M. A. Part - I

SEMESTER - I (CBCS)

PSYCHOLOGY PAPER - III

(CORE COURSE)

STATISTICS FOR PSYCHOLOGY

SUBJECT CODE: PAPSY103

Prof. (Dr.) D. T. Shirke
Offg. Vice Chancellor
University of Mumbai, Mumbai

Prin. Dr. Ajay Bhamare
Offg. Pro Vice-Chancellor,
University of Mumbai

Prof. Prakash Mahanwar
Director,
IDOL, University of Mumbai

Programme Co-ordinator : **Anil R. Bankar**
Head, Faculty Head, Arts and Humanities,
IDOL, University of Mumbai

Course Co-ordinator : **Dr. Naresh Tambe**
Assistant Professor (Psychology),
IDOL, University of Mumbai

Editor : **Dr. Anita Kumar**
Associate Professor (Retired), Visiting Faculty
Acharya and Marathe College, Chembur, Mumbai

Course Writers : **Darshana Sunil Kulkarni**
Assistant Professor,
Maniben Nanavati Women's College,
SNDT University, Vile Pale, Mumbai

: **Mrs. Dinika Sahil More**
Kishinchand Chellaram College,
K. M. Kundnai Chowk, Mumbai

: **Dr. Anita Kumar**
Associate Professor (Retired), Visiting Faculty
Acharya and Marathe College, Chembur, Mumbai

February 2023, Print - 1

Published by : Director,
Institute of Distance and Open Learning,
University of Mumbai,
Vidyanagari,Mumbai - 400 098.

DTP composed and Printed by: Mumbai University Press

CONTENTS

M. A. PART - I

SEMESTER - I (CBCS)
PSYCHOLOGY PAPER - III (CORE COURSE)

STATISTICS FOR PSYCHOLOGY

(CORE COURSE: 4 CREDITS, 15 WEEKS)

SYLLABUS

Objectives:

1. To introduce fundamental concepts about statistical application to psychology
2. To help learners to understand applications of statistics and learn numerical methods associated with them
3. To introduce multivariate methods and computer applications to statistics
4. To be able to use R for all statistical methods taught in the course.

Unit 1. Preliminary Concepts

a. Probability: axioms, random variables, expected value, central limit theorem
b. Distributions: discrete distributions- binomial, poisson; continues distributions: normal, t, F, chi-square, jointly distributed random variables.
c. Inference: estimation theory, statistical hypothesis testing, types of errors. Properties of estimators, methods of estimation: least square, maximum likelihhod. Bayesian inference. CLT; LLN; Cramér–Rao inequality; Rao Blackwell Theorem
d. Descriptive statistics: central tendency and variability, power and effect size. Testing for normality and outliers.

Unit 2. Inferential statistics: inference about location

a. Two group differences: *t* test- independent and dependent samples. Bootstraping.
b. Multi-group differences: one-way ANOVA: independent and dependent samples. two-way ANOVA: independent samples
c. Wilcoxon sign-rank test; median test; U test; Kruskal-Wallis test
d. MANOVA and discriminant function analysis

Unit 3. Association, prediction and other methods

a. Correlation: product moment, partial correlation, special correlations.
b. Linear regression (OLS)
c. Nonparametric correlations: Kendall's tau; Spearman's rho; measures for nominal data, chi square, binomial test, proportions test.
d. Multiple regression, logistic regression.

Unit 4. Factor Analysis and Software Packages

a. Factor analysis: basic concepts, methods of extraction and methods of rotation
b. Confirmatory factor analysis.
c. Structural Equations Modeling.
d. R: syntax, data management, Descriptive; graphs; basic and multivariate statistics in R; R GUI, other software.

Books for Study:

1. Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.
2. Wilcox R. R. (2009). Basic Statistics: Understanding Conventional Methods and Modern Insights. NY: OUP.
3. Minium, E. W., King, B. M., & Bear, G. (2001). Statistical reasoning in psychology and education. Singapore: John-Wiley.
4. Aron & Aron (2008). Statistics for Psychology (5th ed). New Delhi: Pearson

Books for Reference:

1. Daniel, W. W. (1995). Biostatistics. (6th Ed.). N.Y.: John Wiely.
2. Field, A., Miles, J., and Field, Z. (2012). Discovering Statistics Using R. NY: Sage.
3. Gourch, R. L. (1983). Factor Analysis. Lorrence Erlbaum
4. Guilford, J. P., & Fructore, B. (1978). Fundamental statistics for psychology and education. N.Y.: McGraw-Hill.
5. Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Mulivariate data analysis. (5th Ed.). N.J.: Prentice-Hall Inc.
6. Hatekar, N. R. (2009). Principles of Econometrics: An Introduction (Using R). ND: Sage.
7. Loehlin, J. (1998). Latent Variable Models: an introduction to factor, path, and structural analysis. Hillsdale, N.J.: LEA.
8. Marcoulides, A. G. & Schumacker, E. R. (2001). New developments and techniques in structural equation modeling. Hilsdel, New Jersey: Lawrence Erlbaum.
9. R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria:R Foundation for Statistical Computing. (http://www.R-project.org)
10. Sheskin, D. (2011).Handbook of Parametric and Nonparametric Statistical Procedures, (5th ed). Chapman and Hall/CRC.
11. Tabachnick, B. G. & Fidell, L. S. (2001). Using multivariate statistics (4th Ed.). Boston: Allyn and Bacon.
12. Wilcox, R. R. (1996). Statistics for social sciences. San Diego: Academic Press.
13. Wilcox, R. R. (2011). Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction. CRC Press.

Evaluation:
Internal evaluation: 25 marks
Semester end examination: 75 marks

1

PRELIMINARY CONCEPTS –I

Unit Structure :

1.0 OBJECTIVES:

After studying this unit a student will be able to:

- Understand preliminary concepts in statistics
- Understand basics of probability and solve related sums
- Understand what is central limit theorem
- Understand different probability distributions, their formulae and application

1.1 PROBABILITY

In the field of humanities, especially Psychology, research has obtained at most importance. What is setting Psychology apart from common sense is that Psychological theories are backed up by empirical support. Empirical investigation requires unbiased data collection and statistical analysis in order to generalize the results obtained from the data. The analysis that is done on sample is known as statistics whereas analysis done on population is known as parameter. A researcher's aim is to estimate parameters based on statistics. Since the researcher does not have access to the population but can only "estimate" its value from sample, the phenomenon of probability comes into the picture. In simple words, statistics helps you infer what is the 'most probable' population parameter given the current sample data.

Probability refers to the possibility or chance of an event occurring. It is the chance that a specific event will occur out of all the possible events that can occur. For example, probability of getting number 3 when rolling a fair dice. Outcomes of an event are often discussed in probabilistic terms as you can never assert something 100% certain about future.

The most basic way to estimate probability of an event is to divide number of desired outcomes with total number of outcomes possible. For example, for the above stated example, number of desired outcome is only1 (number 3) and number of possible outcomes is 6. Therefore probability of getting 3 when rolling a fair dice becomes 1/6. Probability of any event will always range from 0 to 1.

Some of the important terms to remember when solving probability related problems are:

S: Sample space. All the possible outcomes

Events: A subset of a sample space

P(A)= Probability that event A will occur

For example, outcome numbered 1,2,3,4,5,6 are all the 6 possible outcomes when rolling a fair dice and hence they become its sample space. Obtaining number 3 is a subset or event of this sample space which can be indicated as P(A)= Getting number 3; P(A) = 1/6

An event in probability refers to any occurrence. For example, probability of getting number 3 when rolling a dice is an event, probability of getting heads when tossing a coin is an event, probability of Tina getting selected as class monitor out of 10 candidates is an event.

Empty set or null set	: set with no events in it. Indicated by Ø or { }
Union	: Putting two sets together into a larger set. A or B or both. For example, Red ball or green ball
Intersection	: Finding only the common outcomes between the two sets. A and B. Red ball and green ball.
Complements	: Complement of an event A is all the events from the sample set S, that are not part of A. S= {1,2,3,4,5,6}, A={3,4}, Ac = {1,2,5,6}
Independent events	: events that can occur at the same time without affecting each other
Mutually exclusive events	: when occurrence of on event prohibits occurrence of other
Exhaustive events	: when a set of events includes all the possible outcomes, it is called as an exhaustive event.

We will encounter some more advanced terms as we proceed.

1.1.1: Axioms of probability

The axioms of probability can be thought of as already proven principles or rules that do not require any further evidence. These rules are taken as it is and applied to solve probability related problems. Following are the three of the most important axioms of probability:

(E1 is any event in sample space S)

Axiom 1 : Probability of event E1 is between 0 and 1
$0 \leq E1 \leq 1$

Axiom 2 : Probability of sample space is 1
$P(S)=1$

Axiom 3:

If events in a sample space are mutually exclusive, then the probability of their union is the sum of the probability of those events

$P(E1 \cup E2 \cup E3 \cdots)=P(E1)+P(E2)+P(E3)+\cdots$
$AB = \{\}, P(A \cup B) = P(A) + P(B)$

Some more useful rules of probability based on these axioms include:

1) $P(A)= 1 - P(A^C)$

2) $P(AUB)=P(A \cap B^C) + P(B)$
$= P(B \cap A^C) + P(A)$

3) $P(AUB) = P(A) + P(B) - P(A \cap B)$……..the addition rule

4) $P(AUBUC) = P(A) + P(B)+P(C) - P(A \cap B) - P(A \cap C)- P(B \cap C)$
$+P(A \cap B \cap C)$

5) $P(A \cap B) = P(A)*P(B)$……given that A and B are independent events

6) P(E) </= P(F) if E is subset of F

1.1.2: Random Variables

A random variable is a variable that takes on numerical values according to a random procedure. If X is the random variable, x are the values random variable X can take on.

For example., if you are given two problems to solve. 1 point is given for correct answer. Random variable y = points earned

All the possible outcomes will be :

Both the answers correct SS, points earned will be 2, therefore value of y will be 2.

First is correct second is incorrect SF, value of y will be 01

First is incorrect second is correct FS, value of y will be 01

Both the answers incorrect FF, value of y will be 0

S = (SS,SF,FS,FF)

P(y=0)= 1/4

P(y=1) = 2/4

P(y=2)= 1/4

There are two types of random variables - discrete random variable and continuous random variable. A discrete random variable is the variable that can take only specific values in the given limit. For example, number of students present in class today. If the class strength is 60, there can only be either 0,1,2 or 3…..or 60 students present in the class. There cannot be, lets say, 33.46 students present. This means that the variable of number of students present in class can only take specific values between 0 to 60. On the other hand, if we take variable of amount of water in a bottle with the capacity of 50 litres, at any point, the amount of water can take any possible value between 0 to 50 such as 1 litre, 2 litre, 6.7778492 litres, 42.889530 litres, etc. Such variables are called as continuous variables. Thus, contrary to a discrete random variable, a continuous random variable is a variable that can take any possible value within given limits.

1.1.3: Expected value

In simple words, expected value refers to average of the values of random variable. It is indicated as E(X). With respect to data such as

X = 6,6,6,2,7,7 average or mean can be computed by adding all the values and dividing it with number of values in the data. In this case it would be 6+6+6+2+7+7 / 6 = 5.66. The same example can be expressed as :

$$= \frac{6+6+6+2+7+7}{6}$$

$$= \frac{6(3) + 2(1) + 7(2)}{6}$$

$$= (3/6) + 2\,(1/6) + 7\,(2/6)$$

If you have noticed it, the values in the bracket in the above step is nothing but probability of that value. That is, in the data set of 6,6,6,2,7,7, probability of value 6 is 3/6, probability of value 2 is 1/6 and that of value 7 is 2/6. Thus, we can compute mean or expected value of a random variable X by multiplying each value of X with its own probability and adding it for all the values of X. Therefore, the formula for expected value of a random variable is :

$E(X) = \Sigma\, x.p(x)$……..for a discrete random variable

Expected value of a continuous random variable is expressed as:

$$E[X] = \int_{xmin}^{xmax} xf(x)dx$$

Source: Google images

Variance is another important measure of central tendency after mean. Variance gives an estimate of magnitude by which values in a data set vary. It is the average of the squared deviations of each value of X from its mean. In case of random variables, variance can be expressed as :

Var (X) = $E[(X-\mu)]^2$ ……. μ refers to population mean

or

Var (X) = $E[(X)^{2]} - [E(X)]^2$

1.1.4: Central limit theorem

Central limit theorem is one of the most useful theorems of statistics, especially for social sciences like Psychology. We are able to perform a variety of inferential statistical tests and draw conclusions, owing to this theorem. According to this theorem, if we take multiple samples from a population and compute mean for each of these samples, the histogram of all these means will approximate a normal distribution , irrespective of the shape of its parent population, given that samples are large enough. Several parametric statistical tests require population of the sample in question to be a normal population. However, we do not often have access to the entire population to check its normality. But, based on this theorem, we can use any parametric test by simply ensuring that we take a very large sample for study.

1.2. PROBABILITY DISTRIBUTIONS

In simple words, the shape that a particular data takes when it is plotted on a graph is called as its distribution. A probability distribution is a formula that allows us to compute possibility of specific values that X random variable can take. Based on the types of random variables there are two types of probability distributions : Discrete random distributions and Continuous random distributions. Let us study some of the important distributions under each category. In case of discrete random variables, the probability distribution formula is called as Probability Mass Function (pmf), whereas that for a continuous random variable is called as Probability Density Function (pdf).

1.2.1: Discrete distributions

As stated earlier, distribution of a discrete random variable is called as discrete distribution.

A) Binomial distribution

A situation must meet following conditions for random variable X to have a binomial distribution

- There is a fixed number of trials involving some kind of random process
- There are only two possible outcomes for each trial : success or failure
- The probability of success is same for each trial.
- The trails are independent (one outcome does not influence another outcome)

The probability of success is indicated as p and the probability of failure is indicated as q. Since these situations involve only two possible outcomes which is success and failure, the possibility of failure can also be indicated as (1-p).

An example of binomial variable would be if a class consists of 10 students, what is the probability that 6 students will pass the exam given that probability of passing exam is 0.75 which is same for all the students. This situation is meeting all the above mentioned criteria- there is a fixed number of trials (10 students), there are only two possible outcomes fore each trial (pass/fail), Probability of success is same for all the trails and each trial that is, possibility of each student passing is independent of other student passing the exam. Another example would be the probability of getting 2 heads if one flips a fair coin thrice.

Probability mass function for a binomial variable is :

$$P(x) = \binom{n}{x} \times p^x \times (1-p)^{n-x}$$

Here,

x refers to the value of the random variable whose probability we are trying to estimate

P refers to probability of success

1-p is probability of failures

n is total number of trials.

The expression x trials out of n total trials can be solved as:

$$\frac{n!}{(n-x)! \times x!}$$

The symbol of ! is solved by multiplying the value in question(x) with (x-1), (x-2), (x-3)……1. For example, 6! would be 6 × 5 × 4 × 3 × 2 × 1 = 720.

If the question is to estimate probability of 6 out of 10 students passing in the exam given that the probability of passing in the exam is 0.75, x = 6, n = 10, p = 0.75, (1-p) = 0.25.

$$\frac{10!}{(10-6)! \times 6!} \times (0.75)^6 \times (0.25)^4$$

Putting everything in the formula, the answer turns out to be 0.11.

The expected value and variance of a binomial distribution can be computed as :

$E(x) = np$

$Var(x) = n \times p \times (1-p)$

B) Poisson distribution

The Poisson Distribution was developed by the French mathematician Simeon Denis Poisson in 1837. The Poisson distribution is the discrete probability distribution of the number of events occurring in a time period, given the average number of times the event occurs over that time period.

Conditions for Poisson Distribution:

- An event can happen for any number of times during the given time period.
- Events happen independently. In other words, if an event occurs, it does not affect the probability of another event occurring in the same time period.
- The rate of occurrence is constant; that is, the rate is the same at any time of the day or date
- The probability of an event occurring is proportional to the length of the time period. For example, it should be twice as likely for an event to occur in a 2 hour time period than it is for an event to occur in a 1 hour period

For example, if you know that a hotel inquiry desk gets on an average 12 calls during 4 to 10 pm everyday and you want to estimate probability of them getting 16 calls today, poisson distribution will help you estimate that.

Probability mass function for a poisson distribution is :

$$P(X=x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

Here,

lambda (λ) is the parameter, that is, number of times an even occurs in an average in a given time period.

e is the Euler's constant which is ≈ 2.71

x is the value of random variable x whose probability we want to estimate

Lets solve an example. If a sales woman sales 3 lipsticks a week, what is the probability that she will sell 5 lipsticks this week?

Here, λ is 3 , x is 5 and e as always is 2.71. Adding all these values in equation,

$$P(X=5) = \frac{3^5 \times (2.71)^{-3}}{5!}$$

The answer turns out to be 0.10

Mean and Variance of a poisson distribution can calculated using following equation:

$E(X) = \lambda$

$Var(X) = \lambda$

1.2.2: Continuous distributions

As discussed previously, continuous distribution refers to distribution of a continuous random variable. Some of the most useful of these variable include normal distribution, t distribution, f distribution and chi square distribution

A) Normal Distribution

All of us are familiar with a normal curve. A data is said to be normal when it is spiked at the middle and tapers down to both the ends. It is one of the most useful distributions in statistics.

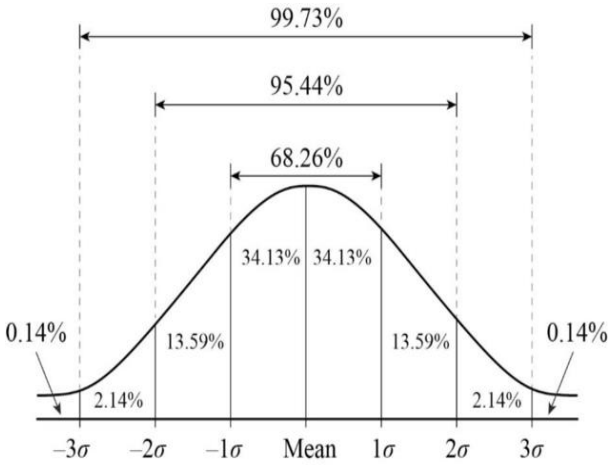


Image source: Google images

The parameters of a normal distribution are μ and σ^2. Some of the important characteristics of a normal distribution include the fact that it is asymptotic (never touches axis at either ends), mean, median and mode of normal curve are the same, it extends from positive infinity to negative infinity and its skewness and kurtosis is zero. Normal distribution is generally expressed as $X \sim N(\mu,\sigma^2)$ Probability density function of a normal distribution is indicated as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Image source: Google images

Here, pi and Euler's constant are two constants involved in the equation. Pi ≈ 3.14, e ≈ 2.71. x is the value of random variable whose probability we want to estimate, μ is mean of the distribution and σ^2 is the variance of the distribution.

B) Chi square distribution

This distribution is at the heart of the very famous chi square test which is used when data falls on nominal scale. Chi square is indicated as χ^2. Chi square distribution is made up of Z distribution. Z distribution is a normal distribution with mean 0 and standard deviation of 1 ; N(0,1). In very simple words, a chi square distribution is the sum of multiple z distributions squared. Degrees of freedom (df) for such a distribution is number of independent z distributions that the chi square distribution is made up of. If a chi square distribution is made up of 6 Z distributions then its df becomes 6.

