean is	Maximum	one	zero	Minimum	Minimum	
Median is the value of item when all the items are in order of magnitude.	First	second	Middle most	last	Middle most	
Find the Median of the following data 160, 180, 175, 179, 164, 178, 171, 164, 176.	160	175	176	180	175	
The position of the median for an individual series is	(N + 1) / 2	(N + 2) / 2	N/2	N/4	(N + 1) / 2	
Mode is the value, which has	Average frequency density	less frequency density	greatest frequency density	graetest frequency	greatest frequency density	
A frequency distribution having two modes is said to	unimodal bimodal		trimodal	modal	bimodal	
Mode has stable than mean.	less	more	same	most	less	
Which of the following is not a measure of dispersion?	Range	quartile deviation	standard deviation	median	median	
Range of the given values is given by	L- S	L+S	S+L	LS	L-S	
Which one of the following is relative measure of dispersion?	Range	Q.D	S.D	coefficient of variation	coefficient of variation	
Coefficient of variation is defined as	(AM * 100)/S.D	(S.D* 100)/A.M	S.D/A.M	(1/S.D)*100	(S.D* 100)/A.M	
If the values of median and mean are 72 and 78 respectively, then find the mode.	16	60	70	76	60	
Find Mean for the following 3, 4, 5.	4	2.25	3	2.28	4	
The coefficient of range	L-S /L+S	L+S /L-S	L-S	L+S	L-S /L+S	
Second quartile is also called as	Mode	mean	median	G.M	median	
If A.M = 8, N=12, then find $\sum X$.	76	80	86	96	96	
If the value of mode and mean is 60 and 66 then, find the value of median.	64	46	54	44	64	
The formula for median for continuous series is	M = (N+1) / 2	M = L + [$(N/2 + cf) / f]$ * i	M =L - (N/2+cf)/f* i	M = L + [(N/2 - cf) / f] * i	M = L + [(N/2 - cf) / f] * i	

Median is	Average point	Midpoint	Most likely point	Most remote point	Midpoint	
Mode is the value which	Is a mid point	Occur the most	Average of all	Most remote Likely	Occur the most	
Is known as positional average	Median	Mean	Mode	Range	Median	
The median of marks 55, 60, 50, 40, 57, 45, 58, 65, 57, 48 of 10 students is	55	57	52.5	56	56	
The middle most value of a frequency distribution table is known as	Mean	Median	Mode	Range.	Median	
The middle most value of a frequency distribution table is known as	Mean	Median	Mode	Range	Median	
Measures of central tendency is also known as	Dispersion	averages	correlation	tendency	correlation	
From the given data 35,40,43,32,27 the coefficient	23	0.23	13	0.13	13	
If $S.D = 6$, then find variance.	6	36	42	12	36	
Which one of the following shows the relation between variance and standard deviation?	var = square root of S.D	S.D = square root of variance	variance = S.D	variance / S.D = 1	S.D = square root of variance	
If variance is 64, then find S.D.	8	13	14	11	8	
Which of the following measures of averages divide the observation into two parts	Mean	Median	Mode	Range	Median	
Which of the following measures of averages divide the observation into four equal parts	Mean	Median	Mode	Quartile	Quartile	
Arithmetic mean of the series 1, 3, 5, 7, 9 is	5	6	5.5	6.5	5	
Arithmetic mean of the series 3, 4, 5, 6, 7 is	5.5	6	5	6.5	5	
The Arithmetic mean for the series 3, 5, 5, 2, 6, 2, 9, 5, 8, 6, is	5	6	5.5	6.5	5	
The median value for the series 3, 5, 5, 2, 6, 2, 9, 5,	6	5	5.5	6.5	5	
The mode for the series 3, 5, 6, 2, 6, 2, 9, 5, 8, 6 is	5	6	5.5	6.5	6	
The Arithmetic mean for the series 51.6, 50.3, 48.9, 48.7, 48.5 is	49.8	50	48.9	49.6	49.8	
The Median for the series 51.6, 50.3, 48.9, 48.7, 49.5, is	49.8	50	48.9	49.6	49.6	

The Mode for the series 51.6, 50.3, 48.9, 48.7, 49.5 is	49.8	50	48.9	49.6	48.9
If standard deviation is 5, then the variance is	5	625	25	2.23068	25
Standard deviation is also called as	Root mean square deviation	mean square deviation	Root deviation	Root median square deviation	Root mean square deviation



KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics							
Subject : Biostatistics and Research Methodology Semester II LTPC							
Subject Code : 19MBP203	Class : I M.Sc Microbiology	4004					

UNIT II - Test of Significance

Sampling parameters: Difference between sample and Population, Censoring, difference between parametric and non-parametric statistics. Sampling Distributions, Standard Error, Testing of Hypothesis, Level of Significance and Degree of Freedom, Confidence Interval;, Small sample test based on t-test, Large Sample Test based on Normal Distribution - Z- test and F test.

Suggested Readings

- 1. Jerrold H. Zar. (2003). *Biostatistical Analysis*. (4th ed.). Pearson Education (P) Ltd, New Delhi.
- 2. Kothari. C.R. (2004). *Research Methodology Methods and Techniques*. (2nd ed.). New Age International Pvt. Ltd, New Delhi.

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 (H₀) 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)

= $P(H_0 \text{ is rejected when it is true})$

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 (H₀ is rejected when it is false) = 1- P (H₀ is accepted when it is false) = 1- P (Committing Type-II error) = 1- β

- 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 $x_1, x_2, x_3, ..., x_n$ 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 $\bar{x} \sim N(\mu, \sigma^2/n)$, Where \bar{x} be the sample mean

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

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

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

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 H₀

If $|Z| < Z_{\alpha}$, accept the null hypothesis H₀

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

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means 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
$$\overline{x}_1 \sim N\left(\mu_1, \sigma_1^2/n_1\right)$$
 and $\overline{x}_2 \sim N\left(\mu_2, \sigma_2^2/n_2\right)$
Therefore $\overline{x}_1 - \overline{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

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

Now calculate |Z|

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

If $\left| Z \right| \! > \! Z_{lpha}$, reject the null hypothesis H_{0}

If $|Z| < Z_{\alpha}$, accept the null hypothesis H₀

Note: If σ_1^2 and σ_2^2 are unknown then we can consider S_1^2 and S_2^2 as the estimate value of σ_1^2 and σ_2^2 respectively..

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

Let $x_1, x_2, x_3, \dots, x_n$ 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. $s \sim N(\sigma, \sigma^2/_{2n})$

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

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

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

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 H₀

If $|Z| < Z_{\alpha}$, accept the null hypothesis H₀

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 σ_1^2 and σ_2^2 respectively

Let $s_{\rm I}$ and $s_{\rm 2}\,{\rm be}$ 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 $\overline{x}_2 \sim N\left(\sigma_2, \sigma_2^2/2n_2\right)$
Therefore $s_1 - s_2 \sim N\left(\sigma_1 - \sigma_2, \frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0: \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$ against the two sided alternative $H_1: \sigma_1 \neq \sigma_2$

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

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

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

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \sim N(0,1)$$
[since $\sigma_1 - \sigma_2 = 0$ from H₀]

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 $|\mathbf{Z}| \! < \! \mathbf{Z}_{\alpha}$, accept the null hypothesis \mathbf{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 $E(\hat{p}_2) = p_2 \Longrightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

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

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

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

$$=\frac{\hat{p}_{1}-\hat{p}_{2}-E(\hat{p}_{1}-\hat{p}_{2})}{S.E(\hat{p}_{1}-\hat{p}_{2})}\sim N(0,1)$$

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

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

M.Jeganathan, Department of Mathematics, KAHE

2019 Batch

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

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 α % l.o.s i.e. Z_{α}

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

If $\left| Z \right| {\leq} Z_{lpha}$, accept the null hypothesis ${
m H}_0$

• As σ is unknown,

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

Step 2: If μ_0 falls into the above confidence intervals, then

do not reject ${m H}_0$. Otherwise, reject ${m H}_0$.

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)

$$H_0: \mu = \mu_0 = 19000$$
 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 $m{H}_0$.

(b)

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

Therefore, **not** reject H_0 .

2019 Batch

(c)

$$H_0: \mu = \mu_0 = 18900$$
 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 $\, lpha = 0.05$. Please test the hypothesis

$$H_0: u = 40 \text{ vs. } H_a: u \neq 40$$

based on

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

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

(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	State H_0 and H_a	$H_0: \mu \ge 16500$						
mathematically and		$H_a: \mu < 16500$						
verbally. Identify the null		$n = 14, \bar{x} = 15700, s = 1250$						
and alternative hypotheses								
Specify the level of	Specify α	$\alpha = 0.05$						
significance								
Identify the degrees of	d.f = n - 1	d.f. = 13						
freedom and sketch the								
sampling distribution								
Determine any critical	Table 5 (<i>t</i> -distribution) in	The test is left-tailed. Since						
values. If test is left tailed,	appendix B	test is left tailed and						
use One tail, $lpha$ column		d.f = 13 , the critical value						
with a negative sign. If test		is $t_0 = -1.771$						
is right tailed, use One tail,								
lpha column with a positive								
sign. If test is two tailed,								
use Two tails, $ lpha $ column								
with a negative and positive								
sign.								
Determine the rejection	The rejection region is	The rejection region is						
regions.	$t < t_0$	<i>t</i> < -1.771						

Find the standardized test statistic	$t = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} \approx \frac{\overline{x} - \mu}{s/\sqrt{n}}$	$t = \frac{15700 - 16500}{1250/\sqrt{14}} \approx -2.39$
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.

Possible Questions PART-B

- 1. How Z-test is used for testing significance of proportions? In a referendum submitted to the student body and 850 men and 550 women voted. Out of these, 530 of men and 310 of women voted 'yes'. Does this indicate a significant difference in opinion on matter between men and women students? (Use $\alpha = 5\%$ and $Z_{(0.05)} = 1.96$)
- 2. Test Median class size for Math is larger than the median class size for English for the following data using Mann Whitney U test.