$\chi^2 = \Sigma Z^2$

Mean of a chi square distribution is k and variance is 2k where k refers to df of the chi square. Addition of two independent chi square distribution also gives another chi square distribution. As its df increases, chi square approaches a normal distribution. Probability density function of chi square distribution is:

$\sqrt{2\chi^2} - \sqrt{(2k-1)}$

C) F distribution

The popular ANOVA test follows f distribution. An f distribution is obtained by dividing two independent chi square distributions(χ^2_1 and χ^2_2) which are in turn divided by their own degrees of freedom.

$$F = \frac{\chi^2_1/df_1}{\chi^2_2/df_2}$$

An f distribution approaches normality as its dfs become large. Mean of f distribution is $df_2/(df_2-2)$ for $df_2 > 2$ and its variance is

$$\frac{2\,df_2^2\,(df_1 + df_2 - 2)}{df_1\,(df_2 - 2)^2\,(df_2 - 4)}$$

D) t distribution

This distribution is at the heart of one of the most popular statistical test which is t test. It can be obtained with the help of an independent Z and chi square distribution with k degrees of freedom.

$$t = \frac{Z}{\chi^2/k}$$

If F distribution has $df_1=1$, then square root of F distribution will provide a t distribution. A t distribution approaches normality as df becomes approximately 30 and more. That is why the traditional rule of thumb suggests researchers to obtain a sample size of at least 30. The expected value of t distribution is 0 and variance is k/(k-2).

1.2.3: Jointly distributed random variables

A joint probability can be defined as two events happening at the same time. For example, probability of a student being a girl and living in Pune or probability of a target being red coloured and triangle. Joint probability of two events X and Y is denoted as p(X,Y). If events X and Y are independent then their joint probability can simply be calculated by multiplying probability of each one of them - p(X) × p(Y). However if the two events are not independent then a more complex procedure is required.

1.3 SUMMARY

- Probability refers to the chance that a specific event will occur out of all the possible events that can occur
- Simple formula for probability is to divide number of desired outcomes with total number of outcomes. For example, when rolling a dice, probability of getting even numbers is 3/6
- Axioms of probability are some basic rules that do not need to be proved further. There are three basic axioms:

  Axiom 1: $0 \leq E1 \leq 1$

  Axiom 2: $P(S) = 1$

  Axiom 3: $AB = \{\}, P(A \cup B) = P(A) + P(B)$
- A random variable is a variable that takes on numerical values according to a random procedure. There are two types of random variables- discrete random variables and continuous random variables. Discrete random variables include those events where x can take only specific values in the given limit whereas continuous random variables can take any possible value in the given limit

- Expected value refers to mean of a random variable and it is estimated by formula $E(X) = \Sigma\ x.p(x)$. Variance of a random variable is estimated by formula : $Var\ (X) = E[(X)^{2]} - [E(X)]^2$

- Central limit theorem is one of the most useful theorem in statistics which states that if we take multiple samples from a population and compute mean for each of these samples, the histogram of all these means will approximate a normal distribution , irrespective of the shape of its parent population, given that samples are large enough.

- Based on the types of random variables there are two types of probability distributions : Discrete random distributions and Continuous random distributions.

- Some of the important distributions under discrete random distributions include binomial and poisson distributions. Binomial distribution is used in cases where an event can have only two outcomes and probability of success is same for all the independent trials. On the other hand poisson distribution is used when we know the average number of times an event occurs in a given time frame and want to estimate probability of that event occurring for a particular number of times in a particular time frame.

- Properties of binomial distribution include $E(X)=np$ and $Var(X) = n*p*(1-p)$. Properties of poisson distribution include $E(X)= \lambda$ and $Var(X) = \lambda$

- Some of the useful continuous random variables include normal distribution, chi square distribution, F distribution and t distribution. Properties of each one of them is discussed in the chapter above

- Joint probability refers to probability of two events happening at the same time which is denoted by P(X

1.4 QUESTIONS

1) Define sample space if I flip a fair coin thrice

2) If three fair coins are flipped, what is the probability of getting at least two tails?

3) If a six sided die is rolled, what is the probability of getting a number that is at most 4?

4) $P(A)=0.4$, $P(B)=0.3$, $P(A \cap B)= 0.2$, Given that all the events are not mutually exclusive, then find:

   P (AUB)

   P (AC)

   P (BC)

   P (AUB)C

P (A∩BC)

P (AC∩B)

P (AUBC)

P (AC UB)

5) Calculate expected value for following data :

x	p(x)
0	0.17
1	0.48
2	0.35

6) Studies have found that 40% of the patients of depression improve after behavioral therapy. If six patients are treated with behavioral therapy, what is the probability that 2 of them show improvements?

7) On an average a college help desk gets 7 phone calls per hour. What is the probability that today they will receive 16 calls between 9 to 10 AM?

8) A statistician records the number of cars that approach an intersection. He finds that an average of 2.7 cars approach the intersection every minute. Assuming the number of cars that approach this intersection follows a Poisson distribution, what is the probability that

   A. 2 cars will approach the intersection?

   B. No car will approach the intersection?

   C. Less than 3 cars will approach the intersection?

   D. 3 or more cars will approach the intersection within a minute?

9) Explain continuous random variables.

10) Describe the relationship between chi square, F and t distribution.

11) Write a short note on joint probability distributions.

1.5 REFERENCES

- Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.
- Belhekar, V. M. (2016). Statistics for Psychology Using R. SAGE publications.

❄❄❄❄❄❄❄

2

PRELIMINARY CONCEPTS - II

Unit Structure :

2.0 OBJECTIVES

After studying this unit a student will be able to:

- Understand process of statistical hypothesis testing
- Understand different statistical estimators along with their properties and methods to estimate them
- Understand what is Bayes theorem and solve related sums
- Understand and apply descriptive statistics
- Understand concepts of power and effects size
- Learn assumption of normality and ways to asses the same

2.1: INFERENCE: ESTIMATION THEORY, STATISTICAL HYPOTHESIS TESTING, TYPES OF ERRORS

The fundamental problem of inference process in statistics is that we often do not have access to the entire population. What we have is a limited representative sample. For example, if a researcher wants to conduct a study on stress among working women of India, it will be impossible to contact each and every working woman in India and assess her stress levels. However what the researcher can do is collect a representative sample of working women from all over the India, assess their stress levels and based on the values obtained estimate probable stress levels of all the working

women in India with the help of suitable statistical test. Various statistical tests allow us to estimate population value based on sample values. Therefore the process of drawing statistical inference is also called as a process of estimation. Sample's values or measurable characteristics are called as "Statistics" (For example, sample mean (X) and Standard Deviation). Population value or measurable characteristics of the population are called as "Parameters" (For example, population mean(μ) or population Standard Deviation). A population parameter is unknown but sample statistic is known. Since the population parameter is unknown, it is estimated using sample statistic, whereas sample statistic is calculated using data. Since the unknown population parameter is estimated using sample statistic, sample statistic is called an "estimator". The process of estimating a population value from sample statistic is called estimation or statistical inference. Population parameter is stable/constant whereas sample statistic changes slightly when another sample is drawn hence it is a variable.

2.1.1 : statistical hypothesis testing

To understand the process of statistical hypothesis testing first of all, understanding the concept of sampling distribution is important. Let us continue with the same example of assessing stress level of working women in India. Suppose we take one sample from a population and that population has mean μ. We calculate statistic on the sample. Lets say we calculate mean. lets call this mean of first sample as $\overline{X}1$. Then we return this sample to the population. Now we obtain another sample from the same population and calculate its mean $\overline{X}2$. Second sample is also returned to the population. We keep on repeating the procedure. Thus, we shall have a large number of sample means with us. If we take 500 such samples we will have 500 means with us. These sample means shall follow a distribution. This distribution is called a distribution of sampling means. The mean of the sampling distribution is population mean. The SD of the sampling distribution is standard error of statistics which is a fundamental requirement for null hypothesis testing. Sampling distribution of means follows a normal distribution. Each mean in the sampling distribution varies slightly from each other it informs us about the chance of getting a specific value. You can compute any other statistic on the collection of samples and plot it on a graph, which will provide you a new sampling distribution.

There are two type of statistics - Null and alternative. Null hypothesis (H_0) essentially asserts that the results obtained in sample are due to chance whereas alternative hypothesis (H_1) asserts that the results are not due to chance but rather due to experimental manipulation. Further, null and alternative hypothesis have sub types called as directional and non directional hypothesis. An alternative directional hypothesis states direction of the difference. For example, mean of group A is higher than mean of group B. Whereas an alternative non directional hypothesis states there is a difference between two groups without stating the direction of that difference. For example, there is a significant difference between mean of group A and group B.

Lets assume that a researcher collects data on stress levels of working women in India and holds null hypothesis that the sample comes from a

population with mean 90 and SD = 15. Mean of 98 is obtained with SD 15 on a sample size of 500. Then the hypothesis are:

$H_0 = \mu = 90$

$H_1 = \mu \neq 90$

Once the null and alternative hypotheses are formed, one needs to work out its sampling distribution. Sampling distribution of mean is normal distribution. If we keep on taking random samples from a population that has $\mu = 90$, there is some chance that we will obtain a sample with mean 98.

The null hypothesis significance would find the probability of obtaining the value of mean 98 or more when the population mean is 90.

Lets say the probability of that is 0.0092. It means: if the population mean is 90 then the chance of obtaining a sample with mean 98 is 0.0092 or approx. 0.01

Researcher has 2 options:

1) Accept that the sample has come from a population with 110 mean, but due to sampling variation its a very unlikely value (accepting the null)

2) Accept that the sample I have drawn has not come from a population with mean 110 as the chance of it is very low (rejecting the null).

The researcher will take this decision based on the prior set alpha level(α). Alpha level can be thought of as the demarcation after which null hypothesis is rejected. In simple words, it is the amount of error that the researcher is okay with committing. If the alpha level is 0.05, then the researcher will accept null hypothesis as long as the probability of obtaining sample value is more than 0.05. If the probability of obtaining the given sample value is less than 0.05, the researcher will reject the null hypothesis. The exact value of the sampling distribution of the statistics at 5% level is called critical value. If calculated value exceeds critical value, the null hypothesis is rejected but if calculated value is lessor than critical value, the null hypothesis is not rejected.

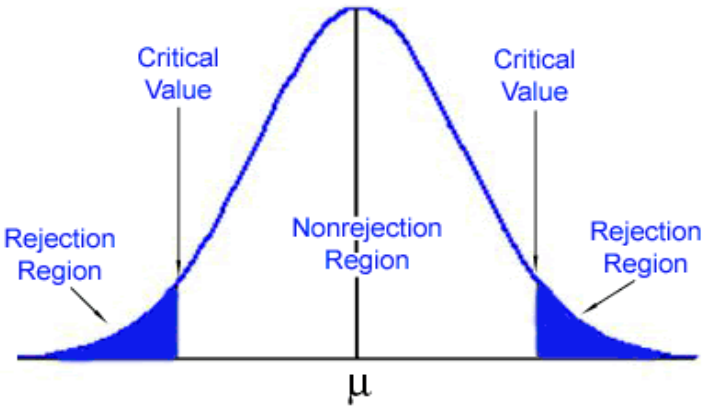


Figure 2.1: Area of rejection an retention.

Image source: Google images

This is how the process of statistical hypothesis testing is carried out.

2.1.2 : Types of errors

Whenever we are testing a statistical hypothesis, we are essential estimating probability of population value being a particular number given the sample value. Since the inference is based on probability and not 100% guarantee, there is always some chance of there being an error in this estimation process. Accordingly, there are two types of errors in hypothesis testing.

Type I error is committed when one rejects the null hypothesis when in reality it shouldn't have been rejected. On the other hand, Type II error is committed when one accepts the null hypothesis when in reality it is false. Probability of type I error is called α and probability of type II error is called β. As we try to reduce type I error, we end up increasing chance of type II and vice versa.

Table of error types		**Null hypothesis (H_0) is**	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Figure 2.2: Types of errors in hypothesis testing

Image source: Google images

2.1.3 : Properties of estimators, methods of estimation: least square, maximum likelihhod

As discussed earlier, since sample values allow us to estimate population values, sample values are also called as estimators. There are two types of estimators: point estimators and interval estimators. Point estimators, as the name suggests, provide a single point based on which population parameter can be estimated. Examples of point estimators include mean, correlation coefficient, etc. On the other hand interval estimators provide a range within which population value is likely to lie. Examples of interval estimators include confidence interval, etc.

An estimator has some properties which make some estimators preferred, more accurate, over others. These properties are further divided into two categories: Small sample properties and Large sample properties. Useful small sample properties include unbiasedness, minimum variance, efficiency, linearity, BLUE and MSE.

An estimator is called as unbiased estimator when expectation or average of sample value is population value. Mean is a well known unbiased estimator. As encountered previously, if we collected several samples from a population, compute mean for each of that sample and plot them on a graph, the mean of such a distribution will turn out to be population mean. An estimator is said to have minimum variance when its variance is smaller than any other available estimator. An estimator among all the unbiased estimator that has lowest variance is called as an efficient estimator. An efficient estimator is called a best estimator. As the name suggests, a linear estimator is the linear function of sample values. The acronym BLUE stands for an estimator that is Best Linear Unbiased Estimator. Mean squared error (MSE) of an estimator is the squared difference between estimator and its corresponding population parameter. It is expected to be as low as possible.

Large sample properties of an estimator essential means how an estimator behaves as sample size increases. These properties include asymptotic unbiasedness, consistency, asymptotic efficiency, asymptotic normality and sufficiency. The term asymptotic refers to the property of a distribution that does not touch the X axis on either ends and extends from minus infinity to plus infinity. An estimator is called as asymptotically unbiased when as sample size nears infinity, expectation of estimator nears the corresponding parameter. If estimator approaches true value of parameter as the sample size increases, the estimator is called to be a consistent estimator. An asymptotically unbiased estimator that has smaller variance among the consistent estimators is called as asymptotically efficient estimator. When sampling distribution of the estimator approaches normality as sample size increases, it is said to have the property of asymptotic normality. Lastly, when an estimator is a function of sample observations and uses all the information available about the sample, it is called as a sufficient estimator.

Two of the competing methods of obtaining estimators include method of maximum likelihood and the method of least squares. R.A.Fisher developed the principal of maximum likelihood estimation. This method uses something called as likelihood function. Likelihood function is the function of parameter with data values being fixed. The method of maximum likelihood tries to obtain estimator in such a way that this probability is maximized. On the other hand, method of least squares tries to obtained estimators in such a way that the squared difference between sample data (estimator) and population parameter is as least as possible. The very useful regression analysis is based on this method of estimation.

2.2 BAYESIAN INFERENCE

Bayes theorem or Bayesian inference is one of the most fundamental theorems of statistics. It uses conditional probability- the probability that is expressed in terms of if-then statements. Lets say that P(A) and P(B) are probabilities associated with events A and B. P(A|B) is a conditional probability, it is the probability of observing A given that B has occurred and P(B|A) is a conditional probability, it is the probability of observing B given that A has occurred. Then, according to Baye's theorem,

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Let is solve a sum. Floods are rare (1%) but heavy rain is fairly common (10%) . 90% of floods make rains. Then estimate the probability of there being a flood given that today there was heavy rain.

P(A) = Floods = 0.01

P(B)= heavy rains = 0.1

P(B|A)= 0.9

Using Baye's theorem,

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$$P(A|B) = \frac{0.9 \times 0.01}{0.1}$$

$$P(A|B) = 0.09$$

Therefore, the probability of flood occurring given that it rained heavily today is 0.09.

The Baye's equation can be derived from basic law of conditional probability.

According to the law of conditional probability,

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

Now,
$P(A \cap B) = P(B \cap A)$ .....(i)

Also,
$P(A \cap B) = P(A|B)*P(B)$ $P(B \cap A) = P(B|A)*P(A)$ ..... (ii)

Using (i) and (ii),

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

$$= P(A|B) \times P(B) = P(B|A) \times P(A)$$

$$= P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

An alternative Baysian formula to estimate probability of B given A is

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B|A)*P(A) + P(B|AC)*P(AC)}$$

Let us solve a sum. 1% of people have a certain genetic defect. 90% of tests of the genetic defect turn out to be true positive. 9.6% of the tests are false

positives.If a person gets a positive test result, what are the odds they actually have the genetic defect?

Here,

P(A)= Genetic defect = 0.01 , $P(A^{c)} = 0.99$

P(B)= Positive test results

P(B|A)= 0.9

$P(B|A^{c)} = 0.096$

Putting all the values in the formula, the answer turns out to be 0.086.

2.2.1 : Cramér–Rao inequality; Rao Blackwell Theorem

The principle of Cramer-Rao inequality was presented by Herald Cramer and Calympudi Radhakrishna Rao. It postulates that the highest amount of variance any unbiased estimator can have is as least as Fisher's information. Fisher's information is a method of estimating how much information a sample estimator carries about population parameter. The one that has lowest MSE out of all the unbiased estimators is then called as Minimum variance unbiased estimator.

The principle of Rao Blackwell theorem was given by Calympudi Radhakrishna Rao and David Blackwell. It states that if g(X) is estimator of any population parameter theta and T(X) is a sufficient estimator then conditional expectation of g(X) given T(X) is a better estimator of population parameter theta. This theorem can be used to improve any estimator by obtaining its conditional expectation given any other sufficient statistic.

2.3 DESCRIPTIVE STATISTICS: CENTRAL TENDENCY AND VARIABILITY

Measures of central tendency provide Provide measures of average or central point of the data. The three measures of central tendency include mean, median and mode. Mean is the most useful among them and it is obtained by dividing the sum of scores (Σ) by number of scores (n). For example for a data set of 5,5,6,7,3,2 mean would be 5+5+6+7+3+2/ 6 = 4.66. Mean reflects each score in the data set but is very sensitive to extreme scores. It allows for calculation of a lot of inferential statistics such as t test, ANOVA, etc. It is estimator of population mean μ. Median is the middle score of the data. It is obtained by arranging data in ascending or descending order and then identifying the mid point that lies at N/2th position. It is the 50th percentile of the data. It is less sensitive to extreme scores but has limited usage for further algebraic and inferential statistics; it is useful for ordinal data. Mode is that score in the distribution that has highest frequency. A data can have more than one mode which is called as a bimodal or multi modal data. A normal distribution, however is always unimodal. This statistic can be used with nominal data however cannot be

used in majority of the inferential statistics. Unlike mean, mode does not represent entire data. It can only be used to describe sample properties.

Measures of variability help us understand how the data is dispersed in a sample. Provides a measure of spread of the data around mean. Four of the most popular measures of variability include range, quartile deviation, average deviation and standard deviation. Range is the most simple statistics that reflects the difference between the lowest score and highest score of the data. It heavily depends on extreme scores. Interquartile range solves the problem of heavy dependence on extreme scores. It is computed by eliminating upper 25% and lower 25% of scores: Q_3-Q_1. If we divide interquartile range by 2,we get what is called as semi quartile range. It is less sensitive to extreme scores and more useful in skewed distributions however it is affected by sampling variation. Average deviation calculates deviation around mean by using formula:

$$\Sigma|(X_i - \overline{X})| / n$$

Sample variance (S^2) is the average of the squared deviation from the mean. When used as an estimator of population variance, denominator changes to n-1. Square root of variance is called as standard deviation. That is, standard deviation is the square root of the average of the squares of the deviations of each score from the mean. It is the most stable and reliable measure of variability which is used in several inferential statistical tests such as t test and ANOVA.

2.4 POWER AND EFFECT SIZE

Power of a test is its ability to reject null hypothesis when null hypothesis is false in reality. It is denoted by 1-β. Power depends on following factors:

1) Probability of type I error - power is higher for 0.05 than 0.01 alpha level
2) As sample size increases power also increases
3) True alternative hypothesis produces greater power
4) Power for one sided test is higher than two sided test

Null hypothesis significance testing informs us whether difference between two groups (populations) is non zero or not. The question of how large is that difference can be answered by effect size. The popular method to estimate effect size is Cohen's D given by Cohen (1960). It can be computed by using formula:

$$d = (\overline{X}_1 - \overline{X}_2) / Sp$$

Here, $\overline{X}_1$ and $\overline{X}_2$ refer to means of group 1 and 2 that are being compared. Sp refers to pooled variance which is average variance of group 1 and 2.

A d value between 0.2 to 0.5 is considered as the small effect size, d value between 0.5 to 0.8 is medium effect size, d value of 0.8 and above is considered to large effect size

2.5 TESTING FOR NORMALITY AND OUTLIERS

The assumption of normality underlies majority of the inferential statistical tests, especially the parametric tests. The assumption states that in order to use any parametric test such as t test, ANOVA, etc, the population from where study sample is derived should be normally distributed. This assumption can be assessed by using graphical or statistical methods. Popular graphical methods include QQ plots, GG plots, histogram, box whisker plot, etc. Examples of statistical methods include skewness test, Jarque-Bera test (JB test), W/S test, Shapiro Wilk's test, etc.

Histogram is a test of normalcy where in if the joint bars take an approximate bell shape, the data is assumed to be normal. However it is a very crude method of assessing normality. The preferred graphical method is QQ plots. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Q-Q plots take the sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. If a set of observations is approximately normally distributed, a normal QQ plot of the observation will result in an approximately straight line. QQ plot moves in an upward direction when the data is positively skewed, it moves in downward direction when the data is negatively skewed. On the other hand, in the box whisker plot the bottom and top of the box is first and third quartile which contains 75% of the observations. Horizontal line in the box indicates median. Whiskers extend to the observations in each tail of the data which is fourthest from the fourths but less than or equal to 1.5 times the fourth spread (Q1-1.5IQR, =10 Q3+1.5IQR = 90) IQR = Q3-Q1. The outliers are plotted as individual points.

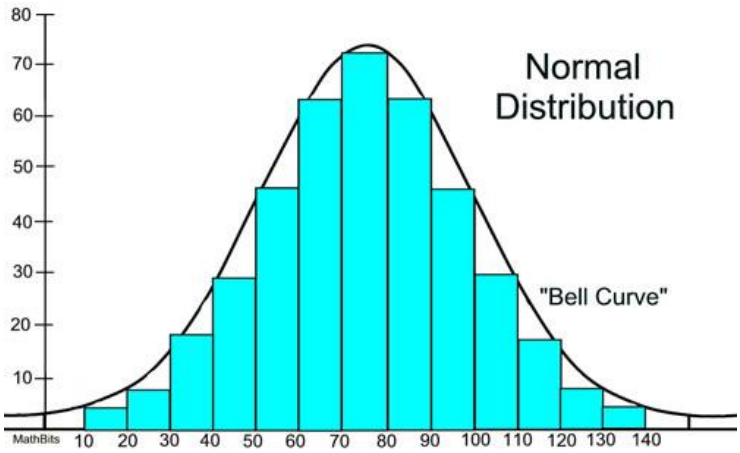


Figure 2.3: Histogram

Image source: Google images

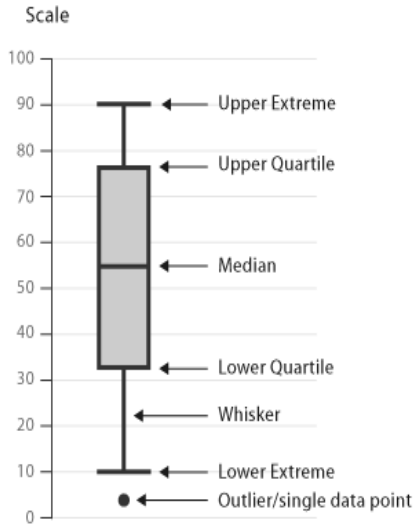


Figure 2.4: Box whisker plot.

Image source: Google images

The statistical method of skewness is based on the characteristic of normal distribution that its skewness is zero. It uses the formula $g1= m^3/m2^{3/2}$. When $g1 > 0$ then the data is skewed to the right, when $g1 < 0$ then the data is skewed to the left. Method of skewness and kurtosis for assessing normality are also called as methods of moments. The JB test on the other hand combines information from skewness and kurtosis test and follows chi square distribution. The W/S test of normality makes use of studentised q statistic.

Traditionally, outliers are defined as those data points which are 3 standard deviations above or below the mean. These values are extreme and do not represent the true nature of the population. The outliers can be identifies with the help of box whisker plots or a line graph. The outliers being extreme scan affect further statistical calculations. Hence the popular practice is to remove such outliers. However at times the researcher may not be in a position to compromise on data after removal of outliers, especially if it is a small sample size. In such cases the method of trimming or winsorization can be used. The method of trimming involves discarding some percentage of observations in the extremes of the data whereas the method of winsorizing involves replacing most extreme values by immediate less extreme values.

2.6 SUMMARY

- In social sciences, we often do not have access to population but we have access to sample. We collect a representative sample and try to estimate population value based on sample values. Sample values are called as statistics whereas population values are called as parameters. As sample value estimates population value it is also known as estimator.

- The process of hypothesis testing involves defining hypotheses first. There are two types of hypotheses. Null hypothesis states that the obtained sample results are due to chance whereas alternative hypothesis states that obtained sample results are not because of chance but because of experimental manipulation. Once hypothesis is defined the next step is to work out sampling population. Based on the prior set alpha level and obtained probability of getting sample values, researcher either rejects or does not reject the null hypothesis.
- Two types of errors are inherently involves in null hypothesis testing. Type I error is when a researcher rejects null hypothesis when it should not have been rejected whereas Type II error is when null hypothesis is accepted when in reality it is false.
- An estimator has some properties which make some estimators preferred, more accurate, over others. These properties are further divided into two categories: Small sample properties and Large sample properties. Useful small sample properties include unbiasedness, minimum variance, efficiency, linearity, BLUE and MSE.
- The two popular methods of deriving estimators are maximum likelihood method and method of least squares.
- Bayesian inference is one of the most fundamental theorems of statistics that allows us to compute conditional probability of event A given B based on probabilities of event A, event B and event B given A.
- Cramér–Rao inequality; Rao Blackwell Theorem further guide selection of the most accurate estimator
- Measures of central tendency and measures of variability are two types of descriptive statistics. Measures of central tendency includes mean, median and mode whereas measures of variability includes range, quartile deviation, average deviation, variance and standard deviation.
- Power of a test is its ability to reject null hypothesis when null hypothesis is false in reality. Effect size helps us estimate how large is the difference between two groups given that the difference between two groups has turned out to be significant.
- The assumption states that in order to use any parametric test such as t test, ANOVA, etc, the population from where study sample is derived should be normally distributed. Q-Q plots, histogram, box whisker plot are some graphical methods to assess normality and methods of moments, Shapiro wilk's test, JB test are some of the statistical methods of assessing normality.

2.7 QUESTIONS

1) What is estimation theory?

2) Describe the process of null hypothesis significance testing

3) Write a short note on types of errors

4) What is the assumption of normality and how to assess that?

5) In a particular pain clinic, 10% of patients are prescribed psychotropic drugs. Overall, five percent of the clinic's patients are addicted to narcotics (including pain killers and illegal substances). Out of all the people prescribed psychotropic drugs, 8% are addicts. If a patient is an addict, what is the probability that they will be prescribed psychotropic drug?

6) An advertising executive is studying television viewing habits of married men and women during prime time hours. Based on the past viewing records he has determined that during prime time wives are watching television 60% of the time. It has also been determined that when the wife is watching television, 40% of the time the husband is also watching. When the wife is not watching the television, 30% of the time the husband is watching the television. Find the probability that if the husband is watching the television, the wife is also watching the television?

2.8 REFERENCES

- Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.
- Belhekar, V. M. (2016). Statistics for Psychology Using R. SAGE publications.

❄❄❄❄❄❄❄

3

INFERENTIAL STATISTICS: INFERENCE ABOUT LOCATION - I

Unit Structure:

3.0 OBJECTIVES

After studying this unit students will be able to:

- Understand which inferential statistical tests are appropriate for which data and study design
- Solve sums based on t test and ANOVA
- Differentiate between different types of t and ANOVA tests
- Understand the technique of bootstrapping

3.1 TWO GROUP DIFFERENCES: T TEST- INDEPENDENT AND DEPENDENT SAMPLES

On many occasions in the field of Psychology researchers are interested in knowing whether an independent variable has a significant effect on dependent variable. This effect can often be inferred by studying whether two groups that differed with respect to treatment received differ from each other or not. For example, a researcher might be interested in knowing whether using a particular method of memory enhancement actually leads to improved learning among students. She may collect two groups of students, lets say of size 100 each. Both the groups can be presented with the same material to learn however one group will be instructed to remember the presented material using rote learning method whereas other group will be instructed to use imagery method. Further both the groups can be asked to recall the material presented. The researcher may compute mean recall scores for both the groups to know which group has performed better. However this conclusion will be limited to the specific sample studied. If

the researcher wants to generalize these results to the entire population of students she can do so with the help of inferential statistics of t test.

t test or students t test is one of the most popular inferential statistical test that is used to find out significance of difference between two means or two groups. It is used when the study has only one independent variable which is manipulated at only two levels. The test was given by William Sealy Gosset in 1908. It is a parametric test which means it can be used when the data or the dependent variable of the study lies on interval or ratio scale and assumptions of normality and homogeneity of variances are met.

Types of t tests:

There are essentially 3 types of t tests: single sample t test, independent samples t test and dependent samples t test. Single sample t test is used when we want to know whether a sample at hand belongs to a particular population or not. For example, a school teacher has measured IQ of 8^{th} standard students in her school, mean for which is found to be 110. She wants to know whether the sample belongs to the population of 8^{th} standard students whose IQ is 100. Single sample t test will help answer this question.

Independent samples t test used when design of the study is randomized groups design whereas dependent samples t test is used when design of the study is repeated measures design. Stoop experiment is a famous example of dependent samples t test in which same sample of participants are exposed to both congruent as well as incongruent colour words and the number of errors and reaction time is measures as dependent variable to assess amount of interference experienced in both the conditions. On the other hand, If a researcher wants to know which of the treatments - psychoanalysis or CBT is better for depression, they may select two separate group of participants suffering from depression who will then be administered either of the treatments. The significance of difference between depression scores of two groups at the end of the treatment will indicate which one is more effective than the other.

Assumptions underlying t test:

1) Assumption of normality:

   Assumption of normality requires that the parent population from where the study samples are drawn from should be normally distributed.

2) Equality of variances

   This assumption is also known as homogeneity of variances or homoscadasticity. This assumption requires variances of populations from where study samples are taken from to be equal

   Both the above-mentioned assumptions can be tested via methods described in chapter I

3) Samples are independent

   It is essential that the scores in the sample selected are independent of each other.

4) Samples are drawn from the population at random

   Only when the samples are selected at random, they becomes representative of the population and allow generalization of sample results to the population.

Steps involved in t test calculation:

Following steps can be followed when performing t test, based on null hypothesis significance testing procedure described in chapter I

1) First, the researcher should formulate null and alternative hypothesis for the stud. Null hypothesis of t test states that there is no significant difference between means of the two groups whereas alternative hypothesis states that there is a significant difference between the means two groups.

   H_0: $\mu_1 = \mu_2$

   H_1: $\mu_1 \neq \mu_1$

2) Next step is to determine the level of significance that is the alpha level. Popular alpha levels are 0.05 or 0.01.

3) The researcher then has to decide whether to use one tailed or two tailed test.

   One tailed test is used when researcher expects a direction of difference, when alternative hypothesis of the study is directional. For example, in the above given example of techniques to enhance memory, if the researcher expects students receiving imagery method to show higher recall than those using rote learning method, she would have to use one tailed t test where the difference is expected to lie at the positive end of the normal curve. Similarly, two tailed test is used when hypothesis is non directional. If in the same example the researcher was only expecting there to be a difference in the mean recall scores of two groups without any specific direction, such a study would require two tailed t test where the difference is expected to lie at either side of the normal curve.

4) Based on the design of the study the researcher then has to decide whether to use independent samples t test or dependent samples t test.

5) Collect the required data by administering levels of independent variable.

6) Once the data is collected researcher needs to compute mean and standard deviation for the two groups

7) The calculated mean and standard deviation can then be added to the formula to compute calculated t value

8) Calculated t value is then compared to the critical t value obtained from t table of significance.

9) Lastly, if the calculated t value is greater than critical t value, the researcher will reject the null hypothesis; if calculated value is less than critical value the researcher will not reject the null hypothesis.

Formula for independent samples t test:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Figure 3.1: Formula for independent samples t test

Source: google images
Site: https://vitalflux.com/two-sample-t-test-formula-examples/

Here, $\overline{X}_1$ and $\overline{X}_2$ stand for mean of group 1 and 2 respectively.

n_1 and n_2 are sample sizes of group 1 and 2 respectively.

Sp is pooled variance or average variance of group 1 and 2. If sample sizes of the two groups are equal then pooled variance can be computed by using formula:

$$Sp^2 = \frac{S_1^2 + S_2^2}{2} \quad \text{.....(i)}$$

If the sample sizes of the two groups are unequal, then following formula can be used to compute pooled variance:

$$Sp^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad \text{..... (ii)}$$

Formula for dependent samples t test:

$$t = \frac{\overline{d}}{S_d / \sqrt{n}}$$

Figure 3.2: Formula for dependent samples t test

Source: google images
Site: https://247amend.com/2016/03/paired-sample-t-test-illustration.html

The formula for computing degrees of freedom for independent samples t test is $(N_1 - 1)+(N_2-1)$ or $N_1 + N_2 - 2$. Degrees of freedom for dependent samples t test can be calculated by using formula N-1

Solved sums:

1) A researcher wanted to find out whether there is a difference in subjective well being of individuals living in city A vs those living in city B . He collected data from a total of 20 individuals as follow. Use appropriate test statistic to find out whether the two groups differ in their level of subjective well being.

City A	City B
45	34
38	22
52	15
48	27
25	37
39	41
51	24
46	19
55	26
46	36

Note: This data is taken from
http://faculty.webster.edu/woolflm/ttest.html

Here, it can be inferred that the design of the study is randomized group design and hence independent samples t test will be used.

$\overline{X}_1$ and $\overline{X}_2$ are 44.5 and 28.1 respectively. Standard deviations of City A and City B are 8.6 and 8.5 respectively. As number if participants in both the groups are equal, formula (i) can be used to compute pooled variance.

Adding all the values in the formula, t_{cal} turns out to be 4.25 and t_{crit} is 2.1 at $\alpha = 0.05$ and df of 18. As the calculated value is greater than critical value, we can reject the null hypothesis and state that the two cities differ significantly in their level of subjective well being.

2) A teacher wanted to know if extensive training in arithmetic will improve students performance on statistics test or not. He gave students a test of statistics before and after the training program (pre and post questions different of course). Use appropriate test statistic to find out whether the training has any effect on performance in statistics test or not.

Student no.	Pre scores	Post scores
1	18	22
2	21	25
3	16	17
4	22	24
5	19	16
6	24	29
7	17	20
8	21	23
9	23	19
10	18	20
11	14	15
12	16	15
13	16	18
14	19	26

Student no.	Pre scores	Post scores
15	18	18
16	20	24
17	12	18
18	22	25
19	15	19
20	17	16

Note: this data is taken from:
www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf

First, we will compute a column of difference between the two post and pre scores.

Student no.	Pre scores	Post scores	d
1	18	22	4
2	21	25	4
3	16	17	1
4	22	24	2
5	19	16	-3
6	24	29	5
7	17	20	3
8	21	23	2
9	23	19	-4
10	18	20	2
11	14	15	1
12	16	15	-1
13	16	18	2
14	19	26	7
15	18	18	0
16	20	24	4
17	12	18	6
18	22	25	3
19	15	19	4
20	17	16	-1

Calculated mean and standard deviation of the difference turns out to be:

$\overline{d}$=2.05 and sd=2.837

Therefore, $S_d / \sqrt{n} = 2.837 / \sqrt{20}=0.634$

t = 2.05 / 0.634 = 3.23 at df of 19.

Thus, as t_{cal} is 3.23 and t_{crit} is 2.09 at $\alpha = 0.05$, we can reject the null hypothesis.

3.2 BOOTSTRAPPING

Bootstrapping is a procedure that is used to enable usage of t test when faced with problems such as non normality, heterogeneous variances, small sample etc. that violate its basic assumptions. In this procedure, multiple samples are taken out of the original sample by using sampling with replacement. The statistic of interest such as mean and SD is computed on each of this sample. Then the computed statistic is used to calculate t value for each sample. This process is repeated multiple number of times. Lets say the process is repeated 500 times so now we have 500 t values with us. A distribution of these t values can then be created. This distribution can further be used to estimate confidence interval. The obtained t values are ordered from smallest to largest in order to test the null hypothesis.

3.3 MULTI-GROUP DIFFERENCES: ONE-WAY ANOVA: INDEPENDENT AND DEPENDENT SAMPLES

As discussed earlier, psychologists are often interested in knowing whether an independent variable has a significant effect on dependent variable. At times an independent variable has more than two levels. In this case the researcher has to compare means of more than two groups. As the name suggests, this test utilizes variance. Variance is nothing but square of the standard deviation (σ^2). As the variance has additive property, the total variance can be broken down into sub groups and can then be compared. Two of the variances analysed are within group variance and between group variance. Within group variance refers to how much scores of a single participant varies across different treatments they receive. This variance for all the participants is averaged to obtain within groups variance. Between group variance refers to how much do means of different treatment groups vary from grand mean which is mean of scores of all the participants across all the groups. The final statistic is received by taking ratio of between groups variance to within groups variance.

Let us understand this by extending previous example. For example, a researcher might be interested in knowing whether using a particular method of memory enhancement actually leads to improved learning among students. Lets say this time she is interest in testing effect of techniques of rote learning, imagery and method of loci. She may collect three groups of students of size 100 each. All the groups can be presented with the same material to learn however one group will be instructed to remember the presented material using rote learning method, second group will be instructed to use imagery method and third group will be asked to use method of loci. Further all the groups can be asked to recall the material presented. The researcher may compute mean recall scores for all the groups to know which group has performed better. However this conclusion will be limited to the specific sample studied. If the researcher wants to generalize these results to the entire population of students she can do so with the help of inferential statistics of ANOVA test.

Analysis of Variance (ANOVA) test is based on f distribution. Students may find discussion of f distribution in previous chapter. ANOVA test is used when one has to compare more than two means. This can happen in two ways. Either the experiment has one independent variable which has more than two levels or the experiment has more than two independent variables with each having at least two levels.

ANOVA is also a parametric test hence can be used only when dependent variable lies on interval or ratio scale and when following assumptions are met:

1) Dependent variable should be normally distributed in population

2) The populations from where samples are drawn should have equal variances

3) All the observations should be independent of each other.

4) Participants should be randomly selected.

5) The contributions to variance in total sample is cumulative.

The two broad types of ANOVA tests include one way ANOVA and two way ANOVA. One way ANOVA is used when the study has only one independent variable which has more than two levels. One way ANOVA is further divided into within groups ANOVA which is used when the design of the study is repeated measures and between groups variance is calculated when it is randomized groups design.

The null hypothesis and alternate hypothesis of ANOVA test are as follow:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_n$$

H_1: Not null

Solved sums:

1) A manager divided 20 of his employees into 4 equal groups and asked them to go through 4 different kinds of communication skills training. At the end of the training, communication skills of all the employees were measured on a likert scale. Use appropriate statistics and draw conclusion.

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Note: The above data is taken from
https://atozmath.com/example/CONM/anova.aspx?he=e&q=anova1

First of all, as the design of the study is randomized groups design and there is only one IV with 4 levels that is Training method with levels A,B,C,D, appropriate test will be One way ANOVA of independent samples.

In the procedure described below, $\overline{X}$ refers to grand mean / mean of all the scores in the data, $\overline{X}_i$ refers to mean of a particular group and X_{ij} refers to score of an individual participant.