Class size (Math, M)	23	45	34	78	34	66	62	95	81
Class size (English, E)	30	47	18	34	44	61	54	28	40

Hypothesis Testing

- Test of Hypothesis
- The null hypothesis
- The alternative hypothesis
- Steps in hypothesis testing
- Type-I & II errors
- One and Two Tailed Tests

What is a Hypothesis?

- A hypothesis is a claim (assumption) about the population parameter
 - Examples of parameters are population mean or proportion
 - The parameter must be identified before analysis

I claim the mean GPA of this class is $\mu = 3.5!$



Test of Hypothesis

- A hypothesis test allows us to draw conclusions or make decisions regarding population data from sample data.
- A hypothesis test allows us to draw conclusions or make decisions regarding difference between two population mean or sample mean.

The Null Hypothesis, H₀

- States the assumption (numerical) to be tested
 - e.g.: The average number of TV sets in U.S. Homes is at least three $(H_0: \mu \ge 3)$
- Is always about a population parameter (*H*₀ : µ ≥ 3), not about a sample statistic (*H*₀ : X̄ ≥ 3)
The Null Hypothesis, H₀

- Begins with the assumption that the null hypothesis is true
 - Similar to the notion of innocent until proven guilty
- Refers to the status quo
- Always contains the "=" sign
- May or may not be rejected



The Alternative Hypothesis, H₁

- Is the opposite of the null hypothesis
 - e.g.: The average number of TV sets in U.S. homes is less than 3 ($H_{\rm 1}$: μ < 3)
- Challenges the status quo
- Never contains the "=" sign
- May or may not be accepted
- Is generally the hypothesis that is believed (or needed to be proven) to be true by the researcher

Hypothesis Testing Process

Assume the population mean age is 50. ($H_0: \mu = 50$)

Identify the Population



Reason for Rejecting H₀

Sampling Distribution of \overline{X}

It is unlikely that we would get a sample mean of this value Therefore, we reject the null hypothesis that m = 50.

... if in fact this were the population mean.

20 $\mu = 50$ If H_0 is true

Level of Significance, α

- Defines unlikely values of sample statistic if null hypothesis is true
 - Called rejection region of the sampling distribution
- Is designated by α, (level of significance)
 Typical values are .01, .05, .10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

The significance level of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis H_0 , if it is in fact true.

It is the probability of a type I error and is set by the investigator in relation to the consequences of such an error. That is we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims.

The significance level is usually denoted by a

Significance Level = $P(type \ I \ error) = \alpha$ Usually the significance level is chosen to be a = 0.05 = 5%. Level of Significance and the Rejection Region



Errors in Making Decisions

Type-I error

Rejecting a null Hypothesis when it is True is called as a Type-I error. The probability of occurrence of type I error is denoted by a

Type-II error

Accepting a null Hypothesis when it is false is called as a Type-II error. The probability of occurrence of type II error is denoted by β

Type-I & II errors

- Type I Error
 - Rejects a true null hypothesis
 - Has serious consequences
 - The probability of Type I Error is $\,\alpha\,$
 - Called level of significance
 - Set by researcher
- Type II Error
 - Fails to reject a false null hypothesis
 - The probability of Type II Error is β
 - The power of the test is $(1-\beta)$

Result Probabilities

		<i>H</i> ₀ : Ini	nocent			
	Jury T	rial	Нуро	othesis	Test	
	The Truth			The Truth		
Verdict	Innocent	Guilty	Decision	H ₀ True	H ₀ False	
Innocent	Correct	Error	Do Not Reject H ₀	1 - α	Type II Error (β)	
Guilty	Error	Correct	Reject H ₀	Type I Error (α)	Power (1 - β)	

Type I & II Errors Have an Inverse Relationship

α

If you reduce the probability of one error, the other one increases so that everything else is unchanged.

The z-test for the Mean of a Normal Population

We want to test, μ , denote the mean of a normal population

The test statistic

$$z = \sqrt{n} \, \frac{\overline{x} - \mu_0}{\sigma} \approx \sqrt{n} \, \frac{\overline{x} - \mu_0}{s}$$

- Acceptance Region – Accept H_0 if: $-z_{\alpha/2} \le z \le z_{\alpha/2}$
- Critical Region

- Reject
$$H_0$$
 if: $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

With this Choice

 $P[\text{TypeI Error}] = P[\text{Reject } H_0 \text{ when true}]$ $= P[z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2}] = \alpha$



Sometimes the *Critical* Region is broken into two parts and lies in **both tails** of the sampling distribution of the test statistic (when H_0 is true).



When the *Critical* Region is broken into two parts and lies in both tails the test is called a **two-tailed test**



Sometimes the *Critical* Region is lies entirely in **one tail** of the sampling distribution of the test statistic (when H_0 is true).



When the *Critical* Region is lies entirely in **one** tail, the test is called a **one-tailed test**



The Critical Region

- The set of values of the test statistic that indicate H_A is true.
- Whether the test is one or two tailed depends on H_A

Usually If *H_A* is **two sided** than the test is **twotailed**.

If H_A is **one sided** than the test is **one-tailed**.

- 1. $H_A: \mu \neq \mu_0$ is **two sided** ($\mu < \mu_0$ or $\mu > \mu_0$)
 - 2. H_A : $\mu < \mu_0$ is one sided
 - 3. H_A : $\mu > \mu_0$ is **one sided**

Whether one uses a one or two tailed depends on the objectives of the researcher The alternative hypothesis, H_A , is the research hypothesis

Different researchers could choose different alternatives because their objectives are different

Steps in Hypothesis Testing

- I. Specify the Null Hypothesis (H_0) and the alternative Hypothesis (H_1)
- II. Select the significance level (z)
- III. Compute an appropriate test statistic

$$Z=rac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$$

where

- \overline{X} : sample Mean
- IV. Locate the Critical Region
- V. Statistical Conclusion

Example

In a manufacturing company a metal container that is suppose to weigh m = 42.0 Kilograms. A sample of n = 50 containers were selected and found to weigh an average of

 $\overline{x} = 42.3$ Kilograms

with a standard deviation of

s = 1.1 Kilograms

We want to test $H_0: \mu = 42.0$ Kilograms against

$$H_A$$
: $\mu \neq 42.0$ Kilograms

Test statistic
$$z = \sqrt{n} \left(\frac{\overline{x} - 42.0}{s} \right)$$

Using $\alpha = 0.05$, we would reject H_0 if $z < -z_{\alpha/2} = -1.960$ or $z > z_{\alpha/2} = 1.960$

Now

$$z = \sqrt{n} \left(\frac{\overline{x} - 42.0}{s}\right) = \sqrt{50} \left(\frac{42.3 - 42.0}{1.1}\right) = 1.928$$

Since

 $-1.960 \le z \le 1.960$

we would accept H_0

H_0 is accepted.



the **p-value** is defined to be the probability that the test statistic is **as or more extreme** than the observed value (assuming that the *Null Hypothesis* (H_0) is true) Now **p-value** = $P[z \le -1.928] + P[z \ge 1.928] = 0.054$



"Students" t-test

The Situation

- Let $x_1, x_2, x_3, \ldots, x_n$ denote a sample from a normal population with mean μ and standard deviation σ . Both μ and σ are unknown.
- Let

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \text{the sample mean}$$

$$\sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} = \text{the sample standard deviation}$$

• we want to test if the mean, μ , is equal to some given value μ_0 .

For small sample sizes:

The sampling distribution of



is called "students" t distribution with n-1 degrees of freedom

Properties of Student's t-distribution

- Similar to Standard normal distribution
 - Symmetric
 - unimodal
 - Centred at zero
- Larger spread about zero.
 - The reason for this is the increased variability introduced by replacing σ by *s*.
- As the sample size increases (degrees of freedom increases) the t distribution approaches the standard normal distribution

t distribution

standard normal distribution



The Alternative Hypothesis <i>H_A</i>	The Critical Region		
$H_A: \mu \neq \mu_0$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$		
$H_A: \mu > \mu_0$	$t > t_{\alpha}$		
$H_A: \mu < \mu_0$	$t < -t_{\alpha}$		

 t_{α} and $t_{\alpha/2}$ are critical values under the *t*-distribution with (n-1) degrees of freedom

Example

- Let x_1 , x_2 , x_3 , x_4 , x_5 , x_6 denote weight loss from a **new diet** for n = 6 cases.
- Assume that x_1 , x_2 , x_3 , x_4 , x_5 , x_6 is a sample from a normal population with mean μ and standard deviation σ . Both μ and σ are unknown.
- we want to test:

 $H_0: \mu \le 0$ New diet is not effective

versus $H_A: \mu > 0$ New diet is effective

The Test Statistic



The Critical region:

Reject if $t > t_{\alpha}$



1	2	3	4	5	6
2.0	1.0	1.4	-1.8	0.9	2.3

The summary statistics:

$\bar{x} = 0.96667$ and s = 1.462418

The Test Statistic



The Critical Region (using $\alpha = 0.05$)

Reject if $t > t_{0.05} = 2.015$ for 5 d.f.

<u>Conclusion:</u> Accept H_0 :
The z-test for Proportions

Testing the probability of success in a binomial experiment

Test of significance for Proportions

The Standard error of proportion is given by

$$\sigma_{\overline{p}} = \sqrt{\frac{\stackrel{P_{\rightarrow} \times q}{H_{0}}}{n}} \qquad \therefore Z = \frac{\stackrel{\overline{P} - P_{\rightarrow}}{H_{0}}}{\frac{\sigma_{-}}{p}}$$

where

- $P_{H_{u}}$: Hypothesized value of population proportion of success
- q_{H_0} : Hypothesized value of population proportion of failure
- \overline{P} : Sample Proportion of Success
- \overline{q} : Sample Proportion of failure
- n : Sample Size

Chi-Square Test

: expected frequency

$$\chi^2 = \sum \frac{\left(f_0 - f_e\right)^2}{f_e}$$

where,

 $f_{\scriptscriptstyle 0}$: observed frequency

 f_{e} : expected frequency

Chi-square test procedure

- I. Setting up of the null and the alternative hypothesis
- II. Listing up of the observed frequencies
- III. Calculating the expected frequencies if the data followed a given theoretical distribution
- IV. Obtaining the differences between the Observed and the corresponding expected Frequency. Adding all the fractions obtained above

- V. Comparing the value with the appropriate (χ 2) value from the tables (at the specified level of significance)
- V. Accepting the Null Hypothesis if the value thus computed for the given degree of freedom and the level of significance is lesser than the critical value otherwise reject the hypothesis

Chi-square as a test of goodness of fit.