Step 1 is to compute sum of squares between groups (SS_b)

$$SS_b = n\Sigma(\overline{X}i - \overline{X})^2 = 50$$

Step 2 is to compute sum of squares within (SS_w)

$$SSw = \Sigma(X_{ij} - \overline{X}_i)^2 = 208$$

Step 3 is to compute total sum of squares

$SS_b + SS_w = 258$

Step 4 is to calculate variance between samples

$MS_b = SS_b / df_b$

Df for between groups is k-1 where k is number of groups. Hence df_b here will be 4-1=3

$MS_b = 50 /3 = 16.66$

Step 5 is to calculate variance within samples

$MS_w = SS_w / df_w$

Df_w is n-k. here it will be 20-4=16

$MS_w = 208 / 16 = 13$

Lastly, we calculate f statistic for ANOVA using formula

$F= MS_b/ Ms_W = 16.66/13 = 1.28$

F value has two types of degrees of freedom. Df numerator is same as df for between groups variance and df denominator is same as df for within groups variance.

Thus, the summary table for ANOVA will be:

Source	SS	df	MS	F	p
Between	5	3	16.66	1.28	>0.05
Within	208	16	13		
Total	258	19			

Thus, null hypothesis is not rejected.

2) In an experiment studying levels of processing model, five participants were shown words at shallow, phonemic and semantic levels of processing. Once all the words were shown randomly the

participants were asked to free recall whichever words they remembered. Calculate appropriate statistics to find out whether levels of processing has any effect on recall.

Participant	Level 1	Level 2	Level 3
1	30	28	16
2	14	18	10
3	24	20	18
4	38	34	20
5	26	28	14

Note:The above data is taken from https://www.statology.org/repeated-measures-anova-by-hand/

Here, as there are only 5 participants who have gone through all the levels of IV, appropriate inferential statistics would be Repeated measures one way ANOVA.

First, we will calculate the total sum of squares (SST)

$SS_T = \Sigma(\overline{X}_{ij} - \overline{X})^2$

$SS_T = 899.7$

Second step is to compute SS_b, like previous sum

SS between is also known as SS treatment

$SS_b = n\Sigma(\overline{X}_i - \overline{X})^2$

$SS_b = 362.1$

Next step is to calculate sum of squares for subjects

$\overline{X}_{sub}$ is obtained by computing mean of all the scores of one particular participant

$SS_{sub} = k\Sigma(\overline{X}_{sub} - \overline{X})^2$

$SS_{sub} = 441.1$

Then we have to compute sum of squares for error term

$SS_{error} = SS_T - (SS_{treat} + SS_{sub})$

$SS_{error} = 96.5$

Further, MS for treatment and error can be computed by using formula

$MS_b = SS_b / df_b$.....computation same as previous sum

$MS_{error} = SS_{error}/df_{error}$ and $df_{error} = df_b \times df_{sub}$

Lastly, f value is obtained by dividing MS_b with MS_{error}

Summary table:

Source	Sum of Squares (SS)	df	Mean Squares (MS)	F	p
Between	362.1	2	181.1	15.006	<0.05
Subject	441.1	4	110.3		
Error	96.5	8	12.1		

As F critical value is 4.45, the null hypothesis can be rejected.

3.4 POST HOC TESTS

ANOVA tests let us know whether the three means differ or not. However if the researcher is interested in knowing pairwise comparison then a procedure called as post hoc tests. Tucky's HSD, Bonferroni Dunn test, Studentized range statistics, etc.

3.5 TWO-WAY ANOVA: INDEPENDENT SAMPLES

Essentially, ANOVA test is used to study differences between more than two groups. One of the ways in which this could occur is when the study has more than one independent variable(IV) with each IV having at least two levels of its own. In such cases Two way ANOVA is used. One of the benefits of using two way ANOVA is that it provides us with two effects: main effect and interaction effect. Main effect is the effect of each individual IV on dependent variable (DV). Interaction effect is effect of a particular level of IV 1 when combined with levels of IV2. Interaction effect is absent when impact of one IV on DV is constant for all levels of another IV

There are three sub types of two way ANOVA:

- Two way completely randomized ANOVA

  - Used when all the IVs in the study have randomized design

- Two way completely repeated ANOVA

  - Used when all the IVs in the study have repeated measures design

- Two way mixed ANOVA

  - Used when one IV has randomized design and another has repeated design.

Two way ANOVA follows same assumptions as one way ANOVA

Lets solve an example to understand Two way ANOVA better.

A teacher wanted to know whether time of the lecture and different teachers have significant effect on students learning or not. She selected a total of 40 students. Half of them attended lectures in the morning whereas half of them attended lectures in the afternoon. Within 20 students who attended morning lectures, 5 students were taught by 4 different teachers each. Same was done with afternoon batch as well. Following data was obtained:

Lecture time	**Type of teacher**			
	A	B	C	D
Morning lectures	4.8	5	6.4	6.3
	4.4	5.2	6.2	6.4
	3.2	5.6	4.7	5.6
	3.9	4.3	5.5	4.8
	4.4	4.8	5.8	5.8
Evening lectures	4.4	4.9	5.8	6
	4.2	5.3	6.2	4.9
	3.8	5.7	6.3	4.6
	3.7	5.4	6.5	5.6
	3.9	4.8	5.5	5.5

Note: The above data is taken from https://www.statology.org/two-way-anova-by-hand/

Note: for repeated terms such as SS_w, SS_T etc. , same formula that is written on above solved sums of one ANOVA to be used

Step 1 is to calculate sum of squares for each IV

Grand mean = 5.15

For calculating SS of IV1,

Mean of scores for morning lectures= 5.15

Mean of scores for evening lectures = 5.15

We can compute SS by using formula

$\Sigma n(\overline{X}j - \overline{X})^2$

= 0.00025

Similarly for IV2,

Mean of scores for Teacher A = 4.07

Mean of scores for Teacher B = 5.1

Mean of scores for Teacher C = 5.8

Mean of scores for Teacher D = 5.5

SS for IV2 by using same formula would be 18.76

Next step is to compute sum of squares within

SS for morning lectures with teacher A: 1.5

SS for morning lectures with teacher B: 0.9

SS for morning lectures with teacher C: 1.7

SS for morning lectures with teacher D: 1.6

SS for afternoon lectures with teacher A:0.3

SS for afternoon lectures with teacher B: 0.5

SS for afternoon lectures with teacher C: 0.6

SS for afternoon lectures with teacher D:1.2

Adding all the above , sum of squares within turns out to be 8.68

Next step is to calculate total sum of squares which is 28.45

Calculate sum squares interaction

$SS_{Interaction} = SS_T - SS_{IV1} - SS_{IV2} - SS_W$

$SS_{Interaction} = 1.01$

df of IV1 = 2-1= 1

df of IV2 = 4-1= 3

df interaction = (j-1)×(k-1) = 3

df within = n - (j*k) = 40 – (2*4) = 32

df total: n-1 = 40-1 = 39

MS = SS/ df

Source of variation	**SS**	**df**	**MS**	**F**	**P value**
Lecture timing	0.00025	1	0.00025	0.0009	NS
Type of teacher	18.76	3	6.25	23.04	>0.05
Interaction effect	1.01	3	0.33	1.24	NS
Total	28.45	39	0.72	1.24	

3.6 SUMMARY

- Inferential statistical tests help us generalize results from sample to population.
- t test is used to find out differences between two groups. It is parametric test that can be used when DV lies on interval or ratio scale.
- Independent samples t test is used when design of the study is randomized groups design , dependent samples t test is used when study has repeated measures design.

- ANOVA is used to find out differences between more than two groups. When it is used for One IV having more than two groups, One way ANOVA is used. Depending on the design of the study one may either use independent samples ANOVA or dependent samples ANOVA.
- When one wants to study pair wise comparisons a procedure called post hoc test is used.
- When study has more than one IVs with each having at least two levels each, one uses Two way ANOVA.
- Two ANOVA provides main effect which is effect of each IV on DV and interaction effect is effect of one IV when combined with level of another IV on DV.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Figure 3.2: Critical values table for t test

Image source: Google images
Site: http://www.ttable.org/

F-table of Critical Values of α = 0.10 for F(df1, df2)																		
	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Figure 3.3: Critical values table for F test

Image source: Google images
Site:https://statisticsbyjim.com/hypothesis-testing/f-table/

3.7 QUESTIONS

1) A researcher wanted to study if males and females differ with respect to their leadership abilities. Following data was collected. Draw appropriate conclusions.

M	F
128	127
118	115
144	142
133	140
132	131
111	132
149	122
139	119
136	129
126	128

Note: Above data is taken from
https://www.statsdirect.com/help/parametric_methods/single_sample_t.htm

2) Three types of treatments are used on three groups of clients for 5 weeks. We want to check if there is a difference in the mean growth

of each group. Using the data given below apply a one way ANOVA test at 0.05 significant level.

Treatment 1	Treatment 2	Treatment 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Note: Above data is taken from https://www.cuemath.com/anova-formula/

3) Write a note on t test

4) Write a note ANOVA test

3.8 REFERENCES

- Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.
- Belhekar, V. M. (2016). Statistics for Psychology Using R. SAGE publications.
- Mangal, S. K. (2002). Statistics in psycholohy and education. PHI Learning Pvt. Ltd
- https://vitalflux.com/two-sample-t-test-formula-examples/
- https://247amend.com/2016/03/paired-sample-t-test-illustration.html
- http://faculty.webster.edu/woolflm/ttest.html
- www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf
- https://atozmath.com/example/CONM/anova.aspx?he=e&q=anova1
- https://www.statology.org/repeated-measures-anova-by-hand/
- https://www.statology.org/two-way-anova-by-hand/
- http://www.ttable.org/
- https://statisticsbyjim.com/hypothesis-testing/f-table/
- https://www.statsdirect.com/help/parametric_methods/single_sample_t.htm
- https://www.cuemath.com/anova-formula/

❄❄❄❄❄❄❄

4

INFERENTIAL STATISTICS: INFERENCE ABOUT LOCATION - II

Unit Structure:

4.0 OBJECTIVES

After studying this unit the students will be able to:

- Understand difference between parametric and non parametric tests
- Understand application of most useful non parametric tests
- Understand Multivariate Analysis of Variance
- Get introduced to differential functional analysis

4.1 NON PARAMETRIC TESTS

Statistical tests can broadly be divided into two categories- descriptive statistics and inferential statistics. Descriptive statistics as the name suggests describes sample characteristics whereas inferential statistics allows researchers to generalize findings of sample to the population. Inferential statistics can further be divided into parametric and non parametric tests.

Parametric tests are used when assumptions of normality and homogeneity of variances is met, dependent variable lies on interval or ratio scale and sample size is large. Examples of parametric tests include t test, ANOVA, Pearson Product Moment correlation, etc. Non parametric tests are also called as distribution free tests as they make no assumption about nature of population from where samples are drawn. The non parametric tests are used when:

1. Sample is small

2. Assumption of normality and equality of variances is not met

3. Data lies on nominal and ordinal scale

Examples of non parametric tests include Wilcoxon sign-rank test, Median test, U test, Kruskal-Wallis test, Kendall's Tau etc. Most of the non parametric tests are based on median rather than mean. For the tests dealing with ordinal ranks, the data is translated into ranks on which then tests are carried out. They asses whether median of two or more populations are equal or not.

Parametric tests	**Non parametric tests**
Used when DV lies on interval or ra tio scale	Used when DV lies on nominal or ordinal scale
Assumption of normality needs to b e met	No assumptions required
Assumption of homogeneity of vari ances needs to be met	No assumption required
Require large sample	can be used with small samples
More powerful	Less powerful
Examples: t test, ANOVA, Pearson product moment correlation, etc.	Examples: chi square, median t est, mann whitney U test, etc

Figure 4.1: Parametric vs non parametric tests

4.1.1 Wilcoxon sign-rank test

Wilcoxon sign rank test is used when study has one independent variable with two levels and the design of the study is matched groups. In matched groups design participants are screened for some ability such as intelligence, working memory and matched.

Let us understand its procedure through an example. Lets say two teachers have marked students on their level of classroom participation. Following data has been received.

Participant number	**Teacher A**	**Teacher B**	**Difference**	**Rank**
1	42	39	+3	3
2	40	36	+4	4
3	63	40	+23	5
4	61	67	-6	2
5	58	61	-3	1

Once the data is collected, the difference between the two sets of data is computed. This difference is then ranked with least difference receiving rank of one and highest difference getting highest rank. If difference between two scores is zero then it carries no rank. Ranks are added with their respective direction of the difference. Positive and negative ranks are then separated and their respective total is obtained. Summation of all the positive ranks is indicated by T+ and summation of all the negative ranks is

indicated ad T-. Thus, the test takes into account the direction of the difference as well as rank of the difference.

It uses statistic T, which is smaller of the two sums : T+ and T-. Significance of T is then calculated by comparing it to critical T value. If the obtained value or T is equal to or smaller than critical value for T, the null hypothesis is rejected. If the obtained T value is greater than critical T value, null hypothesis about difference between the groups is not rejected.

A test similar to Wilcoxon sign rank test is Sign test. It uses a test statistic of r. It considers only the number of signs in one direction. It operates on the assumption that if the two groups are truly equal in their means then the pairs of scores having difference in one direction will be approximately same as number of pairs having difference in another direction. For small N, binomial distribution is used to compute r. If value of r obtained from data is greater than critical r value, null hypothesis is rejected.

Researchers often prefer Wilcoxon sign rank test over sign test as it considers not only sign of the difference but magnitude of difference as well which represents true nature of the data. The parametric test of matched pairs t test corresponds to this non parametric test. The major difference is that t uses actual values whereas Wilcoxon sign rank test uses ranks of the given values.

4.1.2 Median test

When a researcher has to study difference between more than two means but requirements of non parametric tests are not met, median test is the appropriate statistic of choice. As the name suggests, this test is based on median rather than mean. In nutshell, it computes median of the groups being compared and counts number of frequencies above and below the median for all the groups. If the groups being compared are equal then they should have more or less equal number of frequencies in each half. Thus the null hypothesis of the test states that there will be no difference in the amount of frequencies above and below the mean across all the groups whereas alternative hypothesis states that frequencies above and below the median across all groups will be non equivalent.

Unlike sign test or Wilcoxon sign-rank test, median test does not use ranks but uses rather proportion of frequencies. Hence it uses chi square to check significance of the data received. Chi square is another useful non parametric test of goodness of fit which compares observed frequencies of the data against expected frequencies. A general formula for chi square test is:

$$\chi^2 = \frac{\Sigma (f_0 - f_e)^2}{f_e}$$

Let understand median test in detail with the help of an example.

A teacher has scored students of class A and B on their sincerity level. Use median test and draw appropriate conclusion using following data.

"Scores of class A:
79,86,40,50,75,38,70,73,50,40,20,80,55,61,50,80,60,30,70,50

Scores of class B:

85,80,50,55,65,50,63,75,55,45,30,85,65,80,55,75,65,50,75,62"

Data source:
https://atozmath.com/example/CONM/NonParaTest.aspx?he=e&q=mt

Step 1: stating null and alternative hypothesis

Null hypothesis: there is no significant difference in the sincerity levels of class A and B

Alternative hypothesis: Class A and B differ significantly with respect to their sincerity levels.

Step 2 : is to rank all the scores of two groups. Rank of 1 is assigned to the highest value and so on.

Step 3: Once all the scores across both the groups are ranked, grans median is computed by using all the scores from both the groups. In this case it turns out to be 61.5.

Step 4: Is to compute chi square statistic. Here observed frequencies f_o will be number of frequencies obtained in both the groups above and below the mean as it is. Expected frequencies f_e will be equal number of frequencies above and below the mean for the two groups. A table similar to this can be formulated :

		Class A	**Class B**
Above median	f_o	8	12
	f_e	10	10
Below median	f_o	12	8
	f_e	10	10

Step 5: By using above mentioned formula, calculated chi square value turns out to be 0.9

Step 6 : Like any other test of inferential statistics, once the statistical value is computed we have to calculate degrees of freedom in order to test the hypothesis

df of chi square = (C-1)×(r-1) = 1……..c is number of columns and r is number of rows.

Step 7: compare obtained chi square value to critical chi square value from the table. Critical value in this case is 3.84

Step 8: Draw conclusion

As calculated chi square value is less than critical value in this data, we will not reject the null hypothesis and conclude that class A and B are same with respect to their sincerity level.

4.1.3 U test

U test or Mann-Whitney U test is a preferred choice of statistics when the study has one IV with two levels and the design of the study is randomized groups design. It is equivalent to parametric test of independent samples t test. On the same lines as independent samples t tests, U test also allows us to identify difference between two groups however it does not use mean to compute this difference, Rather it compares medians of the two groups to draw the conclusion.

Like previously described non parametric tests, U test also starts with ranking the given data. The values from the two groups are combined and ranked with lowest value getting rank of 1 and so on. For example, lets say an employer has scored 6 male and 9 female employees on their level of punctuality and wants to know whether the two groups differ significantly. In this case, the scores of all the 15 employees will be ranked with rank of 1 given to the least punctual employee and rank of 15 to the most punctual one. Ranks of the two groups are then summed up and denoted as T_1 and T_2 for each group respectively.

Once we obtain T_1 and T_2 they are transformed into U by using formula :

$U_1 = N_1+N_2+[N_1(N_1+1)]^2-T_1$

$U_2 = N_1+N_2+[N_2(N_2+1)]^2-T_2$

Now, one should ensure that

$U_1 + U_2 = N_1 \times N_2]$

Lastly, the smaller of the U value is considered to be calculated U statistic which is compared to critical U value. If Calculated U is lower than critical U, the null hypothesis is rejected.

4.1.4: Kruskal-Wallis test

When a researcher wants to conduct one way randomized ANOVA but assumptions of parametric tests are not met, researcher can use Kruskar Wallis test. It is used when study has one IV with more than two levels and the design of the study is randomized groups design.

This test shares similarities with U test wherein it also deals with ranks and sum of ranks. However the test statistic used in Kruskal-Wallis test

Is H. Formula for H is:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Figure: 4.2 Formula for H statistic

Image source: Google images
Site: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kruskal-wallis/

Here,

N is total number of participants.

T_j is sum of the ranks in one group

N_i is number of participants in each group.

H statistic is based on chi square distribution. Degrees of freedom is calculated as k-1 where k is number of groups or levels of IV being compared. The null hypothesis of this test states that all the groups being compared come from the populations that have equal median whereas alternative hypothesis states that the medians of the populations from where samples are drawn are different.

Kruskal Wallis test is less powerful than ANOVA in the sense it requires differences between the groups to be larger for test statistic to reject the null hypothesis.

4.2 MANOVA

Multiple Analysis of Variances (MANOVA) is an extension of ANOVA that is used when study has more than one dependent variable(DV) and the DVs are inter correlated. For example, if a researcher wanted to study whether different types of exercises - yoga, cardio and stretching have different effects on mood and general well being, he/she may use MANOVA. Since it is a parametric test, it can be used only when DV lies on interval or ratio scale and rest of the assumptions for parametric tests are met.

An alternative in such situations, when one wants to see effect of levels of IV on more than one DV is to use multiple t tests/ANOVAs. In the above example as well the researcher could have done two ANOVAS: one with mood as DV and another with general well being as DV. However, such analysis increases chance of type I error. Such an analysis also gives two separate pieces of information rather than one comprehensive picture. On the other hand MANOVA studies differences between groups using a linear combination of DVs, thereby guarding against increased type I error. Once the evaluation and turned out to be significant, it also allows researchers to asses which which individual DVs are significantly different across levels and which ones are not.

The null hypothesis of MANOVA test states that all populations being studied have equal vector of means whereas alternative hypothesis claims a difference between population means.

Assumptions underlying MANOVA:

1) Multivariate normal distribution

Similar to ANOVA's assumption of normality, even MANOVA expects distribution of population means for all the DVs to be equal. It can be tested via tests of normality such as Mardia's MVN test, Royston's MVN test, etc. Eliminating outliers is another way of boosting normality. However one should be careful to eliminate multivariate and not univariate outliers in this regard.

2) Homogeneity of variance-covariance

An extenstion of ANOVA's assumption of homogeneity of variances is MANOVA's population variance-covariance matrix. The matrix of population variances of DVs and the matrix of covariance among DVs is assumed to be equal. These matrices are also used in actual calculation of test statistic.

3) Multicollinearity

A complete linear relationship of DVs among each other is discouraged in MANOVA. As a rule of thumb, a correlation above 0.8 is considered to be an indicator of multicollinearity. At the same time, 0 relationship among DVs is also not desired as it will obstruct forming linear combination of DVs which is needed for analysis

4) Linearity

DVs should be linearly related to each other. There should not be any curvilinearity involved. For example, time of the day and sunlight are not linearly related as although initially as time of the day progresses sunlight also increases after afternoon even though time increases, sunlight decreases.

5) Homogeneity of regressions

The regressions among all the DVs are assumed to be equal

6) Sample size

As a rule of thumb, sample size is expected to be 6-10 times the number of DVs in each group. Greater sample size leads to greater power.

Thus, MANOVA is a useful statistic to be used when dealing with more than one DV in the same study.

4.3 DISCRIMINANT FUNCTIONAL ANALYSIS (DFA)

Discriminant functional analysis is related to the family of tests in ANOVA and regression. It helps to predict membership of participant in a particular group. It uses combination of variables to predict group membership. It is a data reduction technique. In a typical study using DFA, a linear combination of variables that maximize differences between groups is used.

In DFA analysis, IV is continuous and DV is categorical. It is also used as a follow up for MANOVA. For example, DFA analysis can help predict if based on a bunch of symptoms an individual is likely to be diagnosed with ADHD or not. In other words, if the person belongs to the group of those having ADHD or not. The two types of DFA include descriptive investigations and predictive investigations. When a set of independent variables predict outcome variable, there is assumed to be an underlying dimension which can be identified using linear combination of DVs called as 'variate'. Linear combination of these variate is then used to predict different groups membership. That is why this analysis is called as 'discriminant functional analysis'.

DFA uses maximization principle to analyze data. It involves deriving ratio of systematic to asystematic variance. The three types of DFA include Direct DFA, Hierarchical DFA and Stepwise DFA. In direct DFA, all the predictor variables are simultaneously added to the equation and only those contributing significantly to the function are retained. In hierarchical DFA, variables are entered in a particular order based on some theory. Similarly in step wise DFA, variables are entered in a particular order but based on some statistical criteria. Formula for DFA :

$D= a+ v_1X_1+v_2X_2+v_3X_3+\ldots\ldots+v_nX_n$

Here,

D= discriminant function

A= equation constant

X_1= participants score on first predictor

V = weight given to the predictor variable

Assumptions of DFA:

1) Multi normality

Like parametric tests, DFA requires normality among all the populations of all the variables being studied. Studying large samples is one way of protecting against non normality.

2) Homogeneity of variance-covariance matrices

The variance and covariance must be equally distributed across all the groups. Outliers must be weeded out as they pose a threat to this assumption.

3) Multicollinearity

Unlike MANOVA, DFA prefers singularity over multicollinearity. This means that a linear relationship or inter dependence among independent variables is to be avoided.

4) Scale of measurement

The data should be lying on interval or ratio scale.

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
5	--	--	--	--	--	0	2
6	--	--	--	--	0	2	3
7	--	--	--	0	2	3	5
8	--	--	0	2	3	5	8
9	--	0	1	3	5	8	10
10	--	1	3	5	8	10	14
11	0	3	5	8	10	13	17
12	1	5	7	10	13	17	21
13	2	7	9	13	17	21	26
14	4	9	12	17	21	25	31
15	6	12	15	20	25	30	36
16	8	15	19	25	29	35	42
17	11	19	23	29	34	41	48
18	14	23	27	34	40	47	55
19	18	27	32	39	46	53	62
20	21	32	37	45	52	60	69
21	25	37	42	51	58	67	77
22	30	42	48	57	65	75	86
23	35	48	54	64	73	83	94
24	40	54	61	72	81	91	104
25	45	60	68	79	89	100	113
26	51	67	75	87	98	110	124
27	57	74	83	96	107	119	134

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
28	64	82	91	105	116	130	145
29	71	90	100	114	126	140	157
30	78	98	109	124	137	151	169
31	86	107	118	134	147	163	181
32	94	116	128	144	159	175	194
33	102	126	138	155	170	187	207
34	111	136	148	167	182	200	221
35	120	146	159	178	195	213	235
36	130	157	171	191	208	227	250
37	140	168	182	203	221	241	265
38	150	180	194	216	235	256	281
39	161	192	207	230	249	271	297
40	172	204	220	244	264	286	313
41	183	217	233	258	279	302	330
42	195	230	247	273	294	319	348
43	207	244	261	288	310	336	365
44	220	258	276	303	327	353	384
45	233	272	291	319	343	371	402
46	246	287	307	336	361	389	422
47	260	302	322	353	378	407	441
48	274	318	339	370	396	426	462
49	289	334	355	388	415	446	482
50	304	350	373	406	434	466	503

Firgure 4.3: Critical values table for Kruskal Wallis test

Image source: Google images
Site: https://www.real-statistics.com/statistics-tables/wilcoxon-signed-ranks-table/

Table 3 Critical values of U (5% significance).

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Figure 4.4: Critical values table for Mann Whitney U test

Image source: Google images
Site: https://accendoreliability.com/mann-whitney-u-test/

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Figure 4.5: Critical values of chi square test

Image source: Google images
Site : https://passel2.unl.edu/view/lesson/9beaa382bf7e/8

4.4 QUESTIONS

1) Write a short note on MANOVA

2) What is differential functional analysis?

3) Differentiate between parametric and non parametric tests

4.5 REFERENCES

- Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.
- Belhekar, V. M. (2016). Statistics for Psychology Using R. SAGE publications
- Snodgrass, J. G., Levy-Berger, G., & Haydon, M. (1985). Human experimental psychology (Vol. 395). New York: Oxford University Press.
- https://atozmath.com/example/CONM/NonParaTest.aspx?he=e&q=mt
- https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kruskal-wallis/
- https://www.real-statistics.com/statistics-tables/wilcoxon-signed-ranks-table/
- https://accendoreliability.com/mann-whitney-u-test/
- https://passel2.unl.edu/view/lesson/9beaa382bf7e/8

❄❄❄❄❄❄❄

5

ASSOCIATION, PREDICTION AND OTHER METHODS - I

Unit Structure:

5.0 OBJECTIVES

After studying this unit, you will understand

a.) what is product moment correlation

b.) what is meant by partial correlation

c.) what are the other special type of correlations

d.) what is the use of linear regression and how it is computed.

5.1 INTRODUCTION

Very often, in psychological research, the psychologists are interested to find out the whether there is any association between two variables, and if yes, then what type of association it is. In this unit, we will be discussing tools for measuring this association or relationship between two variables, that have been measured by interval or ratio scales. Though in common language the words 'association' and 'relationship' appear to have same meaning but in statistics, their meanings are different. Howell (2002) explained that The word *correlation* is used to describe the situation in which both *X* and *Y* are random variables that measure the correlations. These are the variables where some sort of order can be assigned to each of the variables, e.g., anxiety, stress, while measures of association are those statistical procedures that are used for variables that do not have property of order. They are categorical or nominal variables, e.g., nationality, gender,

etc. A simple bivariate correlation analyses can measure the strength and direction of the relationship between two variables. The strength of the relationship is determined by looking at the extent to which the values of the two variables co-vary. The direction signifies whether the values of one variable tend to move in the same direction as the values of the second variable, or whether they tend to move in opposite direction.

Correlation analyses are used for a variety of purposes in research. For example, it is used in descriptive research to determine the strength of the relationship between two variables of interest with regard to the sample or population being studied. In cases where the correlation between two variables has been shown to be very strong, the value of one variable can be used to predict the value of the second variable. It is also used in the development of measurement instruments, and to calculate the reliability of measurement instruments. It can be used in twin studies, factor analysis and in the procedure of regression analysis.

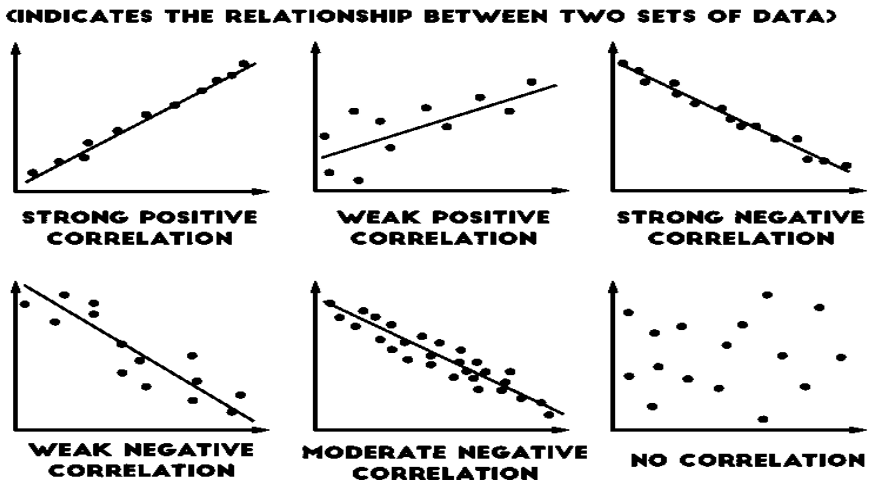
However, one must keep in mind that sampling weakness can artificially increase or decrease a correlation coefficient. Yet its usefulness cannot be undermined.

5.2 SCATTERGRAM

Scattergram or scatterplot is a graphical representation of the relationship between two variables. In Scattergram, each participant of the experiment is represented by a dot in two-dimensional space. The pairs of scores are plotted, one for variable X and one for variable Y. Generally, X variable is represented on the abscissa and is known as predictor or independent variable while variable Y is represented on ordinate or Y-axis and Y variable is known as criterion variable. Each individual (X,Y) pair is plotted as a single point.

In Fig.1 below, you will notice that a straight line passes through most of the dots. This line is known as regression line of Y predicted on X.

If the pairs of scores of two variables are not scattered randomly, instead form a consistent pattern, we can say that there is a relationship between the two variables. If the scores are scattered randomly, it means that there is no relationship or very weak relationship between the two variables. Scattergram can depict linear and non-linear (curvilinear) relationship. It can also indicate whether the relationship is positive or negative or no relationship. Positive relationships have points that incline upwards to the right. As *x* values increase, *y* values increase too. As *x* values decrease, *y* values decrease too. Negative relationships have points that decline downward to the right. As *x* values increase, *y* values decrease. As *x* values decrease, y values increase.



See Fig 1 **Scattergram of linear relationships**

Source: Wikipedia

Correlation would be called non linear or curvilinear if the amount of change in one variable does not bear a constant ratio with the amount of change in the other variable.

When graphed, a nonlinear relationship will NOT result in a straight line. The graph may be curved, U-shaped, or V-shaped. Nonlinear relationships do not have a constant rate of change. Various types of curvilinear relationships cubic, quadratic, polynomial, exponential, etc.

For example, see fig 2

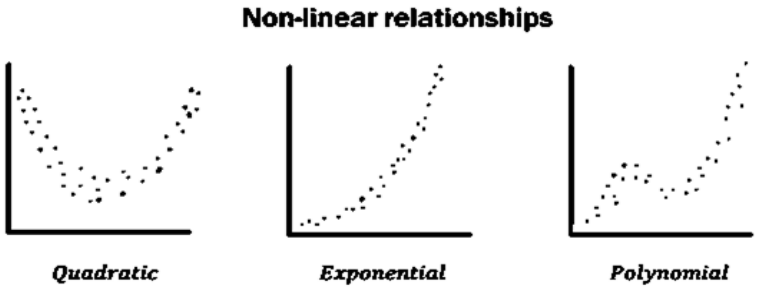
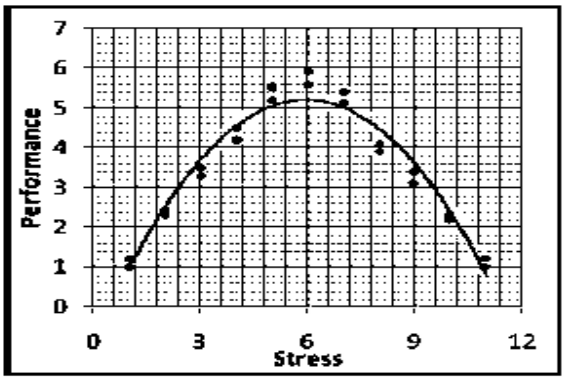


Fig.2 Non-linear Relationships

Two variables may have a strong, or even perfect, relationship, yet the relationship is not at all linear. The correlation coefficient might be zero. For example see fig 2



i: Typical relationship between stress and performance. The performance is poor at extremes and improves with moderate stress. This is one type of curvilinear relationship.

5.3 THE COVARIANCE

The covariance is the number that reflects the degree to which two variables vary together. It is denoted as

cov_{xy}

The covariance between X and Y (or *S XY*) can be stated as

$$Cov_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{N-1}$$

Looking at the formula of covariance, it is obvious that it is similar to formula of variance. In case of positive relationship between two variables, the covariance will be positive and in case of negative relationship, the covariance of two variables will also be negative. If there is no relationship between two variables, the products of the deviations will be positive for half of the scores and negative for the other half of the scores. Thus their sum of products will be zero and the covariance will indicate no relationship.

The Correlation Coefficient :

While Scattergrams provide a way of displaying the relationship between two variables, they can only provide an estimation of the strength of the relationship. In order to calculate the exact strength and statistical significance of the relationship, the correlation coefficient must be determined. The correlation coefficient is a descriptive statistic, like the mean or standard deviation. The correlation coefficient describes the linear relationship between two variables. The correlation coefficient provides a numerical indicator of both the strength and the direction of the relationship between the paired values of two interval or ratio-level variables. Because the correlation coefficient provides information about the strength or magnitude of the relationship between the variables, it is considered to be a measure of effect size.

5.4 PRODUCT MOMENT CORRELATION

Correlation analyses can be used to measure relationship between two variables or even more than two variables. When it is measuring relationship between two variables, it is called bivariate correlation. Pearson's Product Moment Correlation is an example of bivariate correlation.

If the amount of change in one variable is constantly proportional to the change in other variable then the correlation is said to be linear.

If the quantities of one variable increases, the quantity of other variable either increases or decreases at a constant rate.

When graphed, a linear relationship will create a straight line. There will be no curve or bend in the line. Linear relationships have a constant rate of change (slope).

it is called 'Product Moment Coefficient of Correlation'. It is a parametric test and is denoted by the letter 'r'. Pearson's *r* only considers linear correlations between variables.

Conditions for Pearson's Product Moment Correlation Coefficient –

It is assumed that -

1. The variables X and Y are continuous variables.
2. The data is collected from truly random sample that is representative of population of interest.
3. The data contains paired scores. Each subject must have both variable X and Variable Y scores.
4. Each subject's scores on both variables are independent of scores of others on those two variables.
5. The data of these two variables is in normal distribution
6. There is homogeneity among the two variables.
7. There is linear relationship between the two variables.
8. There are no outliers in the data.

Outliers are extreme score on one or both the variables. Outliers may be errors. The strength of the correlation gets affected by the presence of outlier, as outliers can greatly influence the sample mean and variance. See Fig. 4

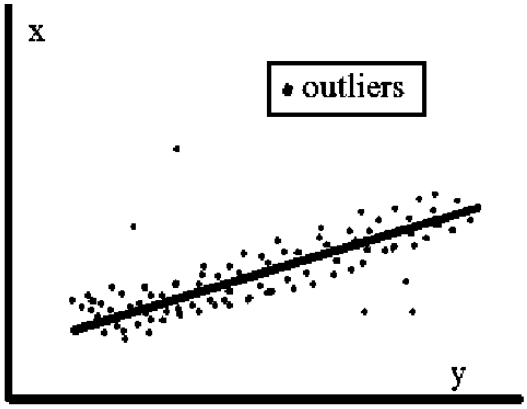


Fig.4

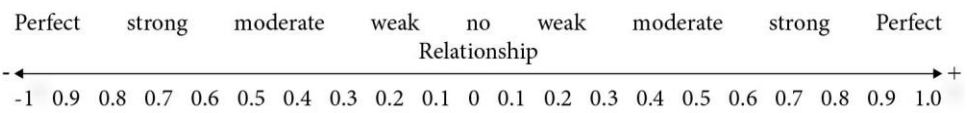
(Source: Image taken from Modi, K., & Oza, P.B. (2017). Outlier Analysis Approaches in Data Mining.)

The values of the Pearson's *r* correlation coefficient, expressed as *r,* can range from − 1.0, signifying a perfect negative linear correlation, to +1.0, signifying a perfect positive linear correlation. It is not possible to obtain a value less than -1.0 and more than +1.0 A correlation coefficient of 0 using Pearson's *r* indicates no linear correlation or relationship between the variables at all. The plus sign is usually not displayed before a positive correlation coefficient. The positive nature of the correlation is implied. The closer the value of the coefficient is to +1.0, the stronger the positive correlation between the two variables. The closer the value is to − 1.0, the stronger the negative correlation. The closer the value is to 0, the weaker the correlation between the variables. One must understand that the negative correlations are not inherently weaker than positive correlations. It simply indicates that the values co-vary in opposite directions rather than in the same direction.

Cohen (1988) has developed a system for interpreting the strength of a linear correlation based on the Pearson's *r* correlation coefficient. He suggests that a correlation coefficient greater than or equal to .1 but less than .3 is a weak correlation, a correlation greater than or equal to .3 but less than .5 is a moderate correlation, and a correlation greater than or equal to .5 is a strong correlation.

See fig 3

Perfect strong moderate weak no weak moderate strong Perfect



It is important to point out that even a very strong correlation between two variables does not indicate that the values of one of the variables actually *cause* the values of the second variable. "correlation does not equal causation." If, however, it is determined that one variable has a causal

relationship with another variable, these variables must also be correlated with one another.

Determining statistical significance

Sample size is an important factor when determining the statistical significance of a correlation. A strong correlation of, say, .62, might not be statistically significant with a small sample size, because sampling error may be able to explain the relationship. On the other hand, a weak correlation, for example .25, might be statistically significant with a large sample size, because although the correlation is a weak one, the relationship between the variables could not be explained by sampling error. So, one need to look at the strength and direction of a correlation, in addition to its statistical significance, when interpreting the outcome of a correlation analysis.

Computation of Pearson's Product Moment Correlation

Example – Suppose we want to calculate the strength and direction of the correlation between the intelligence and the marks obtained in statistics by a group of college students. There are two methods of computing Product Moment Correlation.

a) Through deviations

b) through raw scores only

a) The formula given below is for computing correlation through deviations.

Formula – $\mathbf{r} = \frac{\Sigma xy}{n\,Sx\,Sy}$

Let us take a numerical example for using this formula

	x	y	x-$\bar{x}$ = x	Y – $\bar{y} = y$	(x*x)= x^2	y^2	xy
A	2	10	2-5=-3	10-6=4	9	16	-12
B	3	7	3-5=-2	7-6=1	4	1	-2
C	4	8	-1	8-6=2	1	4	-2
D	7	2	2	-4	4	16	-8
E	8	3	3	-3	9	9	-9
F	9	1	4	-5	16	25	-20
G	2	10	-3	4	9	16	-12
H	3	10	-2	4	4	16	-8
I	4	7	-1	1	1	1	-1
J	8	2	3	-4	9	16	-12
	Σ=50	Σ= 60			Σ=66	Σ=120	Σ= -86
N =	10	10					
	$\bar{x} = 5$	$\bar{y} = 6$					

$$\boldsymbol{Sx} = \sqrt{\frac{x^2}{n}} = \sqrt{\frac{66}{10}} = \mathbf{2 \cdot 57}$$

$$\text{Sy} = \sqrt{\frac{y^2}{n} = \frac{120}{10}} = 3.46$$

Formula

$$\mathbf{r} = \frac{Covxy}{\mathbf{n}\, Sx\, Sy} \qquad \mathbf{r} = \frac{-86}{10(2.57 * 3.46)} = \frac{-86}{88.92} = \mathbf{-0.967}$$

The results can be interpreted by saying that there is a strong but negative relationship between variable X and Y.