Situation

- A success-failure experiment has been repeated n times
- The probability of success π is unknown. We want to test
 - $-H_0$: $\pi = \pi_0$ (some specified value of π)

Against

 $\pi \neq \pi_0$

The Data

- The success-failure experiment has been repeated n times
- The number of successes *x* is observed.

$$\hat{\pi} = \frac{x}{n} =$$
 the poportion of successes

 Obviously if this proportion is close to π₀ the Null Hypothesis should be accepted otherwise the null Hypothesis should be rejected.

The Test Statistic

 To decide to accept or reject the Null Hypothesis (H₀) we will use the test statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0 \left(1 - \pi_0\right)}{n}}}$$

- If *H*₀ is true we should expect the test statistic *z* to be close to zero.
- If H_0 is true we should expect the test statistic *z* to have a standard normal distribution.
- If *H_A* is true we should expect the test statistic *z* to be different from zero.

- Acceptance Region – Accept H_0 if: $-z_{\alpha/2} \le z \le z_{\alpha/2}$
 - Critical Region - Reject H_0 if: $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
 - With this Choice

 $P[\text{TypeI Error}] = P[\text{Reject } H_0 \text{ when true}]$ $= P[z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2}] = \alpha$

Example

- In the last election the proportion of the voters who voted for the Liberal party was 0.08 (8 %)
 - The party is interested in determining if that percentage has changed
 - A sample of n = 800 voters are surveyed

We want to test $-H_0$: $\pi = 0.08$ (8%)

> Against $-H_A$: $\pi \neq 0.08$ (8%)

The Test

- 1. Decide on $\alpha = P[Type \ I \ Error] = the significance level of the test Choose (<math>\alpha = 0.05$)
- 2. Collect the data
- The number in the sample that support the liberal party is x = 92

$$\hat{\pi} = \frac{x}{n} = \frac{92}{800} = 0.115 \ (11.5\%)$$

3. Compute the test statistic



- 4. Make the Decision $z_{\alpha/2} = z_{0.025} = 1.960$
 - Accept H_0 if: $-1.960 \le z \le 1.960$
 - Reject H_0 if: z < -1.960 or z > 1.960

Since the test statistic is in the Critical region we decide to Reject H_0

Conclude that H_0 : $\pi = 0.08$ (8%) is false

There is a significant difference ($\alpha = 5\%$) in the proportion of the voters supporting the liberal party in this election than in the last election

Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
The regression line cut each other at the	Average of X only	Average of Y	Average of X	the median of X	Average of X and
point of		only	and Y	on Y	Y
Given the coefficient of correlation being	0.98	0.64	0.66	0.54	0.64
0.8, the coefficient of determination will be					
0.9, the coefficient of determination will be	0.98	0.81	0.66	0.54	0.81
If the coefficient of determination being 0.49 , what is the coefficient of correlation	0.7	0.8	0.9	0.6	0.7
Given the coefficient of determination being					
0.36, the coefficient of correlation will be	0.3	0.4	0.6	0.5	0.6
Which one of the following refers the term Correlation?	Relationship between two values	Relationship between two variables	Average relationship between two variables	Relationship between two things	Relationship between two variables
If $r = +1$, then the relationship between the given two variables is	perfectly positive	perfectly negative	no correlation	high positive	perfectly positive
If $r = -1$, then the relationship between the given two variables is	perfectly positive	perfectly negative	no correlation	low Positive	perfectly negative
If $r = 0$, then the relationship between the given two variables is	Perfectly positive	perfectly negative	no correlation	both positive and negative	no correlation
Coefficient of correlation value lies between	1 and –1	0 and 1	0 and ∞	0 and –1.	1 and -1
While drawing a scatter diagram if all					
points appear to form a straight line getting Downward from left to right, then it is inferred that there is	Perfect positive correlation	simple positive correlation	Perfect negative correlation	no correlation	Perfect negative correlation
The range of the rank correlation coefficient is	0 to 1	-1 to 1	0 to ∞	$-\infty$ to ∞	-1 to 1
If $r = 1$, then the angle between two lines of regression is	Zero degree	sixty degree	ninety degree	thirty degree	ninety degree

Regression coefficient is independent of	Origin	scale	both origin and scale	neither origin nor scale.	Origin
If the correlation coefficient between two variables X and Y is negative, then the Regression coefficient of Y on X is	Positive	negative	not certain	zero	negative
If the correlation coefficient between two variables X and Y is positive, then the Regression coefficient of X on Y is	Positive	negative	not certain	zero	Positive
There will be only one regression line in case of two variables if	r =0	r = +1	r = -1	r is either +1 or -1	r =0
The regression line cut each other at the point of	Average of X only	Average of Y only	Average of X and Y	the median of X on Y	Average of X and Y
If b_{xy} and b_{yx} represent regression coefficients and if $b_{yx} > 1$ then b_{xy} is	Less than one	greater than one	equal to one	equal to zero	Less than one
Rank correlation was discovered by	R.A.Fisher	Sir Francis Galton	Karl Pearson	Spearman	Spearman
Formula for Rank correlation is	1- ($6\Sigma d^2 /(n(n2-1)))$	1- ($6\Sigma d^2 / (n(n2+1)))$	$1+ (6\Sigma d^2 / (n(n2+1)))$	1 /(n(n2-1))	1- ($6\Sigma d^2 / (n(n2-1)))$
With $b_{xy}=0.5$, $r = 0.8$ and the variance of Y=16, the standard deviation of X=	6.4	2.5	10	25.6	2.5
The coefficient of correlation r =	$(b_{xy.} b_{yx})^{1/4}$	$(b_{xy}, b_{yx})^{-1/2}$	$(b_{xy}, b_{yx})^{1/3}$	$(b_{xy.} b_{yx})^{1/2}$	$(b_{xy.} b_{yx})^{1/2}$
If two regression coefficients are positive then the coefficient of correlation must be	Zero	negative	positive	one	positive
If two-regression coefficients are negative then the coefficient of correlation must be	Positive	negative	zero	one	Positive
The regression equation of X on Y is	X = a + bY	X = a + bX	X= a - bY	Y = a + bX	X = a + bY
The regression equation of Y on X is	X = a + bY	$X = \overline{a + bX}$	X= a - bY	Y = a + bX	Y=a+bX
The given two variables are perfectly positive, if	r = +1	r = -1	$\mathbf{r} = 0$	$r \neq +1$	r = +1
The relationship between two variables by plotting the values on a chart, known as-	coefficient of correlation	Scatter diagram	Correlogram	rank correlation	Scatter diagram

If x and y are independent variables then,	$cov(x,y) \neq 0$	cov(x,y)=1	cov(x,y)=0	cov(x,y) > 1	cov(x,y)=0	
Correlation coefficient is the of	Mode	Geometric mean	Arithmetic mean	median	Geometric mean	
the two regression coefficients.	111040				Geometrie mean	
$b_{xy} = 0.4, b_{yx} = 0.9$ then r =	0.6	0.3	0.1	-0.6	0.6	
$b_{xy}=1/5$, r=8/15, s _x = 5 then s _y =	40/13	13/40	40/3	3	40/3	
The geometric mean of the two regression	Correlation	regression	coefficient of	coefficient of	Correlation	
coefficients.	coefficient	coefficients	range	variation	coefficient	
If two variables are uncorrelated, then the	De met emiet		Parallel to each	perpendicular to	perpendicular to	
lines of regression	Do not exist	coincide	other	each other	each other	
If the given two variables are correlated	a 1			<i>m</i> / ±1	. 1	
perfectly negative, then	r = +1	r = -1	$\mathbf{r} = 0$	$I \neq \pm I$	r = -1	
If the given two variables have no		. 1			0	
correlation, then	r = +1	r = -1	$\mathbf{r} = 0$	$r \neq \pm 1$	$\mathbf{r} = 0$	
If the correlation coefficient between two						
variables X and Y is, the Regression	Needin	positive	not certain	zero	positive	
coefficient of Y on X is positive	Negative					
If the correlation coefficient between two						
variables X and Y is, the Regression	Negative	positive	not certain	zero	Negotine	
coefficient of Y on X is negative	Negative				Negative	
is independent of origin and	Correlation	regression	coefficient of	coefficient of	Correlation	
scale.	coefficient	coefficients	range	variation	coefficient	
The angle between two lines of regression is	r – 2	r = 0	r _ 1	r — 1	r _ 1	
ninety degree, if	1 - 2	I = 0	1 – 1	1 – -1	1 - 1	
is used to measure closeness of	Degracion	maan	Dank correlation	actualition	correlation	
relationship between variables.	Regression	mean	Kalik correlation	correlation	correlation	
If r is either $+1$ or -1 , then there will be only						
one line in case of two variables	Correlation	regression	rank correlation	mean	regression	
When $b_{xy} = 0.85$ and $b_{yx} = 0.89$, then	0.09	0.5	0.69	0.97	0.97	
correlation coefficient r =	0.98	0.3	0.08	0.07	0.07	

If b_{xy} and b_{yx} represent regression coefficients and if $b_{xy} < 1$, then b_{yx} is	less than 1	greater than one	equal to one	equal to zero	greater than one
While drawing a scatter diagram if all points appear to form a straight line getting Downward from left to right, then it is inferred that there is	Perfect positive correlation	simple positive correlation	Perfect negative correlation	no correlation	Perfect negative correlation
If r =1, the angle between two lines of regression is	Zero degree	sixty degree	ninety degree	thirty degree	ninety degree
Regression coefficient is independent of	Origin	scale	both origin and scale	neither origin nor scale.	Origin
There will be only one regression line in case of two variables if	r =0	r = +1	r = -1	r is either +1 or -1	r =0



KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics						
Subject : Biostatistics and Research Me	thodology Semester II	LTPC				
Subject Code : 19MBP203	Class : I M.Sc Microbiology	4004				

UNIT III Analysis of Variance

Distribution-free test - Chi-square test; Basic Introduction to Multivariate statistics, etc. Test of significance: Tests based on Means only-Both Large sample and Small sample tests - Chi square test - goodness of fit. Analysis of Variance – one way and two way classification, CRD, RBD Designs.