The correlation coefficient is simply a point on the scale between -1 and +1, and the closer it is to either of those limits, the stronger is the relationship between the two variables.

b.) Formula for computation of correlation with raw scores –

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n} \cdot \sum y^2 - \frac{(\sum y)^2}{n}}}$$

Let us take a numerical example – Suppose a researcher is trying to find the relationship between social skills and stress of college students. Both are continuous variables and normally distributed.

Students	Social skills (X)	Stress (Y)	X^2	Y^2	XY
1	10	6	100	36	60
2	20	2	400	4	40
3	30	4	900	16	120
4	40	10	1600	100	400
5	50	12	2500	144	600
6	60	8	3600	64	480
Total	210	42	9100	364	1700

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n} \cdot \sum y^2 - \frac{(\sum y)^2}{n}}}$$

$$\text{r} = \frac{1700 - \frac{(210)(42)}{6}}{\sqrt{\left(9100 - \frac{(210)^2}{6}\right)\left(364 - \frac{(42)^2}{6}\right)}} = \frac{230}{350} = 0.657$$

The result can be interpreted by saying that there is a positive but moderate relationship between the two variables.

Adjusted r

Pearson's correlation coefficient that is calculated on the sample is not an unbiased estimate of population coefficient(*p)*. The smaller the sample size, more biased the estimate of population will be. So, we need to adjust the correlation coefficient to reduce this bias. The formula for adjusted r is

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}}$$

Applying this formula to above example of correlation of coefficient of raw scores

$$r_{adj} = \sqrt{1 - \frac{(1-.657^2)(6-1)}{6-2}} = .842$$

The results indicate that it is a relatively biased estimate of the population correlation coefficient. In our example, the sample size is very small (N=6), therefore there is a big difference between r and $_{radj}$.

5.5 PARTIAL CORRELATION

Partial correlation is a linear relationship between two random variables, after excluding /controlling the effect of one or more independent variables. The correlation coefficient of two variables will give misleading results if there is another, confounding, variable that is numerically related to both variables. To avoid that partial correlation is used.

Example

Researcher is interested to compute correlation between anxiety and achievement by controlling the impact of intelligence.
To compute partial correlation, we need data on all 3 variables.

Formulae for Partial Correlation

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{13.2} = \frac{r_{13} - (r_{12})(r_{23})}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{23.1} = \frac{r_{23} - (r_{12})(r_{13})}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

Order of Partial Correlation:

1. **The first order partial correlation,** $r_{12.3}$ indicates the correlation between first and second variable while partialling out the third variable. For example, a researcher is interested to find out the relationship between academic achievement motivation and emotional intelligence by controlling the impact of parental pressure.

2. **The second order correlation,** $r_{12.34}$ indicates relationship between first and second variable by partialling out the third and fourth variable. Partial correlation can be computed with not only three variables but many more.

3. **The third order partial correlation ,** $r_{12.345}$ indicates that correlation is computed between first and second variable while partialling out the third, fourth and fifth variables. .

The number to the right of the decimal point are the variables whose effect is controlled or ruled out and the number to the left of the decimal point are the variable whose relationship the researcher is interested in. One cannot compute second order partial correlation without first computing the first order correlation.

Computation of Partial Correlation: Suppose we want to find out the relationships between three variables – (1.) academic achievement motivation (2.) emotional intelligence (3.) parental pressure. The correlation coefficient values are as under

r_{12} = .90 , r_{23} = .60 , r_{13} = .50

$$r_{12.3} = \frac{.90-(.50)(.60)}{\sqrt{1-.50^2}\sqrt{1-.60^2}} = .86$$

$$r_{13.2} = \frac{.50-(.90)(.60)}{\sqrt{1-.90^2}\sqrt{1-.60^2}} = -.05$$

$$r_{23.1} = \frac{.60-(.90)(.50)}{\sqrt{1-.90^2}\sqrt{1-.50^2}} = .18$$

The results indicate that there is a strong relationship between academic achievement motivation and emotional intelligence if we keep parental pressure constant. There is negative and very weak relationship between academic achievement motivation and parental pressure if we keep emotional intelligence constant. Negative relationship is indicating that as the parental pressure goes up, the academic achievement motivation comes down. There is weak but positive relationship between emotional intelligence and emotional intelligence if we keep academic achievement motivation constant.

Assumptions of Partial Correlation:

1. All variables used are continuous variables
2. Scores of variables have unimodal or have fairly symmetrical distribution without any significant skewness.
3. The paired scores of each pair of variables are independent of all other scores in the sample.
4. There is a linear relationship between the scores of every pair.
5. There are no outliers.
6. There can be one or more variables that one wants to control.

5.6 SPECIAL CORRELATIONS

We need to use special kind of correlation methods when the data does not meet the preconditions of Pearson's product moment correlation. In other words, we use special types of correlations when any of the following characteristics are there in the data–

a) the data is not normally distributed

b) Both the variables are not the continuous variables.

c) there is no homogeneity among the variables.

d) there is no linear relationship between the variables.

Some of the special types of correlations are variations of Pearson correlation and some are non-Pearson correlations. Special type Pearson correlations are Point-Biserial Correlation and Phi coefficient while the non-Pearson type correlations are Biserial and Tetrachoric correlations. Now let us look at each one of them.

Point Biserial Correlation (r_{pb})

Point Biserial Correlation (r_{pb}) is a Pearson's Product moment correlation between one truly dichotomous variable and the other variable being continuous variable. Algebraically, the rpb = r. Point biserial correlation is calculated in similar manner as Product Moment Correlation. It can be denoted as rpb. it is simply Pearson's r applied to a special kind of data. It is not necessary to have equal number of cases in both categories, e.g., we can have 5 male and 6 female, or 5 black male and 8 white male, etc.

Assumptions for Using Point Biserial Correlation

1. One of the two variables should be measured on a **continuous** scale. Examples of **continuous variables** include intelligence, exam performance (measured from 0 to 100), weight (measured in kg), etc. This continuous scale should be either interval scale or ratio scale.

2. The other variable should be **dichotomous.** Examples of **dichotomous variables** include gender (two groups: male or female), employment status (two groups: employed or unemployed), athlete (two groups: yes or no),etc. There should not be any underlying continuum between the groups. In other words, dichotomy should not be artificially created in a continuous data.

3. There should be no outliers for the continuous variable for each category of the dichotomous variable.

4. The continuous variable should be **approximately normally distributed** for each category of the dichotomous variable.

5. The continuous variable should have **equal variances** for each category of the dichotomous variable.

Example - X Y

Test Score	Gender
90	1
85	1
78	1
93	1
79	1
85	0
79	0
89	0
90	0
95	0

Males ("1") tend to score lower than females ("0")

Calculation of Point Biserial Correlation (r_{pb})

Let us understand the calculation of point biserial through an example. Let us take a data of 20 college students, out of which 9 are male and 11 are female. We obtain their scores on a stress test and then we want to see whether there is any relationship between sex and stress. Sex is dichotomous- male and female, stress is a continuous variable.

Step 1 – Code the dichotomous variable, e.g., we code males as '0' and female as '1'.

Step 2 – Compute mean and standard deviation for group X and Y

Step 3 - Apply the formula

$r_{pbis} = \frac{Mp-Mq}{\sigma}\sqrt{pq}$

M_p = mean of category p

M_q = mean of the category q

• = the proportion of the sample in first group denoted as category p

q = the proportion of the sample in second group denoted as category q.

σ = is the standard deviation of the entire sample.

Let us take a numerical example of point biserial correlation.

Table 1: Data showing the sex and scores of stress test for 20 college students. (**Female is coded as 1, Male is coded as 0)**

Sr. no.	**Male(p) Female(q)**	**Scores of Stress (X)**	**X- $\bar{x}_{total}$**	**(X- $\bar{x}_{total}$)^2**
1.	0	6	6-8.06 =-2.06	4.24
2.	0	10	10-8.06 =1.94	3.76
3.	0	8	8-8.06 =-0.06	0.003
4.	0	12	12-8.06=3.94	15.52
5.	0	6	6-8.06 =-2.06	4.24
6.	0	5	5-8.06 =-3.06	9.36

Sr. no.	Male(p) Female(q)	Scores of Stress (X)	X- $\bar{x}_{total}$	(X- $\bar{x}_{total}$)^2
7.	0	12	12-8.06=3.94	15.52
8.	0	13	13-8.06 =4.94	24.40
9.	0	9	9-8.06 =0.94	0.88
10.	1	7	7-8.06=-1.06	1.12
11.	1	8	8-8.06 =-0.06	0.003
12.	1	6	6-8.06 =-2.06	4.24
13.	1	6	6-8.06 =-2.06	4.24
14.	1	8	8-8.06 =-0.06	0.003
15.	1	5	5-8.06 =-3.06	9.36
16.	1	8	8-8.06 =-0.06	0.003
17.	1	8	8-8.06 =-0.06	0.003
				Total = 96.895

Np (male) = 9

Nq (female) = 8

M_p = Total scores of males / n = 81/9 = 9.00

M_b = Total scores of females / n = 56 / 8 = 7.00

$M_{total} = \frac{81 + 56}{9+8} = 8.06$

P = Proportion of males= 9 / 17 =.529,

q = Proportion of females = 8 /17 = .471

$\sigma_{total} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{96.90}{16}} = 2.39$

$r_{pbis} = \frac{Mp-Mq}{\sigma} \sqrt{pq}$

$r_{pb} = \frac{9-7}{2.39} X \sqrt{.529\ X.471}$ = 0.84 X .50 =0.42

r_{pb} = 0.42

The computed r value will be either positive or negative sign. The sign of r needs to be interpreted depending upon the coding of the dichotomous group. Suppose, if we had coded male as 1 and female as 0 then rpb would have come with a negative sign. But this negative sign can be ignored as it is arbitrarily decided by us. A positive correlation indicates that the mean of the group coded 1 is more than the mean of the group coded 0.

Significance Testing of Point Biserial Correlation Coefficient

Since r_{pb} is Pearson's correlation, the significance testing is also similar to it.

The *t*- distribution is used for this purpose with $n - 2$ as *df*.

$$t = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1-r_{pb}^2}}$$

Phi coefficient

If both the variables (X and Y) are truly or naturally dichotomous, then the Pearson's Product Moment Correlation calculated is called as Phi coefficient (phi). If the classification of the variables into two categories is truly discrete, then we cannot have more than two categories. For example, items that can be categorized as yes/no, pass/fail, living/dead, etc. and no third option is allowed, it is called truly dichotomous variable.

Phi coefficient is used in item analysis when we want to know the item to item correlation. The phi is a nonparametric statistic and is *also called the mean square contingency coefficient. It makes no assumption about the form of distribution in dichotomous variables.*

The Phi Coefficient can be calculated as:

$$\emptyset = \frac{a.d - b \cdot c}{\sqrt{e \cdot f \cdot g \cdot h}}$$

Similar to a Pearson Correlation Coefficient, a Phi Coefficient takes on values between -1 and +1 where:

- **-1** indicates a perfectly negative relationship between the two variables.
- **0** indicates no association between the two variables.
- **1** indicates a perfectly positive relationship between the two variables.

The further away a Phi Coefficient is from zero, the more

evidence there is for some type of systematic pattern between the two variables.

Computation of Phi Coefficient:

The phi-coefficient related to 2 X 2 table:

Variable 2	**Variable 1 (Gender)**		**Total**
(smoking)	**Male**	**Female**	
yes	a	b	a+b (e)
No	c	d	c+d (f)
Total	a+c(g)	b+d (h)	(a+b+c+d)

Let us take a numerical example –

	Variable 1 (Gender)		**Total**
Variable2 (smoking)	**Male**	**Female**	
yes	50 (a)	40 (b)	90 a+b= (e)
No	40 (c)	60 (d)	100 c+d = (f)
Total	90 (a+c)=(g)	100 (b+d)= (h)	190 (a+b+c+d)

Formula = $\emptyset = \frac{a.d-b\cdot c}{\sqrt{e\cdot f\cdot g\cdot h}}$

Substituting the values in the formula, we get

$\emptyset = \frac{50X60-40X40}{\sqrt{90X100X90X100}} = \frac{3000-1600}{9000} = \frac{1400}{9000} = 0.15$

Ans. $\emptyset$ coefficient = 0.15

The relationship is function of the way we have assigned the number 0 and 1 to each variable. Assigning these numbers helps us to interpret sign of the r. To calculate the proportion of variance shared by the two variables we need to compute

$$r^2 = \phi^2$$

The significance of phi is tested by computing the chi square value from phi for corresponding degree of freedom. The relationship used is as follows:

$$\Phi^2 = \chi^2/N$$

Non-Pearson Correlation Coefficie

Biserial correlations and tetrachoric correlation are two non Pearson's correlation coefficients.

Biserial correlations

The only difference between point biserial and biserial correlation is that we use point biserial when a variable has true dichotomy (e.g. male/female) and we use biserial correlation when a continuous variable is artificially divided into two categories on the basis of criteria determined by the researcher. For example, we have the data of marks (ranging from 0 to 100) that students have scored in psychology test. We decide to divide these into two categories such as pass and fail on the basis of cut off score of 40. So all those who have scored 40 and above are considered as pass and those who have scored less than 40 are considered to have failed in the test. It is known as artificial dichotomy because the cut off point was decided by the researcher and was not naturally occurring dichotomous. The researcher can

shift the cut off point of 40 to 35 or to 50 depending on the data and the objective of his study or purely on the basis of his convenience. The artificial dichotomy assumes that variable underlying this dichotomy is continuous and normally distributed.

In other words, we can say that biserial correlation is a correlational index that estimates the strength of a relationship between an artificially dichotomous variable and a true continuous variable.

Conditions / Assumptions:

1. Both variables must be measured on a continuous scale.

2. One of the variables is artificially made a dichotomous variable

3. There are no outliers for both continuous variables.

4. Both variables are normally distributed.

5. Both have equal variance

6. Sample should be large

All conditions must be satisfied before one decides to use biserial correlation coefficient.

Computation of Biserial Correlation Coefficient

Formula

$$\mathbf{r_{bis}} = \frac{Mp - Mq}{SDt} X \frac{pq}{y}$$

p= Proportion of cases in one of the categories of dichotomous variable

q = Proportion of cases in the other category of dichotomous variable.

Mp = Mean of the values of first category X .

Mq = Mean of the values of second category X .

SDt = standard deviation of the entire group.

y = height of the ordinate of the normal curve separating the portion of p and q

Now let us take a numerical example –

Following are the marks scored by the students on a test of psychology. These marks are divided into two groups, of pass and fail, on the basis of predetermined criteria.

Scores on a test of achieve-ment	Marks Pass (n_1)	Marks Fail (n_2)	Total	Pass group (n_1)		Fail group (n_2)		Entire group		
				X'	fx'	Y'	fy'	Z'	fz'	fz'^{2}
185 - 194	7	0	7	4	28	4	0	4	28	112
175 - 184	16	0	16	3	48	3	0	3	48	144
165 - 174	10	6	16	2	20	2	12	2	32	64
155 - 164	35	15	50	1	35	1	15	1	50	50
145 – 154	24	40	64	0	0	0	0	0	0	0
135 - 144	15	26	41	-1	-15	-1	-26	-1	-41	41
125 - 134	10	13	23	-2	-20	-2	-26	-2	-46	92
115 - 124	3	5	8	-3	-9	-3	-15	-3	-24	72
105 -114	0	5	5	-4	0	-4	-20	-4	-20	80
Total	120	110	230		∑fx'=87		∑fy'=-60		∑fz'=27	∑fz'^{2}=655

(Source: Statistics in Psychology and Education – S.K. Mangal)

$$r_{bis} = \frac{Mp-Mq}{SDt} X \frac{pq}{y}$$

Step 1. p = proportion of cases in the pass group

$$= \frac{n_1}{N} = \frac{Those\ who\ passed}{Total\ No.of\ students} = \frac{120}{120+110} = .52$$

Step 2. q = 1 -p = 1 -.52 = .48

Step 3. *y* = Height of the normal curve separating the portion of p and q

= .3984 **(**critical value as given in the critical value table of biserial correlation, table is not given in this study material**)**

Step 4. To find out Mp = Mean of the values of first category X , Mq = Mean of the values of second category X , SDt = standard deviation of the entire group.

$$Mp = A + \frac{\sum fx'}{N}\ x\ i$$

A = Assumed mean = $\frac{145+154}{2} = 149.5$

(you have already learnt before that to find out the assumed mean, you need to take the class interval that is in the middle, total up the two ends of that class interval and divide by 2).

$$X' = \frac{X-A}{i} = \frac{Mid-value\ of\ X\ scores-assumed\ mean}{class\ interval}$$

N= Total number of students =120

i = size of class interval

Mp = $149.5+\frac{87}{120} x\ 10$ =156.75

$$Mq = A + \frac{\sum fx'}{N}\ x\ i$$

Mq= 149.5 + $\frac{(-60)}{110} x$ 10 = 144.05

Mp -Mq = 156.75 -144.05 = 12.70

$$\sigma = i\sqrt{\frac{\text{fz}'^{2}}{N} - \left(\frac{fz\prime}{N}\right)^{2}}$$

$$\sigma = 10\sqrt{\frac{655}{230} - \left(\frac{27}{230}\right)^{2}} \quad = \quad 16.83$$

$r_{bis} = \frac{Mp-Mq}{SD\text{t}} X \frac{pq}{y} = \frac{12.7\ x\ .52\ x\ .48}{16.83\ x\ .3984} = 0.47$

Ans.: Coefficient of biserial correlation, r_{bias} = o.47

Tetrachoric correlations

When both the variables are artificially dichotomous and both of them cannot be expressed in scores, then we cannot use correlation methods explained above and we need to use tetrachoric correlation method. It estimates what would be the correlation between two binary variables if you could measure variables on a continuous scale. It assumes that if it is possible to get scores for these variables then both variables would be normally distributed and have linear relationship.

For example, suppose a researcher wants to find the relationship between anxiety and social skills. Then, though the scores on anxiety scale are continuous ones, he can artificially make them dichotomous by dividing them into high anxiety and low anxiety based on some cut off score. Similarly, social skills may be artificially categorized into two categories- having social skills and not having social skills.

Though tetrachoric r is less reliable than the Pearson's r, it becomes more reliable if –

- N is large
- The division in the two categories are near the medians.

The value for a tetrachoric correlation can range from -1 to 1 where:

-1 indicates a strong negative correlation between the two variables.

0 indicates no correlation between the two variables.

1 indicates a strong positive correlation between the two variables.

Formula –

$r_t = \cos\left(\frac{180^0 x\sqrt{BC}}{\sqrt{AD}+\sqrt{BC}}\right)$

A,B,C,D are frequencies in 2X2 table.

Formula – When AD > BC,

$r_t = \cos\left(\frac{180^0 x\sqrt{BC}}{\sqrt{AD}+\sqrt{BC}}\right)$

The value of r_t will always be positive.

Formula – when BC > AD,

$r_t = \cos\left(\frac{180^0 x\sqrt{AD}}{\sqrt{AD}+\sqrt{BC}}\right)$

The value of r_t will always be negative

Formula -when BC=AD,

$r_t = \cos\left(\frac{180^0 x\sqrt{AD}}{2\sqrt{AD}}\right) = \cos 90^0 = 0$

The value of r_t will be always 0.