Suggested Readings

- 1. Jerrold H. Zar. (2003). *Biostatistical Analysis*. (4th ed.). Pearson Education (P) Ltd, New Delhi.
- 2. Kothari. C.R. (2004). *Research Methodology Methods and Techniques*. (2nd ed.). New Age International Pvt. Ltd, New Delhi.

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 t	he Chi-Square Goodness-	of-Fit Test (p 496)
Procedure	Equations	Example (p 497)
Identify the claim. State the	State $H_{_0}$ and $H_{_1}$	H_0 :
null and alternative		Classical 4%
hypothesis.		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	$\varphi_{0.01}^2(d.f=5) = 15.086$
	6 Appendix B	
Identify the rejection region	$\chi^2 \ge \chi^2_{\alpha}$	$\chi^2 \ge 15.086$
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 $P(X \ge 22.713) < 0.01$ 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
Рор	75	0.18	90
Rock	125	0.29	145

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

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

Next add up the values in C5 and the sum is the test statistic as follows: Click on Calc->Column

Statistics. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

Sum of C5

Sum of C5 = 22.7132

Next calculate the P-value: Click on Calc->Probability Distributions->Chi-square. Select

Cumulative Probability and enter 5 Degrees of Freedom Enter the value of the test statistic

22.7132 for the Input Constant. Click OK.

The following is displayed on the Session Window.

Cumulative Distribution Function

Chi-Square with 5 DF x P(X <= x)

22.7132 0.999617

P(X≤22.7132) = 0.999617 So the P-value = 1 − 0.999617 = 0.000383. This is less that α =

0.01 so we reject the null hypothesis.

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

Chi-Square with M&M's

<i>H</i> ₀ : Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: $\alpha = 0.05$
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^{2}_{0.05}(d.f.=5) = 11.071$
Rejection Region: $\chi^2 \ge 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where <i>O</i> 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_i.

	A ₁	A ₂	A ₃	A ₄	Α
B ₁	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B ₂	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B ₃	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
В	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

If A and B are independent, we'd expect

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

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

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	State H_0 and H_1	H_0 : CEO's ages are		
hun and alternative		independent of company		
nypotneses.		size		
		H_1 : CEO's ages are		
		dependent on company size.		
Specify the level of	Specify α	$\alpha = 0.01$		
significance				
Determine the degrees of	d.f. = (r-1)(c-1)	d.f. = (2-1)(5-1) = 4		
freedom				

2019 Batch

Find the critical value.	χ^2_{α} : Obtain from Table 6,	$\chi^2_{\alpha} \ge 13.277$
	Appendix B	
Identify the rejection region	$\chi^2 \geq \chi^2_{lpha}$	$\chi^2 \ge 13.277$
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 $P(X \ge 77.0) < \alpha$ 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		CEO's ages and company
context of the original claim		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) H₀: $\pi = 0.5$ and H_A: π^{-1} 0.5
- 2) We will use the *Z*-distribution

- 3) We will use the 5%-level, thus $\alpha = 0.05$
- 4) The test statistic is $z = (0.25 0.5) / \sqrt{0.25 / 20} = -2.24$
- 5) Table A-4 shows that $P(|Z| > 2.24) \gg 0.025$.
- 6) Because PROB-VALUE < α , we reject H₀. 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 $P(X \ ^3 \ 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$ $P(X \ ^3 \ 8) = 0.1208 + 0.0537 + 0.0161 + 0.0029 + 0.002 = 0.1937$ Adopting the 5% uncertainty level, we see that PROB-VALUE > α . Thus we fail to reject H₀. We cannot conclude students prefer Fontanez.

- Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs. *Solution:*
 - (a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is

0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

- (b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference. We conclude the samples were drawn from different populations.
- 6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

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

Solution:

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

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

Using the χ^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 (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 $2P(X \ge 8)$ 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	10
	(39)	(21)
No Insurance	15	25
	(26)	(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:

	Percentage		Percentage		Percentage
Day	of sunshine	Day	of sunshine	Day	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

Unit III	Analysis of Variance					Batch
	6 84	4 16	90	26	90	
	7 90	0 17	6	27	100	
	8 1	8 18	0	28	90	
	9 9	0 19	22	29	58	
1	0 10	0 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

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	$S(x_i)$	$F(x_i) / S($	(x_i) - $F(x_i)/$
4.2	0.020	0.015	0.005
4.3	0.040	0.023	0.017

0.028	0.032	0.060	4.4
0.088	0.692	0.780	5.9
0.000	0.960	0.960	6.7
0.008	0.972	0.980	6.8
0.019	0.981	1.000	6.9

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	\mathbf{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.

Possible Questions PART-B

1. According to the IQ level and the economic conditions of their homes 1000 students at a college were graded. Use χ^2 test to find out whether there is any association between economic

were graded. Use λ test to find out whether there is any association between economic condition at home and IQ.

Economic	IC	Total	
Conditions	High	Low	Total
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

(Note: The level of significance is 0.05 and table value is 3.84).

2. Test Median class size for Math is larger than the median class size for English for the following data using Mann – Whitney U test.

U									
Class size (Math, M)	23	45	34	78	34	66	62	95	81
Class size (English, E)	30	47	18	34	44	61	54	28	40

3. Mr. Gowtham, Personal Manager is concerned about absenteeism. He decides to sample the records to determine if absenteeism is distributed evenly throughout the six-day work-week. The null hypothesis to be tested is: absenteeism is distributed evenly throughout the week. The sample results are as follows:

Day	Number of Absentees
Monday	12
Tuesday	9
Wednesday	11
Thursday	10
Friday	9
Saturday	9

i. Using χ^2 test of significance, compute χ^2 value.

- ii. Is the null hypothesis rejected?
- iii. Specifically, what does this indicate to the Personal Manager? (*Note: The level of significance is 0.01 and table value is 15.086*).

Analysis of Variance

Contents

- Primary Scales of Measurement
- Measures Independent vs. Repeated
- Stacked Data vs. Un stacked Data
- Deciding the Statistical Tool
- Exploring the Data
 - Mean Plot
 - Standard Deviation Plot
 - Box Plot

Contents

- ANOVA
- One-way ANOVA
- Two-way ANOVA
| Scale | Basic
Characteristics | Common
Examples | Marketing
Examples | Permissible statistics | | |
|---------|--|--|--|------------------------|---|--|
| | | | Examples | Descriptive | Inferential | |
| Nominal | Numbers identify
and Classify
objects | Social Security
numbers,
numbering of
football
players | Brand
numbers,
store types,
gender
classification | Percentages,
mode | Chi-square,
binomial
test | |
| Ordinal | Numbers indicate
the relative
positions of the
objects but not
the magnitude of
differences
between them | Quality
rankings,
rankings of
teams in a
tournament | Preference
rankings,
market
position,
social class | Percentile,
median | Rank-order
correlation,
Friedman
ANOVA | |

Scale	Basic Characteristics	Common Examples	Marketing Examples	Permissible statistics		
				Descriptive	Inferential	
Interval	Differences between objects can be compared; zero point is arbitrary	Temperature (Fahrenheit, Centigrade)	Attitudes, opinions, index numbers	Range, mean, standard deviation	Product- moment correlation, t-tests, ANOVA, regression, factor analysis	
Ratio		Length, weight	Age, income, costs, sales, market shares	Geometric mean, harmonic mean	Coefficient of variation	

Mean Plot

- Purpose: Detect changes in location between group.
- Mean plots are formed by:
 - Vertical axis: Group mean
 - Horizontal axis: Group identifier
- A reference line is plotted at the overall mean.

Sample Plot

Mean Plot



Mean Rank

DRAFT69.DAT

Questions

- The mean plot can be used to answer the following questions.
- Are there any shifts in location?
- What is the magnitude of the shifts in location?
- Is there a distinct pattern in the shifts in location?

Standard Deviation Plot

- Purpose: Detect Changes in Scale Between Groups
- Standard deviation plots are formed by:
 Vertical axis: Group standard deviations
 Horizontal axis: Group identifier
- A reference line is plotted at the overall standard deviation.

Sample Plot



Questions

- The standard deviation plot can be used to answer the following questions.
- Are there any shifts in variation?
- What is the magnitude of the shifts in variation?
- Is there a distinct pattern in the shifts in variation?

Box Plot

- Purpose: Check location and variation shifts
- The box plot is an important EDA tool for determining if a factor has a significant effect on the response with respect to either location or variation.
- Box plots are formed by
 - Vertical axis: Response variable
 - Horizontal axis: The factor of interest

Sample Plot

 This box plot reveals that machine has a significant effect on energy with respect to location and possibly variation



Questions

- The box plot can provide answers to the following questions:
- Is a factor significant?
- Does the location differ between subgroups?
- Does the variation differ between subgroups?
- Are there any outliers?

The one-way ANOVA is useful when we want to compare the effect of multiple levels of one factor and we have multiple observations at each level.

The factor can be either discrete (different machine, different plants, different shifts, etc.) or continuous (different gas flows, temperatures, etc.).

- ANOVA allows us to compare the effects of multiple levels of multiple factors .
- ANOVA splits the data into components .

- Suppose we have a turning operation in a machine shop where we are turning pins to a diameter of .125 +/- .005 inches.
- Throughout the course of a day we take five samples of pins and obtain the following measurements: .125, .127, .124, .126, .128.

We can split these data values into a common value (mean) and residuals (what's left over) as follows: ••.125 .127 .124 .126 .128

*****.126 .126 .126 .126 .126

+

✤-.001 .001 -.002 .000 .002

- From these data, also called overlays, we can easily calculate the location and spread of the data as follows:
- ▶ mean = .126
- \succ std. deviation = .0016.

One-Way ANOVA - Data

- Suppose we have a turning operation in a machine shop where we are turning pins to a diameter of .125 +/- .005 inches.
- we take five random samples from each machine to obtain the following diameter data:

Machine								
1	2	3	4	5				
.125	.118	.123	.126	.118				
.127	.122	.125	.128	.129				
.125	.120	.125	.126	.127				
.126	.124	.124	.127	.120				
.128	.119	.126	.129	.121				

Hypothesis Formulation

- H_o : There is no statistically significant difference in the pin diameters across the machines on which they were manufactured.
- H₁: There is statistically significant difference in the pin diameters across the machines on which they were manufactured.

One-Way ANOVA – Results

• Using ANOVA software we summarize the data into an ANOVA table as follows:

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-value
Factor Levels	0.00013744	4	3.436E-05	5.21> 2.87
Residuals	0.000132	20	6.6E-06	
Corrected Total	0.00026944	24		

One-Way ANOVA - Interpretation

- By dividing the Factor-level mean square by the residual mean square, we obtain a F-value of 5.21 which is greater than the cut-off value of 2.87 for the F-distribution at 4 and 20 degrees of freedom and 95% confidence.
- Therefore, there is sufficient evidence to reject the null hypothesis that the levels are all the same.

One-Way ANOVA - Conclusion

 From the analysis of these data we can conclude that the factor "machine" has an effect. There is a statistically significant difference in the pin diameters across the machines on which they were manufactured.

Two-Way ANOVA - Example

 Let's extend the one-way machining example by assuming that we want to test if there are any differences in pin diameters due to different types of coolant. We still have five different machines making the same part and we take five samples from each machine for each coolant type to obtain the following data:

Two-Way ANOVA - Data

	Machine						
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>		
	.125	.118	.123	.126	.118		
Coolant A	.127	.122	.125	.128	.129		
Coolant A	.125	.120	.125	.126	.127		
	.126	.124	.124	.127	.120		
	.128	.119	.126	.129	.121		
	.124	.116	.122	.126	.125		
	.128	.125	.121	.129	.123		
Coolant B	.127	.119	.124	.125	.114		
	.126	.125	.126	.130	.124		
	.129	.120	.125	.124	.117		

Hypothesis Formulation

- H_o : There is no significant main effect of the machine on the pin diameters.
- H₁ : There is significant main effect of the machine on the pin diameters.
- H_o : There is no significant main effect of the coolant on the pin diameters.
- H₁ : There is significant main effect of the coolant on the pin diameters.

Hypothesis Formulation

- H_o : There is no significant interaction effect of the machine and coolant.
- H₁ : There is significant interaction effect of the machine and coolant.

Two-Way ANOVA - Results

ANOVA								
Source of Variation	SS	df	MS	F	P-value	F crit		
Sample-Coolant	3.92E-06	1	3.92E-06	0.4532	0.504700	4.08475		
Columns-Machine	0.000303	4	7.583E-05	8.7665	3.5235E-05	2.60598		
Interaction	1.468E-05	4	3.67E-06	0.4243	0.79017	2.60598		
Within	0.000346	40	0.000009					
Total	0.000668	49						

Two-Way ANOVA - Conclusion

 From the ANOVA table we can conclude that machine is the most important factor and is statistically significant. Coolant is not significant and neither is the interaction. These results would lead us to believe that some tool-matching efforts would be useful for improving this process.

Types of ANOVA

- One-way Independent Measure.
- One-way Repeated Measure.

Factorial ANOVA

- Two-way Independent Measure.
- Two-way Repeated Measure.
- Mixed Design.
- Latin Square.

Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
Completely randomized design is similar					
to	three way	one way	two way	t test	one way
Randomized block design is similar to					
	two way	three way	one way	many	two way
ANOVA is the technique of analysis of					
	standard deviation	variance	mean	range	variance
Under one way classification, the					
influence of only attribute or factor					
is considered	two	three	one	many	one
Under two way classification, the					
influence of only attribute or factor					
is considered	four	two	three	one	two
The word is used to					
indicate various statistical measures like					
mean, standard deviation, correlation etc,					
in the universe.	Statistic	parameter	hypothesis	none of these	parameter
The term STATISTIC refers to the					
statistical measures relating to the					
	Population	hypothesis	sample	universe	sample
A hypothesis may be classified as					
	Simple	Composite	null	all the above	all the above
Level of significance is the probability of			Not committing	any of the	
	Type I error	Type II error	error	above	any of the above
Degrees of freedom are related to				No. of	
	No. of		No. of independent	dependent	
	observations in a	hypothesis	observations in a	observations in	No. of independent
	set	under test	set	a set	observations in a set
A critical function provides the basis for -			no decision about		
	Accepting H ₀	rejecting H ₀	H_0	all the above	all the above

Student's t-test is applicable in case of		for sample of			
		size between		none of the	
	Small samples	5 and 30	Large samples	above	Small samples
Student's t-test is applicable only when	The variate values	the variable is			
	are	distributed	The sample is not		
	independent	normally	large	all the above	all the above
If the calculated value is less than the					
table value then we accept the					
-					
hypothesis.	Alternative	null	both	sample	null
Small sample test is also known as					
	Exact test	t – test	normal test	F-test	t – test
The formula for c^2 is	å(O–E) ² /E	å(E+O) ² /E	å(О-Е) /Е	å(О-Е) ² /О	å(O–E) ² /E
If a statistic 't' follows student's t				c^2 distribution	
distribution with n degrees of freedom	c^2 distribution with	c^2 distribution	c ² distribution with	with $(n+1)$	c^2 distribution with
then t ² follows	(n-1) degrees of	with n degrees	n^2	degrees of	(n-1) degrees of
	freedom	of freedom	degrees of freedom	freedom	freedom
The distribution used to test goodness of					
fit is	F distribution	c ² distribution	t distribution	Z distribution	c ² distribution
Degree of freedom for statistic chi-					
square incase of contingency table of					
order 2x2 is	3	4	2	1	1
Larger group from which the sample is					
drawn is called	Sample	sampling	universe	parameter	universe
Any hypothesis concerning a population				statistical	
is called a	Sample	population	statistical measure	hypothesis	statistical hypothesis
Rejecting Ho when it is true leads					
	Type I error	Type II error	correct decision	either (a) or (b)	Type I error
Accept Ho when it is true leads					
	Type I error	Type II error	correct decision	either (a) or (b)	correct decision
Type II error occurs only if	Reject Ho when it	Accept Ho	Accept Ho when it	reject Ho when	Accept Ho when it
	is true	when it is false	is true	it is false	is false

The correct decision is	Reject Ho when it	Accept Ho	Reject Ho when it		Reject Ho when it is
	is true	when it is false	is false	none of these	false
The maximum probability of committing					
type I error, which we specified in a test					
is		alternative		level of	
known as	Null hypothesis	hypothesis	DOF	significance	level of significance
If the computed value is less than the		Null	Alternative		
critical value, then	Null hypothesis is	hypothesis is	hypothesis is		Null hypothesis is
	accepted	rejected	accepted	population	accepted
If the computed value is greater than the		Null	Alternative		
critical value, then	Null hypothesis is	hypothesis is	hypothesis is		Null hypothesis is
	accepted	rejected	accepted	small sample	rejected
In sampling distribution the standard					
error is	np	pq	npq	sqrt(npq)	sqrt(npq)
If the sample size is greater than 30, then					
the sample is called	Large sample	small sample	population	Null hypothesis	Large sample
If the sample size is less than 30, then				alternative	
the sample is called	Large sample	small sample	population	hypothesis	small sample
Z - test is applicable only when the					
sample size is	zero	one	small	large	large
The degrees of freedom for two samples					
in t – test is	$n_1 + n_2 + 1$	$n_1 + n_2 - 2$	$n_1 + n_2 + 2$	$n_1 + n_2 - 1$	$n_1 + n_2 - 2$
An assumption of t – test is population					
of the sample is	Binomial	Poisson	normal	exponential	normal
The degrees of freedom of chi – square					
test is	(r-1)(c-1)	(r+1)(c+1)	(r+1)(c-1)	(r-1)(c+1)	(r-1)(c-1)
In chi – square test, if the values of					
expected frequency are less than 5, then					
they are					
combined together with the neighbouring					
frequencies. This is known as					
	Goodness of fit	DOF	LOS	pooling	pooling

The expected frequency of chi – square		(RT - CT) /			
test can be calculated as	(RT + CT) / GT	GT	(RT * CT) / GT	(RT*CT)	(RT * CT) / GT
In F – test, the variance of population					
from which samples are drawn are					
	equal	not equal	small	large	equal
If the data is given in the form of a series					
of variables, then the DOF is					
	n	n-1	n+1	(r-1)(c-1)	n-1
The characteristic of the chi-square test				independence	independence of
is	DOF	LOS	ANOVA	of attributes	attributes
If $S_1^2 > S_2^2$, then the F – statistic is					
	$\mathbf{S}_1 / \mathbf{S}_2$	$\mathbf{S}_2 / \mathbf{S}_1$	${\bf S_1}^2 / {\bf S_2}^2$	S_1^{3} / S_2^{3}	${\bf S_1}^2 / {\bf S_2}^2$
The value of Z test at 5% level of					
significance is	3.96	2.96	1.96	0.96	1.96
In, the variance of population from					
which samples are drawn are equal					
	t-test	Chi-Square test	Z-test	F-test	F-test
F – statistics is		Variance			
		within the	Variance between	Variance within	
	Variance between	samples /	the rows /	the rows /	Variance between
	the samples /	variance	variance between	variance within	the samples /
	variance within the	between the	the columns	the columns	variance within the
	samples	samples			samples
Analysis of variance utilizes:					
	t-test	Chi-Square test	Z-test	F-test	F-test
F – test whish is also known as	Chi-Square test	Z-test	varience ratio test	t-test	varience ratio test
The technique of analysis of variance					
refered to as	ANOVA	F – test	Z – test	Chi- square test	ANOVA
The two variations, variation within the					
samples and variations between the					
samples					

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

The Research Process



INFORMATION

- REDUCES UNCERTAINTY
- HELPS FOCUS DECISION MAKING

TYPES OF RESEARCH

- EXPLORATORY
- DESCRIPTIVE
- CAUSAL



UNCERTAINTY INFLUENCES THE TYPE OF RESEARCH


Degree of Problem Definition

Exploratory Research	Descriptive Research	Causal Research	
(Unaware of Problem)	(Aware of Problem)	(Problem Clearly Defined)	
" " " " " " " " " " " " " " " " " " "	"What kind of people are buying our product? Who buys our competitor's product?" "What features do buyers prefer in our product?"	"Will buyers purchase more of our products in a new package?" "Which of two advertising campaigns is more effective?") "

EXPLORATORY RESEARCH

Secondary Data

Experience Survey

Pilot Studies (Focus Groups)

EXPLORATORY RESEARCH

- INITIAL RESEARCH CONDUCTED TO CLARIFY AND DEFINE THE NATURE OF A PROBLEM
- DOES NOT PROVIDE CONCLUSIVE EVIDENCE
- SUBSEQUENT RESEARCH EXPECTED

DESCRIPTIVE RESEARCH

- DESCRIBES CHARACTERISTICS OF A
 POPULATION OR PHENOMENON
- SOME UNDERSTANDING OF THE NATURE OF THE PROBLEM

DESCRIPTIVE RESEARCH EXAMPLE

Weight Watchers average customer:

- Woman about 40 years old
- Household income of about \$50,000
- At least some college education
- Trying to juggle children and a job

DESCRIPTIVE RESEARCH EXAMPLE

Men's fragrance market:

- 1/3 size of women's fragrance market
- But growing at a faster pace
- Women buy 80 % of men's fragrances

CAUSAL RESEARCH

CONDUCTED TO IDENTIFY CAUSE AND EFFECT RELATIONSHIPS



IDENTIFYING CAUSALITY

A causal relationship is very difficult to prove.