Computation of tetrachoric correlation:

Let us take an example, where we want to see the relationship between personality (categorized as introvert/ extrovert) and success/ failure on the job.

Example:

Relation between PERSONALITY (INTROVERT/ EXTROVERT)

X/ Y	0	1	Total
0	a	b	a+b
1	c	d	C+d
Total	A+c	B+d	n

SUCCESS/ FAILURE

ON THE JOB

PERSONALITY (INTROVERT/ EXTROVERT)

X/ Y	0 (introvert)	1 (Extrovert)	Total
0 (success)	40 a	30 b	70 a+b
1 (failure)	10 c	30 d	40 C+d
Total	50 A+c	60 B+d	110 n

SUCCESS/ FAILURE

ON THE JOB

Step 1 – Compute AD – A X D = 40X30 =1200

Compute BC – B X C= 30X10 =300

Here, AD is bigger than BC so we use the formula

Step 2 - Formula – When AD > BC,

$r_t = \cos\left(\frac{180^0 x\sqrt{BC}}{\sqrt{AD}+\sqrt{BC}}\right) = \cos\left(\frac{180^0 x\sqrt{300}}{\sqrt{1200}+\sqrt{300}}\right) = \cos\frac{180^0\ X\ 17.32}{34.64+\ 17.32} = \frac{3117.6}{51.96} = \cos 60^0$

Table value of r_t takes as the cosine of an angle is

$\cos 60^0 = .500$

$r_t = .500$

5.7 LINEAR REGRESSION

Linear regression analysis is actually a form of correlation analysis. Linear regression analysis builds on Pearson's product-moment correlation coefficient, *r*, going beyond simply assessing relationships to actually predicting the value of the DV from the values of one or more IVs. One of the important goals of psychology is to predict behavior in a given situation, in this section, we will look at how we can make predictions about one variable based on the information about another variable. Before that, let us look at some of the basic differences between correlation and linear regression. In correlation, we have two variables (e. g. neuroticism and happiness), we can any of these two variables as X and Y, but while using linear regression to make prediction, we have to specify which variable is a predictor (X) and which variable is being predicted(Y). Generally predictor variable is denoted by X and predicted or outcome or criterion variable as Y. But regression analysis is not restricted to only two variables. When there is only one independent variable and one dependent variable in regression analysis, we call it simple linear regression and if there are more than one predictor variables, we call it multiple regression analysis. In this section, we will look at simple linear regression analysis only.

The Regression Line:

While the Pearson correlation measures the degree to which a set of data points form a straight-line relationship, regression is a statistical procedure that determines the equation for the straight line that best fits a specific set of data. Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression). OLS is another name for finding the best fit line. Before we fit a line of best fit to the observed data we must ensure that there is linear relationship between the

two variables. This can be done by plotting the data on a scatterplot or by computing the correlation coefficient of the data. Both these techniques have been discussed above. The term 'regression' simply means that the average value of y is the function of X, means it changes with X.

The formula for equation of straight line is

$$Y = a+bX$$

If X is known then one can predict Y and if Y is known then one can predict X. We predict Y variable based on X variable but it does not mean Y is caused by X.

Y – Criterion variable / DV/ Outcome variable/ Response variable

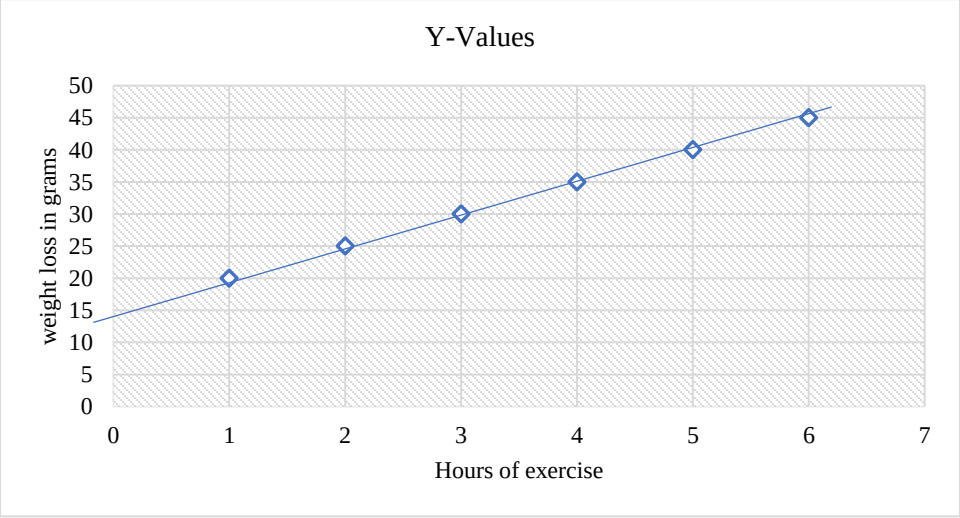
X – Predictor variable /IV / Explanatory variable

b - slope of the line / regression coefficient - The slope is that change in the Y when X changes by one unit. It is the *rate of change*

A - Y intercept/ regression constant - intercept of regression line is that value of Y when the X value is zero. The intercept is the same as the regression constant.

Let us take an example to explain this formula. Suppose we have data of five obese people who have been exercising. We want to find out whether the amount of weight reduction can be predicted on the basis of number of hours spent in exercising.

Obese Persons	Hours spent in exercising (X)	Weight lost in grams (Y)
A	1	20
B	2	25
C	3	30
D	4	35
E	5	40



For this data slope of line can be calculated by using the formula

b (Slope) = $\frac{Y2 - Y1}{X2 - X1}$

Suppose we take Y2 as 25 and Y1 as 20 , similarly we take X2 as 2 and X1 as 1, then the slop will be

Slope = $\frac{25 - 20}{2 - 1} = \frac{5}{1} = 5$

a = The point at which the line passes through Y axis and where X is 0 on X axis is 15 grams. It is also known as Y intercept.

Now if we want to find out how much weight a person will loose if he exercises for 7 hours, then using the equation

$$Y = a + bX$$

We can compute Y = 15 +5(7) = 15+35 = 50

A person will loose 50 grams if he exercises for 7 hours. Using scores outside the range for the predictor variable may give unrealistic (or even impossible) predicted scores for the criterion variable.

However, this equation is applicable only if there is perfect relationship between the two variables, but we know that in real life, the relationships are not so perfect. We need to take into account the error of prediction of the scores on Y variable. It is also called residual. It can be computed as

$$error\ of\ prediction\ = y - \hat{y}$$

Errors are also called deviations. They are vertical distances of all points above or below the line. *The objective is to fit regression line as closely to points as possible. This is done by* making the total of the squares of the deviations as small as possible.

$$Y = a + b\ X + e$$

Y = actual obtained value of Y

X= actual value of the predictor

a= $\bar{y} - b\bar{x}$ [the intercept(the value of $\hat{y}$ when X $= 0$)].

The y-intercept is the height of the line when x is 0. The Y intercept represents the average value of Y when X equals zero. The intercept is the same as the regression constant. The regression constant is a kind of baseline number, the number you start with. Generally, the best baseline number would be the number we predict from a score of 0 on the predictor variable.

b= $\frac{cov_{XY}}{s_X^2}$

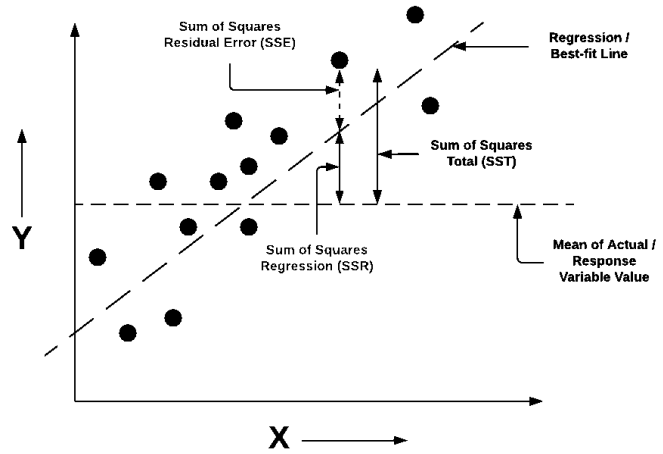
[theslopeoftheregressionline(amountofdifferenceinYwithoneunitofdifferenceinX).]
The slope is called the regression coefficient too.

x intercept - The x intercept is the point where the line crosses the x axis. At this point y = 0.

$\hat{y} = a + bX$

Here $\hat{y}$ is predicted value of Y in sample. It is not an actual value of Y.

Now let us understand the concept of **residual error**.



(Source: Google images)

The data points usually don't fall *exactly* on this regression line; they are scattered around. A residual is the vertical distance between a data point and the regression line. Each data point has one residual. They are:

- Positive if they are above the regression line,
- Negative if they are below the regression line,
- Zero if the regression line actually passes through the point,

The difference between the best fit line and the observed value is called the residual or error. Error just means that there is some unexplained difference.

So we can conclude that there are two important features of regression line –

i) Its distance from base line

ii) Its slope

Ordinary Least Squares (OLS)

It assumes that there are no errors in variable X. Errors are only in prediction of Y. These errors take place because it is not possible to draw a straight line that will pass through all the points. It is possible to draw many lines. So OLS is used to find the best fit. Line of best fit means a line that minimizes the Y error. The $Y - \hat{y}$ is the error in prediction of the Y. It is called obtained residual of the regression. The best line is the one that minimizes this residual. Which means,

$$Y - (a + bX) = e$$

If your line is actually the best fit, then the sum and the mean of residual error will always be zero as they are deviations from regression line. So, we need to square each of these residual errors and find the sum of squares. An attempt to *minimize the sum of the squared errors* is called as least squares. As mentioned above, regression equation means best line of fit. In other words, a line that comes closest to the true scores on Y variable. The difference between predicted scores and actual scores on Y variable is error. To make sure that error component in your prediction is at minimum, you must take smallest sum of squared errors. It is also called least squared criterion.

Now let us see how it is computed

Students	**X**	**Y**	**X = (X-$\bar{x}$)**	**Y = (y- $\bar{y}$)**	x^2	y^2	**xy**
A	2	3	-2	-1	4	1	2
B	3	2	-1	-2	1	4	2
C	2	4	-2	0	4	0	0
D	5	1	1	-3	1	9	-3
E	8	10	4	6	16	36	24
N=5	∑= 20	∑= 20			∑=26	∑=50	∑= 25
	$\bar{x} = 4$	$\bar{y} = 4$					

$cov_{xy} = \frac{\sum(xy)}{n} = \frac{25}{5} = 5$

b $= \frac{cov_{XY}}{S_x^2} = \frac{5}{2.28^2} = 0.963$

$S_x = \sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{26}{5}} = 2.28$

Y= a+ bX+e

a= $\bar{y}$ – b$\bar{x}$ = 4 – (,963) (4) = 0.148

$\hat{y} = a + bX$

= o.148+.963X

Predicted weight loss = 0.148+.963 (hour of exercise)

$\hat{y} = 0.148 + 0.963(2) = 2.074$

When the value of X=2 predicted value of Y will be 2.074.

Assumptions or Conditions to Be Met to Use Regression Analysis:

1. IV/DV – both must be quantitative
2. For each variable, the sample size must be at least 20 cases.

3. identify and remove extreme outliers.

4. Normality – The error terms follow the normal distribution with a mean zero and variance one.

5. The error terms are independent.

6. The population of X and the population of Y follow normal distribution and the population pair of scores of X and Y has a normal bivariate distribution.

7. Linearity – there should be a linear relationship between two variables, use scatter plot to check that.

6. Homoscedasticity – the variance of the residue errors is the same across all values of the predictor (X)

7. Residuals (errors) should be normally distributed and error terms should be independent

8. There should be independence among the pairs of scores

9. The researcher should stay within the range of the data. For example, if the data is from 10 to 60, do not predict a value for 400.

10. The researcher should not make predictions for a population based on another population's regression line.

5.8 SUMMARY

In this unit, we have discussed what is meant by correlation. The word correlation is used to describe the covariance of two variables. Variables that have some order in them. A simple bivariate correlation analyses can measure the strength and direction of the relationship between two variables. Scattergram is presentation of relationship between two variables in pictorial or graphical form. The pairs of scores are plotted, one for variable X and one for variable Y. This relationship between X and Y can be linear or non-linear, positive or negative, strong or weak. The range of correlation coefficient is +1 to -1. Pearson's Product Moment correlation coefficient is biased estimate of population coefficient and to tackle this anomaly, adjusted r is used. Partial correlation is used when we have more than two variables and we want to find out the effect of each variable separately. Special type of correlations are point biserial correlation, biserial correlation, phi coefficient and tetrachoric coefficient.

Point biserial correlation method is used when at least one of the variables are truly dichotomous and biserial correlation is used when continuous variable is made dichotomous artificially. Phi coefficient is used when both the variables are truly dichotomous and tetrachoric correlation is used when both the variables are artificially dichotomous. Biserial and tetrachoric correlations are non-Pearson correlations.

5.9 QUESTIONS

1. What is scatter plot and how it is prepared.
2. Discuss in brief the assumptions underlying the use of Pearson's Product Moment method for computing the correlation coefficient.
3. What is the difference between point biserial correlation and biserial correlation. Discuss the process of their computation with the help of hypothetical data.
4. Write a detailed note on adjusted r and partial correlation.
5. Discuss in detail what is linear regression.
6. What are the conditions to be met to compute linear regression analysis.
7. Explain phi coefficient and tetrachoric correlation.

5.10 REFERENCES

Aron & Aron (2008). *Statistics for Psychology* (5th ed). New Delhi: Pearson

Howell, D. (2009).*Statistical Methods for Psychology* (7th ed.). Wadsworth.

Mangal S.K. "Statistics in Psychology and Education" 2nd ed. PHI Learning Pvt. Ltd., 2010.

❄❄❄❄❄❄❄

6

ASSOCIATION, PREDICTION AND OTHER METHODS – II

Unit Structure:

6.0 OBJECTIVES

After studying this unit, you will understand

a) what is meant by Kendall’s tau

b) What is Spearman ‘s rho

c) what are the measures of nominal data

d) why and how chi square is used

e) What is multiple regression and logistic regression

6.1 INTRODUCTION

In previous unit, you have learnt about very robust statistical techniques, viz., various types of correlations and regression analysis. But there are certain research situations where these techniques cannot be used as they do not satisfy the assumptions of parametric tests. So we will look at non parametric techniques such as Kendall’s tau and Spearman rho. Chi square, binomial test and proportion tests are Similar to each other. While chi square is used to find out the difference between observed frequencies and expected frequencies, binomial test is used to test probabilities of success in repeated trial experiments and proportion test is used to test whether

given proportion of a sample differs from population proportion. There are various measures of nominal data but the most popular one is chi square.

6.2 KENDALL'S TAU

Kendall's tau (**ð)** is one of a number of measures of correlation or association. Kendall's tau is a bivariate measure of correlation/association that is employed with rank-order data. If the data on X or Y or on both the variables are in rank order then Kendall's tau can be used to evaluate the degree of agreement between the rankings of two judges for n subjects/objects. Tau measures the degree of agreement between two sets of ranks with respect to the relative ordering of all possible pairs of subject/objects. We can say that Kendall's tau is a proportion which represents the difference between the proportions of concordant pairs of ranks less the proportion of discordant pairs of ranks.

Just as in Pearson's Product Moment Correlation, for kendall's tau too, the range of possible values is -1.00 to +1.00. The computed value of tau will equal + 1 when there is complete agreement among the rankings (i.e., all of the pairs of ranks are concordant), and will equal -1 when there is complete disagreement among the rankings (i.e., all of the pairs of ranks are discordant).

Concordant pairs – the number of observed ranks below a particular rank which are larger than that particular rank.

Discordant pairs-the number of observed ranks below a particular rank which are smaller in value than that particular rank.

While computing Kendall tau, one set of ranks represents the ranks on the X variable, and the other set represents the ranks on the Y variable. For each subject there will be a pair of ranks (R_x and R_y). Now let us see how it is to be computed.

Computation of Kendall tau (without tied ranks)

Step 1

Below in step 1 we will rank the scores on the X and Y variable starting from rank 1 for the lowest number

Step 2: We will arrange the list of N subjects so that the rank of the subjects on variable X are in their natural order, that is, 1, 2, 3,….N.

Step 3: Next we will place the ranks of Y in the order in which they occur when X ranks are in natural order.

Step 4 : To determine concordant and discordant, we look at only column number 7 . We start with top most row and ask , in this column ' how many numbers are smaller than the number given in that row and how many numbers are bigger than that number in that column ' . The number of bigger numbers are put under column 8 in first row and total number of discordant are put in column number 9. For example, in the data below we

start with number 5 in column number 7 and find that in column number 7, there is no number below 5 that is bigger than 5, so concordant is zero. There are 4 numbers below 5 in that column that are smaller than 5, so we put number 4 in column number 9 in first row. Now we ignore number 5 and again ask the same question for the next number, ' how many numbers are bigger than one and how many numbers are smaller than one in that column. There are 3 numbers bigger than one and no number is smaller than one in that column, so we write 3 in second row of column number 8 and 0 in column number 9. When we come to the last number in column number 7 , i.e., 4, we ask the same question, but there is no number under that, so we write zero in both concordant and discordant columns.

6. Total up the concordant and discordant column and apply the formula

$$\tau au = \frac{concordant - discordant}{concordant + discordant}$$

Column No. 1	2	3	4	5	6	7	8	9
Students	Scores given by Judge X	Rank R_x	Scores given by Judge Y	Rank R_y	Natural rank order of X	Correspond-ing ranks of Y	Concordant	Discordant
A	4	2	3	1	1(E)	5 (E)	0	4
B	7	3	4	2	2 (A)	1 (A)	3	0
C	8	4	7	3	3 (B)	2 (B)	2	0
D	9	5	8	4	4 (C)	3 (C)	1	0
E	3	1	9	5	5 (D)	4 (D)	0	0
							∑=6	∑=4

$$\tau au = \frac{concordant - discordant}{concordant + discordant} = \frac{6-4}{6+10} = .20$$

Kendall Rank-Order Correlation with Ties

If there is tie in the ranks given by either one or both judges, we need to use a different formula. Let us take an example of tied ranks. In this example, scores are already given by the judges and we are going to rearrange those ranks as we did in case of untied ranks example.

As shown in above example, in this example too, we focus on column number 5 and ask the same question ' how many numbers are bigger than 3.5 and how many are smaller than 3.5 in this column'. So there are two numbers that are smaller than 3.5, we put number 2 in discordant (D) and 8 numbers are bigger than 3.5 in that column, so we put number 8 in concordant (C). The number in the second row is same as number in first row in column number 5, so while counting bigger and smaller numbers, we ignore this equal number. After that, we continue with the next number in the column, i.e., the second row and ask the same question. The numbers placed above this number (in row one) are to be ignored.

1	2	3	4	5	6	7
Students	**Judge X Ranks**	**Judge Y Ranks**	**Rearranged ranks of X**	**Rearranged ranks of Y**	**C**	**D**
A	3	1.5	1 (D)	3.5 (D)	8	2
B	4	1.5	2 (C)	3.5(C)	8	2
C	2	3.5	3 (A)	1.5(A)	8	0
D	1	3.5	4 (B)	1.5(B)	8	0
E	8	5	5 (K)	10.5(K)	1	5
F	11	6	6 (H)	8(H)	3	3
G	10	7	7 (I)	9(I)	2	3
H	6	8	8 (E)	5(E)	4	0
I	7	9	9 (L)	12 (L)	0	3
J	12	10.5	10 (G)	7 (G)	1	1
K	5	10.5	11 (F)	6 (F)	1	0
L	9	12	12 (J)	10.5 (J)	0	0
					∑=44	∑=19

Now in next step we need to find the value of Tx and Ty. T means ties. There are no ties among the scores of X ranks, so

Tx = 0.