Evidence of causality:
1. The appropriate causal order of events
2. Concomitant variation--two phenomena vary together
3. An absence of alternative plausible explanations

STAGES IN THE RESEARCH PROCESS

- PROBLEM DISCOVERY AND PROBLEM DEFINITION
- RESEARCH DESIGN
- SAMPLING
- DATA GATHERING
- DATA PROCESSING AND ANALYSIS
- CONCLUSIONS AND REPORT

The Research Process



The Research Process (cont.)



RESEARCH STAGES

- Cyclical process conclusions generate new ideas
- Stages can overlap chronologically
- Stages are functionally interrelated
 - Forward linkages
 - Backward linkages





PROBLEM DISCOVERY AND DEFINITION

- FIRST STEP
- PROBLEM, OPPORTUNITY, OR MONITOR OPERATIONS
- DISCOVERY BEFORE DEFINITION

State the research questions and research objectives

HYPOTHESIS

 A STATEMENT THAT CAN BE REFUTED BY EMPIRICAL DATA



EXPLORATORY RESEARCH TECHNIQUES - TWO EXAMPLES

- SECONDARY DATA (HISTORICAL DATA)
 - Previously Collected
 - Census of Population
 - Literature Survey
- PILOT STUDY

A Number of Diverse Techniques

FOCUS GROUP INTERVIEW





Group dynamics and synergy.
Small samples.
Not just open ended but discussion and analysis.

RESEARCH DESIGN

- MASTER PLAN
- FRAMEWORK FOR ACTION
- SPECIFIES METHODS AND PROCEDURES

BASIC RESEARCH METHODS

Surveys

Experiments

Secondary Data Studies

Observations Studies

Forms of Survey Research





Causal relationships.
Lab or field.
Experimental design, causation and control of extraneous factors.

Observation Research.

Personal - Mystery shoppers, One-way mirrors.
Mechanical - Traffic counters.
Electrical - e.g., video carts, people meters.

[↓]Scanners.

Selecting a Sample

Sample: subset of a larger population.

SAMPLE

POPULATION

SAMPLING

- WHO IS TO BE SAMPLED
- HOW LARGE A SAMPLE
- HOW WILL SAMPLE UNITS BE SELECTED
- PROBABILITY OR NON-PROBABILITY.

DATA GATHERING STAGE



DATA PROCESSING AND ANALYSIS



CONCLUSIONS AND REPORT WRITING



• EFFECTIVE COMMUNICATION OF THE RESEARCH FINDINGS

Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
Study to portray accurately characteristics					
of a particular individual, situation or a					
group is called research				Hypothesis	
	Exploratory	Diagnostic	Descriptive	testing	Descriptive
Critical evaluation made by the					
researcher with the facts and information					
already available is called				Hypothesis	
research.	Analytical	Exploratory	Diagnostic	testing	Analytical
Research to find reason, why people					
think or do certain things is an example of	Quantitative		Qualitative	Fundamental	Qualitative
	Research	Applied Research	research	research	research
Which one is considered a major					
component of the research study	Interpretation	research report	finding	draft	research report
Research task remains incomplete till the				objective and	
has been presented.	Report	objective	finding	finding	Report
What is the last step in a research study	Writing report	writing finding	limitations	research report	Writing report
Which is the final step in report				writing	
writing	Writing report	writing finding	writing drafts	limitations	writing drafts
What is usually appended to the research					
work	Editing	bibliography	coding	research report	bibliography
The is one which gives				writing	
emphasis on simplicity and attractiveness	popular report	research report	article report	limitations	popular report
should slow originality					
and should necessarily be on attempt to					
solve some intellectual problem	Interpretation	research report	finding	draft	research report
The researcher must remain caution about					
the that can possibly arise in					
the process of interpreting results	Analysis	conclusions	findings	error	error
Which one should be considered while			both validity and		
interpreting a given data	Validity	reliability	reliability	technical jargon	reliability

is asking questions face to		mailed		personal	personal
face	Indirect method	questionnaire	through post	interview	interview
Journals, books, magazines etc are useful					
sources of collecting			both primary and		
	Primary data	secondary data	secondary data	objective	secondary data
The collected raw data to detect errors and					
are called,					
	Editing	coding	classification	all the above	Editing
The formal, systematic and intensive					
process of carrying on a scientific method					
of	Research			research	
analysis is	Design	research	interpretation	analysis	research
Refers to the process of assigning					
numerals or symbols to answers of					
response	Coding	editing	classification	all the above	Coding
The research study, which is based on					
describing the characteristic of a					
particular	Experience				
individual or group	survey	Descriptive	Diagnostic	Exploratory	Descriptive
Research is a	Finding	assumption	statement	all the above	all the above
The research, which has the purpose of		1			
improving a product or a process testing					
theoretical concepts in actual problem					
situations isresearch.	Statistical	Applied	Domestic	Biological	Applied
The chart of research process indicates					
that the process consists of a number of	Closely related	unrelated	Closely unrelated	moderately	Closely related
	activities	activities	activities	related activities	activities

The objective of a good design is	Maximize the	Minimize the	Minimize the	Maximize the	Maximize the
	bias	bias and	bias and	bias and	bias and
	andmaximize	minimize	maximize	maximize	maximize
	the reliability of				
	data	data	data	data	data
A is used whenever a full written					
report of the study is required.	Popular report	Technical report	article	monograph	Technical report
The is one which gives					
emphasis on simplicity and attractiveness.					
	Popular report	Technical report	article	monograph	Popular report
Study to portray accurately characteristics					
of a particular individual, situation or a					
group is called research				Hypothesis	
	Exploratory	Diagnostic	Descriptive	testing	Descriptive
Critical evaluation made by the					
researcher with the facts and information					
already available is called				Hypothesis	
research.	Analytical	Exploratory	Diagnostic	testing	Analytical
Research to find reason, why people					
think or do certain things is an example of	Quantitative		Qualitative	Fundamental	Qualitative
	Research	Applied Research	research	research	research
Which one is considered a major					
component of the research study	Interpretation	research report	finding	draft	research report
Research task remains incomplete till the				objective and	
has been presented.	Report	objective	finding	finding	Report
What is the last step in a research study	Writing report	writing finding	limitations	research report	Writing report
Which is the final step in report				writing	
writing	Writing report	writing finding	writing drafts	limitations	writing drafts
What is usually appended to the research					
work	Editing	bibliography	coding	research report	bibliography
The is one which gives				writing	
emphasis on simplicity and attractiveness	popular report	research report	article report	limitations	popular report

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

The chart of research process indicates					
that the process consists of a number of	Closely related	unrelated	Closely unrelated	moderately	Closely related
	activities	activities	activities	related activities	activities
The objective of a good design is	Maximize the	Minimize the	Minimize the	Maximize the	Maximize the
	bias	bias and	bias and	bias and	bias and
	andmaximize	minimize	maximize	maximize	maximize
	the reliability of	the reliability of	the reliability of	the reliability of	the reliability of
	data	data	data	data	data
A is used whenever a full written					
report of the study is required.	Popular report	Technical report	article	monograph	Technical report
The is one which gives					
emphasis on simplicity and attractiveness.					
	Popular report	Technical report	article	monograph	Popular report
The square of the S.D is	Variance	Coefficient of variation	Square of variance	Square of coefficient of variation	Variance
Analysis of variance is a statistical method of comparing the of several populations.	Standard deviations	Means	Variances	Proportions	Means
The analysis of variance is a statistical test that is used to compare how many group means?	Three	More than three	Three or more	Two or more	Two or more
Analysis of variance utilizes:	F-test	Chi-Square test	Z-test	t-test	F-test
What is two-way ANOVA?	An ANOVA with two variables and one factor	An ANOVA with one variable and two factors	An ANOVA with one variable and three factors	An ANOVA with both categorical and scale variables	An ANOVA with one variable and two factors
Which of the following is the correct F ratio in the one-way ANOVA?	MSA/MSE	MSBL/MSE	MST/MSE	MSE/MST	MST/MSE
For validity of F-test in Anova, parent population should be	Binomial	Poisson	Normal	Exponential	Normal

sum of squares measures the variability of the observed values around their respective tabulated values	Treatment	Error	Interaction	Total	Error
The sum of squares measures the variability of the sample treatment means around the overall mean.	Total	Treatment	Error	Interaction	Treatment
If the true means of the k populations are equal, then MST/MSE should be:	more than 1.00	Close to 1.00	Close to -1.00	A negative value between 0 and - 1	Close to 1.00
If MSE of ANOVA for six treatment groups is known, you can compute	Degree of freedom	The standard deviation of each treatment group	Variance	The pooled standard deviation	The pooled standard deviation
To determine whether the test statistic of ANOVA is statistically significant,to determine critical value we need	Sample size, number of groups	Mean, sample standard deviation	Expected frequency, obtained frequency	MSTR, MSE	Sample size, number of groups
Which of the following is an assumption of one-way ANOVA comparing samples from 3 or more experimental treatments?	Variables follow F- distribution	Variables follow normal distribution	Samples are dependent each other	Variables have different variances	Variables follow normal distribution
The error deviations within the SSE statistic measure distances:	Within groups	Between groups	Between each value and the grand mean	Betweeen samples	Within groups
In one-way ANOVA, which of the following is used within the <i>F</i> -ratio as a measurement of the variance of individual observations?	SSTR	MSTR	SSE	MSE	SSE
When conducting a one-way ANOVA, the the between-treatment variability is when compared to the within-treatment variability	More random larger	Smalller	Larger	More random smaller	Smaller
When conducting a one-way ANOVA, the value of F DATA will be tend to be	More random larger	Smalller	More random smaller	Larger	Smaller
--	--	---	---	---	---
When conducting an ANOVA, F DATA will always fall within what range?	Between negative infinity and infinity	Between 0 and 1	Between 0 and infinity	Between 1 and infinity	Between 0 and infinity
If F DATA = 5, the result is statistically significant	Always	Sometimes	Never	Is impossible	Sometimes
If F DATA= 0.9, the result is statistically significant	Always	Sometimes	Never	Is impossible	Never
When comparing three treatments in a one- way ANOVA ,the alternate hypothesis is	All three treatments have different effect on the mean response.	Exactly two of the three treatments have the same effect on the mean response.	At least two treatments are different from each other in terms of their effect on the mean response	All the treatments have same effect	At least two treatments are different from each other in terms of their effect on the mean response
If the sample means for each of <i>k</i> treatment groups were identical,the observed value of the ANOVA test statistic?	1	0	A value between 0.0 and 1.0	A negative value	0
If the null hypothesis is rejected, the probability of obtaining a F - ratio > the value in the F table as the 95th % is:	0.5	>0.5	<0.5	1	<0.5
ANOVA was used to test the outcomes of three drug treatments. Each drug was given to 20 individuals. If MSE =16, What is the standard deviation for all 60 individuals sampled for this study?	6.928	48	16	4	4
Analysis of variance technique originated in the field of	Agriculture	Industry	Biology	Genetics	Agriculture

With 90, 35, 25 as TSS, SSR and SSC , in case of two way classification, SSE is	50	40	30	20	30
Variation between classes or variation due to different basis of classification is commonly known as	Treatments	Total sum of squares	Sum of squares	Sum of squares due to error	Treatments
The total variation in observations in Anova is classified as:	Treatments and inherent variation	SSE and SST	MSE and MST	TSS and SSE	Treatments and inherent variation
In Anova, variance ratio is given by	MST/MSE	MSE/MST	SSE/SST	TSS/SSE	MST/MSE
Degree of freedom for TSS is	N-1	k-1	h-1	(k-1)(h-1)	N-1
For Anova, MST stands for	Mean sum of squares of treatment	Mean sum of squares of varieties	Mean sum of squares of tables	Mean sum of sources of treatment	Mean sum of squares of treatment
An ANOVA procedure is applied to data of 4 samples, where each sample contains 10 observations. Then degree of freedom for critical value of F are	4 numerator and 9 denominator	3 numerator and 40 denominator	3 numerator and 36 denominator	4 numerator and 10 denominator	3 numerator and 36 denominator
The power function of a test is denoted by	M(w,Q)	M(Q,Qo)	P(w,Q)	P(w,Qo)	M(w,Q)
Sum of power function and operation characteristic is	Unity	Zero	two	Negative	Unity
Operation characteristic is denoted by	L(w,Q)	M(w,Q)	L(w,Qo)	M(w,Qo)	L(w,Q)
Operation characteristic is also known as	Test characteristic	Power function	best characteristic	unique characteristic	Test characteristic
The formula to find OC is L(w,Q)=	1-Power Function	2xPower Function	Power Funtion -1	2xConfidance Interval	1-Power Function
Operation Characteristic is of a test is related to	Power Function	Best Test	Unique Test	Uniformally Best Test	Power Function

If the Hypothesis is correct the operation charectristics will be	1	0	-1	0.5	1
If the Hypothesis is wrong the operation charectristics will be	0	1	0.5	0.333333	0
In which test we verify a null hypothesis against any other definite alternate hypothesis?	Best Test	Unique Test	Uniformally Best Test	Unbiased Test	Best Test
A Best Test is a Test such that the critical region for which attains least value for a given α .	Beta	1-Beta	Alpha	1-Alpha	1-Beta
A Test whose power function attains its mean at point $Q = Qo$ is called Test	ttains its ITest Unique Unbiased Power Operation Characteristic		Unique		
A Best Unique Test exist	Always	Never	Sometimes	When Q not = to Qo	Sometimes
Operation Characteristic is related to	Power Function	Unique Test	Best Test	Uniformally Best Test	Power Function
Power is the ability to detect:	A statistically significant effect where one exists	A psychologically important effect where one exists	Both (a) and (b) above	Design flaws	A statistically significant effect where one exists
Calculating how much of the total variance is due to error and the experimental manipulation is called:	Calculating the variance	Partitioning the variance	Producing the variance	Summarizing the variance	Partitioning the variance
ANOVA is useful for:	Teasing out the individual effects of factors on an Independent Variables	Analyzing data from research with more than one Independent Variable and one Dependent Variable	Analyzing correlational data	Individual effects of factors on an Dependent Variables	Analyzing data from research with more than one Independent Variable and one Dependent Variable

What is the definition of a simple effect?	The effect of one variable on another	The difference between two conditions of one Independent Variable at one level of another Independent Variable	The easiest way to get a significant result	Difference between two Dependent Variables	The difference between two conditions of one Independent Variable at one level of another Independent Variable
In a study with gender as the manipulated variable, the Independent Variable is:	Within participants	Correlational	Between participants	Regressional	Between participants
Which of the following statements are true of experiments?	The Independent Variable is manipulated by the experimenter	The Dependent Variable is assumed to be dependent upon the IV	They are difficult to conduct	both (a) and (b)	both (a) and (b)
All other things being equal, repeated- measures designs:	Have exactly the same power as independent designs	Are often less powerful than independent designs	Are often more powerful than independent designs	Are rarely less powerful when compare to than independent designs	Are often more powerful than independent designs
Professor P. Nutt is examining the differences between the scores of three groups of participants. If the groups show homogeneity of variance, this means that the variances for the groups:	Are similar	Are dissimilar	Are exactly the same	Are enormously different	Are similar
Differences between groups, which result from our experimental manipulation, are called:	Individual differences	Treatment effects	Experiment error	Within- participants effects	Treatment effects

Herr Hazelnuss is thinking about whether he should use a related or unrelated design for one of his studies. As usual, there are advantages and disadvantages to both. He has four conditions. If, in a related design, he uses 10 participants, how many would he need for an unrelated design?	40	20	10	100	40
Individual differences within each group	Treatment	Between-	Within-	Individual biases	Within-
of participants are called:	effects	participants error	participants error	marviadar biases	participants error
Calculating how much of the total variance is due to error and the experimental manipulation is called:	Calculating the variance	Partitioning the variance	Producing the variance	Summarizing the variance	Partitioning the variance
The decision on how many factors to keep	Statistical	Theoretical	Both (a) and (b)	Neither (a) nor	Both (a) and (b)
is decided on:	criteria	criteria		(b)	
It is possible to extract:	As many factors as variables	More factors than variables	More variables than factors	Correlation between the actual and predicted variables	As many factors as variables
Four groups have the following means on the covariate: 35, 42, 28, 65. What is the grand mean?	43.5	42.5	56.7	58.9	42.5
You can perform ANCOVA on:	Two groups	Three groups	Four groups	All of the above	All of the above
When carrying out a pretestposttest study, researchers often wish to:	Partial out the effect of the dependent variable	Partial out the effect of the pretest	Reduce the correlation between the pretest and posttest scores	Correlation between the two tests scores	Partial out the effect of the pretest

Using difference scores in a pretest posttest design does not partial out the effect of the pretest for the following reason:	The pretest scores are not normally correlated with the posttest scores	The pretest scores are normally correlated with the different scores	The posttest scores are normally correlated with the different scores	Up normal relationship with the different scores	The pretest scores are normally correlated with the different scores
Experimental designs are characterized by:	Two conditions	No control condition	Random allocation of participants to conditions	More than two conditions	Random allocation of participants to conditions
Between-participants designs can be:	Either quasi- experimental or experimental	Only experimental	Only quasi- experimental	Only correlational	Either quasi- experimental or experimental
A continuous variable can be described as:	Able to take only certain discrete values within a range of scores	Able to take any value within a range of scores	Being made up of categories	Being made up of variables	Able to take any value within a range of scores
In a within-participants design with two conditions, if you do not use counterbalancing of the conditions then your study is likely to suffer from:	Order effects	Effects of time of day	Lack of participants	Effects of participants	Order effects
Demand effects are possible confounding variables where:	Participants behave in the way they think the experimenter wants them to behave	Participants perform poorly because they are tired or bored	Participants perform well because they have practiced the experimental task	Participants perform strongly	Participants behave in the way they think the experimenter wants them to behave

Power can be calculated by a knowledge of:	The statistical test, the type of design and the effect size	The statistical test, the criterion significance level and the effect size	The criterion significance level, the effect size and the type of design	The criterion significance level, the effect size and the sample size	The criterion significance level, the effect size and the sample size
Relative to large effect sizes, small effect	Engine to detect	Harder to detect	As apprete datast	As difficult to	As difficult to
sizes are:	Easier to detect Harder to detec		As easy to detect	detect	detect
Differences between groups, which result from our experimental manipulation, are called:	Individual differences	Treatment effects	Experiment error	Within- participants effects	Treatment effects
Completely randomized design is similar to	three way	one way	two way	t test	one way
Randomized block design is similar to	two way	three way	one way	many	two way
ANOVA is the technique of analysis of	standard deviation	variance	mean	range	variance
Under one way classification , the influence of only attribute or factor is considered	two	three	one	many	one
Under two way classification , the influence of only attribute or factor is considered	four	two	three	one	two



M.JEGANATHAN

M.Sc (Stat)., MBA (HR)., M.Sc (App. Psy)., PGDCA.,



M.Jeganathan Department of Mathematics Karpagam Academy of Higher Education Coimbatore

"There is no way to get experience except through experience."



What Is Statistics?

- 1. The Science of data
- 2. Involves:
 - Collecting
 - Classifying
 - Summarizing
 - Organizing
 - Analyzing
 - Interpreting

What Is Statistics?

- **1.** The science of data
- 2. Involves:
 - Collecting
 - Classifying
 - Summarizing
 - Organizing
 - Analyzing
 - Interpreting

Understanding

Purposes



What Is Statistics?

1. The science of data Why? 2. **Involves:** Understanding Collecting Classifying © 1984-1994 T/Maker Co. **Purposes** Summarizing **Decision-**Organizing Making Analyzing Mana Interpreting











Data Type Examples

Quantitative Interval • What is the temperature? ____ (° Celsius) Ratio How tall are you? (Inches) Qualitative Nominal Do you own savings bonds? ____ Yes ___ No Ordinal • How do you rate this (excellent, good, bad)?

Thinking Challenge

Are they quantitative or qualitative?

Gender

 Male, female

 Weight

 123, 140.2 etc.

 Auto speed

 78, 64, 45 etc.

- 4. Temperature
 - 78, 64, 85 etc.
- 5. # Siblings
 - **0-2, 3-5, 6**+
- 6. Letter gradeA, B, C etc.



Measurement Scales

Nominal Ordinal Interval Ratio

Scale	Basic	Evamplas	Permissible	statistics
	Characteristics	Examples	Descriptive	Inferential
Nominal	Numbers identify and Classify objects	Gender classification, Social Security numbers, numbering of football players Brand numbers, store types	Percentages, Mode	Chi- Square, Binomial Test
Ordinal	Numbers indicate the relative positions of the objects but not the magnitude of differences between them	Quality rankings, rankings of teams in a tournament Preference rankings, market position, social class	Percentile, Median	Rank- order Correlatio n, Friedman ANOVA

Scale	Basic	Evamples	Permissible Statistics		
State	Characteristics		Descriptive	Inferential	
Interval	Differences between objects can be compared; zero point is arbitrary	Temperature (Fahrenheit, Centigrade) Attitudes, Opinions, Index Numbers	Range, Mean, Standard Deviation	Product- moment correlation, t- tests, ANOVA, Regression, Factor Analysis	
Ratio	The characteristic s of the nominal, ordinal and interval scales. In addition ratio scales have a true zero.	Weight and Distance (Length), Age, Income, Costs, Sales, Market Shares	Geometric Mean, Harmonic Mean	Coefficient of Variation	

Sampling

When we undertake a survey, to collect data, we are effectively sampling from a population. It is therefore necessary to define the population and the sampling method (of which there are many).



Characteristics of a good sample.

A sample should have all the characteristics of the population from where it is taken.

Researcher or Manager should not be biased in selecting a sample from the population.

The findings or decisions taken based on the sample study should be applicable to the entire population.

Types of sample

Sample may be either a probability sample or a non-probability sample

Probability sample: Each person in population has an equal, or known, chance of being selected

Non-probability sample: Some people in population have a greater, or unknown, chance of being selected



Probability Sampling

 Items in the sample are chosen based on known probabilities



Simple Random Samples

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement
 (unrestricted) or without replacement
- Samples obtained from table of random numbers or computer random number generators

Systematic Samples

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group
- Select every kth individual thereafter



Systematic Sampling

- Systematic sampling is a special case of cluster sampling, with only one cluster selected
 - This makes it hard to estimate sampling variances
 - Need prior knowledge or assumptions about response patterns

Stratified Samples

- Divide population into two or more subgroups (called strata) according to some common characteristic
- A simple random sample is selected from each subgroup, with sample sizes proportional to strata sizes
- Samples from subgroups are combined into one



Cluster Samples

- Population is divided into several "clusters," each representative of the population
- A simple random sample of clusters is selected
 - All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique



Cluster Sampling

- Typically, face-to-face household surveys involve interviewing several people in each area
- This is an example of a cluster sample, where the areas are the clusters
- This approach is much less costly than an SRS of the same size
- However it will also exhibit higher sampling variability, due to correlations between interviews within a cluster
 - E.g. similar spending patterns due to similar incomes, or a similar range of products being available locally

Difference Between Cluster and Stratified Sampling



Population of *L* strata, stratum *l* contains n_l units



Take simple random sample in *every* stratum



Population of C clusters



Take SRS of clusters, sample every unit in chosen clusters

Advantages and Disadvantages

- Simple random sample and systematic sample
 - Simple to use
 - May not be a good representation of the population's underlying characteristics
- Stratified sample
 - Ensures representation of individuals across the entire population
- Cluster sample
 - More cost effective
 - Less efficient (need larger sample to acquire the same level of precision)

Non-random or Judgment Sampling

Any sampling procedure that does not ensure some non-zero probability for each element in the population to be included in the sample would belong to the category of non-probability sampling.

Judgment Sampling Convenience Sampling Sequential Sampling

Judgment Sampling

The person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population.

It is a <u>nonprobability</u> sampling technique.

<u>Example</u>: A reporter might sample three or four senators, judging them as reflecting the general opinion of the senate.
Judgment Sampling

<u>Advantage</u>: It is a relatively easy way of selecting a sample.

<u>Disadvantage</u>: The quality of the sample results depends on the judgment of the person selecting the sample.

Convenience Sampling

It is a <u>non-probability sampling technique</u>. Items are included in the sample without known probabilities of being selected.

The sample is identified primarily by <u>convenience</u>.

<u>Example</u>: A professor conducting research might use student volunteers to constitute a sample.

Convenience Sampling

<u>Advantage</u>: Sample selection and data collection are relatively easy.

<u>Disadvantage</u>: It is impossible to determine how representative of the population the sample is.

Sequential Sampling

It is a <u>non-probability sampling technique</u>. In this the size of the sample is not fixed in advance but it is decided as the sampling process takes place depending on the results of the first sample.

A number of sample lots are drawn in sequence one after another from the population depending one the results of the previous sample.

This sampling method is used for Statistical Quality Control.

Sequential Sampling

<u>Example</u>: Quality of the inventory- A manager draws a lot from the inventory And tests it for acceptability. If it is acceptable, there will be no further samples required but if it is found unacceptable, the entire stock will be rejected.

<u>Advantage</u>: Sample selection and data collection are relatively easy.

<u>Disadvantage</u>: It is impossible to determine how representative of the population the sample is.

Which sample design do I use?

- 1. A SRS is most suitable for a homogeneous population *e.g. sample of workers in a particular factory*.
- 2. Stratified sampling is most suitable for a heterogeneous population where event of interest varies for different groups.
- 3. Cluster sampling is most suitable when sampling over a large geographical area or where a sample frame of elements is not available.
- 4. Stratified sampling may be feasible for telephone/postal surveys. Cluster/Quota are more appropriate for face to face interviews.

- 5. Stratified sample are most appropriate if accuracy is of particular importance, while quota sampling is not recommended in this case. If you can make do with less accurate estimates, cluster or quota sampling may suffice.
- 6. Cluster/quota sampling may be cheaper and do not require such a detailed sampling frame.

Sources of error when doing sample surveys



- Sample estimates will always differ from the true population figure because of:
- Sampling error: random variation involved when sampling if a random sample this can be predicted
- Non-sampling error: errors arising from all other aspects of survey design e.g. bias in sample (e.g. incomplete sampling frame), measurement errors during fieldwork

Sampling errors

Sampling errors or variations among sample statistics are due to differences between each sample & the population and among several samples. They originate at the time of collecting samples.

Reasons for Sampling errors

- Errors in the selection of the sample
- Bias in reporting of data
- Diversity of population
- •Substitute of sampling unit for convenience
- Faulty demarcation of the sampling universe

Effect of sample size on SE



Non- Sampling errors

Non-sampling errors occur at the time of observation, approximation & processing of data. This error is common to both the sampling & census survey.

The non-sampling error may be due to faulty sampling plan, lack of trained & qualified investigators, inaccuracy in responses collected due to bias of the respondent or the researcher, errors in the design of the survey &finally the errors in compilation or publication.

Reasons for non-sampling errors

Faulty sample plan Lack of trained and Qualified Investigators Inaccuracy in responses Errors in design of survey Errors in publication of data

Sampling distribution Standard error

The standard error (SE) of a statistic, is an estimate of the standard deviation of the statistic based on the data from one or more random samples.

The SE of the mean is used in the computation of confidence intervals & significance tests for the mean.

It is given by
$$\sigma_M = \sqrt{\frac{\sigma^2}{n}}$$

Deciding Which Statistical Test to Use

Deciding the Statistical Test_With Query

And **Good Luck!**

Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
Psychometric Methods book is written	J.P.Guilford	Likert	L.L.Thurstone	Louis Guttman	J.P.Guilford
by					
Respondents are asked to rank their	Comparative scaling	arbitrary scaling	rating scale	differential	Comparative
choices in				scale	scaling
is developed on ad-hoc basis	Differential scale	arbitrary scale	rating scale	ranking scale	arbitrary scale
scale is developed by	Comparative scale	likert scale	differential	rating scale	likert scale
utilizing item analysis approach			scale		
Scalogram analysis is developed	J.P.Guilford	Likert	L.L.Thurstone	Louis Guttman	Louis Guttman
by					
A complete enumeration of all items in	sampling unit	sample design	census inquiry	all the above	census inquiry
the population is known as					
The selected respondents	population	sample	sample size	population size	sample
constitute		1			1
The selection process of respondents is	survey	sampling	sample survey	census inquiry	sampling
called		technique			technique
The survey conducted to select the		sample survey	census inquiry	population size	sample survey
respondents is called	sampling technique				
A sample design is a definite plan for	universe	sample design	population	sample survey	population
obtaining a sample from a					
given					
The number of items in universe can	finite	infinite	both	zero	both
be					
The population of a city, number of	infinite	finite	both	zero	finite
workers in a company					
is					
Source list is also known as	sampling size	sampling size	sampling frame	population size	sampling frame
The size of the sample should be	large	optimum	small	all the above	optimum

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

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

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

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

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

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