On Y scores there are three sets of tied ranks. Two students have tied ranks of 1.5, two other students have tied ranks of 3.5 and another two students have tied ranks of 10.5. In each of these cases tied observations are 2.

Ty = t (t-1)

Ty = 2(2-1) + 2(2-1) + 2(2-1) = 6

S= sum total of concordant and discordant = C-D = 44 – 19 = 25.

N = 12

So now we apply the formula for tau of tied ranks

$$\tau au = \frac{2\,S}{\sqrt{[N(N-1)-Tx][N(N-1)-Ty]}}$$

$$= \frac{2*25}{\sqrt{[12(12-1)-0][12(12-1)-6]}} = \frac{50}{\sqrt{(132)(126)}}$$

$$= \frac{50}{128.97} = .39$$

Interpretation: There is a positive but very weak relationship between the two rankings.

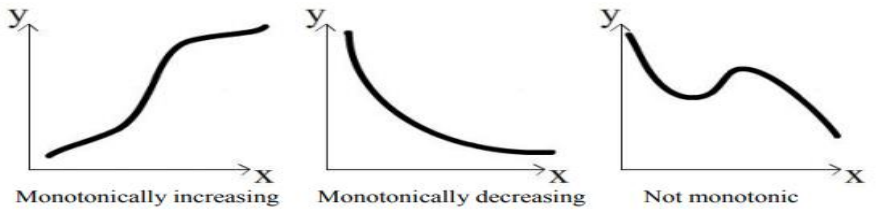
6.3 SPEARMAN'S RHO

Pearson correlation coefficients measure only linear relationships.

The Spearman's rank-order correlation is the nonparametric version of the Pearson's Product Moment Correlation. It measures the strength and direction of association between two ranked variables. If the data on X or Y or on both the variables are in rank order then Spearman's *rho* is applicable. It can also be used with continuous data when the assumptions of Pearson's assumptions are not satisfied. It is used to assess the strength and direction of the monotonic relationship. So, a meaningful relationship can exist even if the correlation coefficients are 0. What is monotonic relationship? See fig 1

Fig.1

In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate.



- Monotonically increasing - as the x variable increases the y variable never decreases;
- Monotonically decreasing - as the x variable increases the y variable never increases;
- Not monotonic - as the x variable increases the y variable sometimes decreases and sometimes increases.

Source: Google images

Formula for Spearman's rank order correlation with untied ranks

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

ρ = Spearman's rank order correlation

D = difference between the pair of ranks of X and Y

n= the number of pairs of ranks

Computation of Spearman's rank order correlation with untied ranks

Step 1 Rank the scores of X variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. In case of our data, the scores are already ranked.

Step 2 Find the difference in the ranks of X and Y and then square the differences and total up the squared differences.

Step 3 apply the formula for spearman's rank order correlation.

X	Y	R_X	R_Y	D= R_X- R_Y	D^2
22	35	2	7	-5	25
23	39	3	8	-5	25
29	22	6	2	4	16
20	40	1	9	-8	64
25	31	4	6	-2	4
27	45	5	10	-5	25
30	30	7	5	2	4
34	28	8	4	4	16
37	25	10	3	7	49
35	20	9	1	8	64
					$\sum$ = 292

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\ X\ 292}{10(100-1)} = 1 - \frac{1752}{990} = -0.77$$

The results show a strong but negative relationship between the two variables.

Spearman's rho with tied ranks

Formula :

$$\rho = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right]}}$$

Let us take an example of tied ranks

Subjects	**X**	**Y**	**R_x**	**R_y**	**$(Rx)^2$**	**$(Ry)^2$**	**$(R_x)(R_y)$**
1	7	8	3.5	2.5	12.25	6.25	8.75
2	11	16	6.5	9.5	42.25	90.25	61.75
3	16	14	9	7	81	49	63
4	9	12	5	5.5	25	30.25	27.5
5	6	8	2	2.5	4	6.25	5.00
6	17	16	10	9.5	100	90.25	95
7	7	9	3.5	4	12.25	16	14
8	11	12	6.5	5.5	42.25	30.25	35.75
9	5	7	1	1	1	1	1
10	14	15	8	8	64	64	64
			$\sum$= 55	$\sum$= 55	$\sum$= 384		Σ=375.75

$$\rho = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right]}}$$

$$\rho = \frac{375.75 - \frac{(55)(55)}{10}}{\sqrt{\left[384 - \frac{(55)^2}{10}\right]\left[383.5 - \frac{(55)^2}{10}\right]}} = 0.902$$

Interpretation - There is a very strong positive relationship between variable X and Y.

6.4 MEASURES FOR NOMINAL DATA

You have already studied that there are four types of data, nominal, ordinal, interval and ratio. Nominal data is the least precise and complex level. The word nominal means "in name," so this kind of data can only be labelled. It does not have a rank order, equal spacing between values, or a true zero value. If the variable is nominal, the mode is the only measure of central tendency to use.

The index of qualitative variation (IQV) is a measure of variability for nominal variables such as race and ethnicity. The index can vary from 0.00 to 1.00. When all the cases in the distribution are in one category, there is no variation (or diversity) and the IQV is 0.00. In contrast, when the cases in the distribution are distributed evenly across the categories, there is maximum variation (or diversity) and the IQV is 1.00.

To calculate the IQV, we use this formula:

$$\text{IQV} = \frac{\boldsymbol{K(100^2 - \Sigma Pct^2}}{\boldsymbol{100^2(K-1)}}$$

where K = the number of categories

∑ Pct2 = the sum of all squared percentages in the distribution

The steps we follow to calculate the IQV:

1. Construct a percentage distribution.
2. Square the percentages for each category.
3. Sum the squared percentages.
4. Calculate the IQV using the formula

Let us take a numerical example

Top Three Racial/Ethnic Groups for Two States by Percentage

Race/Ethnic Group	**Maharashtra**	**Bengal**
Marwari	98.0	35.1
Bengali	1.0	n.d.
Tamilian	1.0	53.3
Maharashtrian	n.d.	11.6
Total	100	100

Squared Percentages for Three Racial/Ethnic Groups for Two States

Race/Ethnic Group	**Maharashtra**		**Bengal**	
	%	**$(\%)^2$**	**%**	**$(\%)^2$**
Marwari	98.0	9604	35.1	1232.01
Bengali	1.0	1	n.d.	n.d.
Tamilian	1.0	1	53.3	2840.89
Maharashtrian	n.d.	n.d.	11.6	134.56
Total	100	9606	100	4207.463

The IQV for Maharashtra is

$$\text{IQV} = \frac{\boldsymbol{K(100^2 - \Sigma Pct^2}}{\boldsymbol{100^2(K-1)}}$$

$$= \frac{\boldsymbol{3(100^2 - 9606}}{\boldsymbol{100^2(3-1)}} = 0.06$$

The IQV for Bengal is

$$\text{IQV} = \frac{\boldsymbol{K(100^2 - \Sigma Pct^2}}{\boldsymbol{100^2(K-1)}}$$

$$= \frac{\boldsymbol{3(100^2 - 4207.463}}{\boldsymbol{100^2(3-1)}} = 0.87$$

In Bengal, where the IQV is 0.87, there is considerably more racial/ethnic variation than in Maharashtra, where the IQV is 0.06.

Crosstabulation (also known as contingency or bivariate tables) is commonly used to examine the relationship between nominal variables **Chi Square tests-of-independence** are widely used to assess relationships between two independent nominal variables.

Apart from chi square, there are many other statistics that can be used to gauge the strength of the association between two nominal variables. They are used as measures of effect size for tests of association for nominal variables.

The statistics *phi* and Cramér's *V* are commonly used. Cramér's *V* varies from 0 to 1, with a 1 indicting a perfect association. *phi* varies from –1 to 1, with –1 and 1 indicating perfect associations. *phi* is available only for 2 x 2 tables.

Cohen's *w* is similar to Cramér's *V* in use, but it's upper value is not limited to 1.

The odds ratio is appropriate for some contingency tables. It is useful when the table describes a bivariate response among groups. For example, disease or no disease among males and females. Or, pass or fail among locations.

Cohen's *h* is used to compare the difference in two proportions. It can be used in 2 x 2 contingency tables in cases where it makes sense to compare the proportions in rows or columns. It can also be used in cases where proportions are known but the actual counts are not. A value of 0 indicates no difference in proportions, and the difference between proportions of 0.00 and 1.00 results in a value of ± *pi* (c. 3.14).

Goodman and Kruskal's *lambda* statistic is also used to gauge the strength of the association between two nominal variables. It is formulated so that one dimension on the table is considered the independent variable, and one is considered the dependent variable, so that the independent variable is used to predict the dependent variable. It varies from 0 to 1.

Another measure of association is Tschuprow's *T*. It is similar to Cramér's *V*, and they are equivalent for square tables (one with an equal number of rows and columns).

Conditions for data

- Two nominal variables with two or more levels each. Usually expressed as a contingency table.
- Experimental units aren't paired.

6.5 CHI SQUARE: CHI-SQUARE TEST FOR A SINGLE SAMPLE

Chi square is a non-parametric test. It is an approximate test of significance for association between two categorical variables when data is in the form of frequency counts and interest focuses on how many subjects fall into different categories. It does not require equality of variances among the study groups or homoscedasticity in the data. It permits evaluation of both dichotomous independent variables, and of multiple group studies.

Distribution of chi square is very different from the normal distribution or t distribution. See fig. 1 and fig 2.

Fig.1

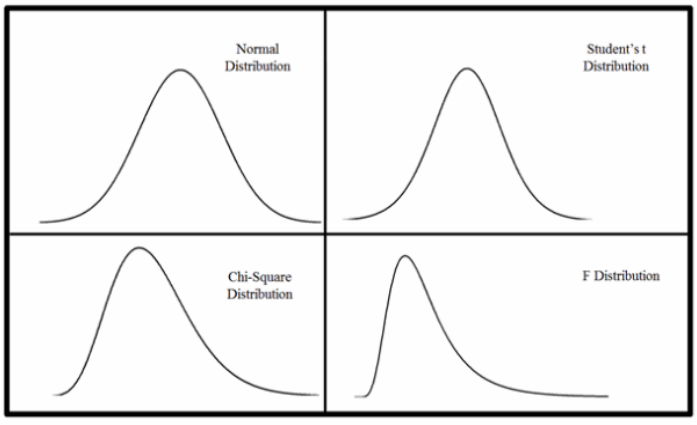
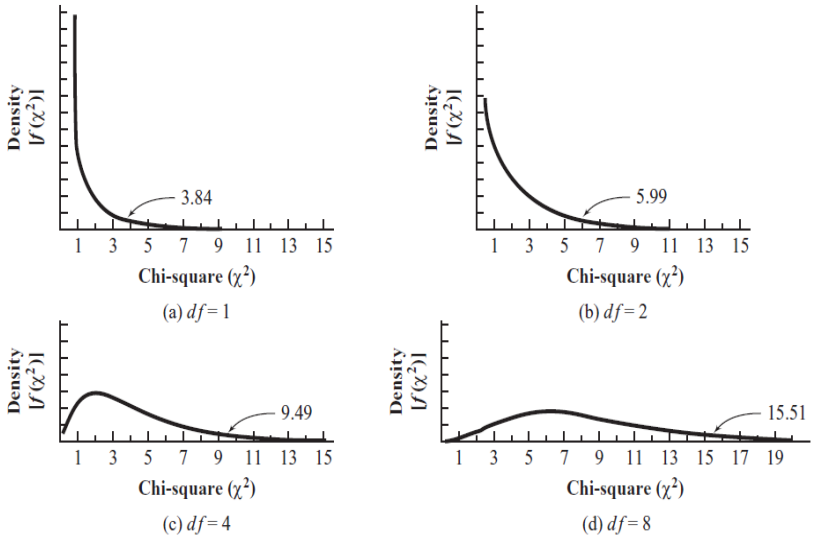


Fig.2



Chi-square distributions for *df* 5 1, 2, 4, and 8.

(Source: Howell, D. (2009)) .*Statistical Methods for Psychology* (7th ed.). Wadsworth)

The chi square distribution is actually not a single curve, instead it is a family of curves, each curve is based on the degrees of freedom. The area under each curve of the chi square distribution equals to one. This type of distribution was first discovered by Helmert in 1875 and later in 1900 Karl Pearson rediscovered it. He used it as a 'goodness of fit' to find out how far the observed frequencies fit the expected theoretical frequencies in some hypothesis.

A one-sample chi-square includes only one dimension, variable, and is usually referred to as a **goodness of fit test.** In other words, it tests just how well do the collected data fits the pattern that was theoretically expected., 50% - 50%?

Let us understand the meaning of observed and expected frequencies through three hypothetical situations.

Hypothesis of chance:

For example, suppose we collect the data from student population on the question, whether cigarette smoking should be banned? The answer can be given either as yes or no. If we have taken a random sample, the expected frequency of yes and no will be 50:50. That is 50% students are likely to say 'yes' and another 50% are likely to say no. Suppose, in actual collected data, we get 60 'yes' and 40 'no' answers, that will be known as observed frequencies.

Hypothesis of Equal Probability:

Suppose there were three options to the question of cigarette smoking instead of two, such as yes, no, not decided. In that case theoretical expectancies will be divided into three parts on the basis of sample size. If

there are 90 students in our sample then theoretically the chances of getting yes, no, or not decided will be equal, i.e., 30 each and that will be the expected frequency of that experiment. So, the expected frequencies are the frequencies you would expect if the null hypothesis were true.

Hypothesis of normal distribution:

Expected frequencies can be determined on the basis of normal distribution of observed frequencies too in the entire population, instead of just the sample size. Suppose we have a sample size of 350 students who had to choose only one colour shirt out of six different colored shirts, and observed frequencies are recorded into five categories as shown in table 1.

Table 1

White	Black	Pink	Green	Purple	Red
40	85	75	65	45	40

If it is assumed that there is equal preference for each colour is normally distributed in large population, then expected frequency in each category will be on the basis of population of normally distributed curve.

Two-sample chi-square includes two dimensions or variables, and is usually referred to as a test of independence. For example, it might be used to test whether preference for certain elective subjects is *independent* of peer pressure and gender.

Participation in one category should not allow participation in another.

The data from all cells of the table should add up to the total count, and no item should be counted twice.

The objective of chi square is to test the statistical significance of the observed (experimental) frequencies relationship with respect to the expected (theoretical) frequencies relationship. The sampling method should be random sampling.

Assumptions for using Chi square goodness of fit test are

1. Sample is randomly selected from a given population.
2. Observations are representative of the populations of interest.
3. Categorical nominal data are used in analysis. Data cannot be in the form of percentages, ratios, codes or anything else.
4. Data should represent actual tallies/counts (and not on percentages, proportions, means, etc. for k mutually exclusive categories,
5. The expected frequency of each cell is at least 5 or greater in at least 80% of the cells and no cell should have an expected frequency of less than one. If the total sample size is ≥ 20, then expected frequencies in one or two cells can be as low as 1 or 2.

6. The categories are mutually exclusive and each subject contributes data to one and only one cell in the chi square. (if the same subject is tested over time and the comparisons are made of time1, time 2, time3, etc., then chi square cannot be used.

7. There are 2 variables and both are measured as categories. But data can be ordinal, interval or ratio that has been collapsed into ordinal categories can be used.

Null Hypothesis for computing chi square will be that there is no actual difference between the observed frequencies (derived from data) and expected frequencies (derived from either chance factor or from some theoretical basis such as hypothesis of normal distribution).

Computation of Chi Square:

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

f_o = observed frequencies in each category

f_e = expected frequencies in the corresponding category

Step 1. Make a contingency table of observed frequencies

Step 2. Compute expected frequencies

Step 3. Find the difference between observed and expected frequencies and square them

Step 4. Find the degrees of freedom by using the formula –
df = (Number of Rows – 1) x (Number of Columns – 1)

Step 5. Compare computed chi square value with table value

Step 6. If the computed chi square value is equal or higher than the table value, reject the null hypothesis.

Let us take a numerical example

H_o = There is no association between the gender and frequency for the replies for question "How often do you exercise?"

Gender	Frequently	Occasionally	Rarely	Never	Total
Male	10	5	4	6	25
Female	20	10	3	2	35
Total	30	15	7	8	60

		Male	**Female**	**Total**	χ^2
Frequently	Observed O	10	20	30	0.50+0.36 = 0.86
	Expected E	30*25/60 = 12.50	30*35/60 = 17.50		
	O-E	10-12.50 = - 2.50	20-17.50 = 2.50		
	$(O-E)^2$	$(-2.50)^2$ = 6.25	$(2.50)^2$ = 6.25		
	$(O-E)^2/E$	6.25/12.50 = 0.50	6.25/17.50 = 0.36		

		Male	**Female**	**Total**	χ^2
Occasionally	Observed O	5	10	**15**	0.25 + 0.18 = 0.43
	Expected E	115*25/60 = 6.25	35*15/60 = 8.75		
	O-E	5 – 6.25 = - 1.25	10-8.75 =1.25		
	$(O-E)^2$	$(-1.25)^2$ = 1.56	$(1.25)^2$ =1.56		
	$(O-E)^2/E$	0.25	0.18		

		Male	**Female**	**Total**	χ^2
Rarely	Observed O	4	3	7	0.40+0.29 =0.69
	Expected E	7*25/60 = 2.92	7*35/60 = 4.08		
	O-E	4 -2.92 = 1.08	3- 4.08 = - 1.08		
	$(O-E)^2$	$(1.08)^2$ = 1.17	$(- 1.08)^2$ = 1.17		
	$(O-E)^2/E$	1.17/2.92 = 0.40	1.17/4.08 =0.29		

		Male	Female	Total	χ^2
Never	Observed O	6	2	8	3.66
	Expected E	3.33	4.67		
	O-E	2.67	-2.67		
	$(O-E)^2$	7.11	7.11		
	$(O-E)^2/E$	2.13	1.52		

$\sum\chi^2$ = 0.86+ 0.43 + =0.69 + 3.66 = 5.64

df= (Number of Rows – 1) x (Number of Columns – 1)

(2-1) X (4-1) = 1 X 3= 3

df= 3

Interpretation: At 0.01 level of significance, calculated chi square value (5.64) is less than critical chi square table value (11.345), so null hypothesis is accepted. Evan at 0.05 level of significance, calculated chi square value of 5.64 is less than critical chi square table value of 7.815, so null hypothesis is accepted. So, we can say there is no association between the gender and frequency for the replies for question "How often do you exercise?"

6.6 BINOMIAL TEST

Binomial distribution was first discovered by James Bernoulli in 1700. It deals with situations in which each of a number of independent trials results in one of two mutually exclusive outcomes. Such a trial is called a Bernoulli trial. The binomial distribution is an example of discrete rather than continuous distribution. For example, out of 100 students, we can have 70 passed and 30 failed students but not 68.97 passed.

The binominal test is used when there can be exactly two mutual exclusive outcomes of a trial, e.g. a child is either male or female, a person is either alive or dead, a coin has either head or tail face, etc. All such dichotomous possibilities are labelled as 'success' or 'failure'. These are arbitrary labels for the two alternative outcomes.

To understand it clearly, suppose a random experiment is performed repeatedly and each repetition is called a trial. The occurrences of an event in each trial are called a 'success' and non-occurrence is called a 'failure. There are finite number of independent Bernoulli trials in which the probability of success is constant for each trial, then q = 1-p will be the probability of failure in any trial. The two independent constant n and p in the distribution are known as the parameter of the distribution. 'n' is also

known as the degree of binomial distribution. Any random variable that follows binomial distribution is known as binomial variate.

The Properties of Binomial distribution are

1. Fixed number of trials , n, means that the experiment is repeated a specific number of times.

2. The n trials are independent, i.e., what happens in one trial does not influence the outcome of other trials.

3. There are only two outcomes, called 'success' and 'failure'.

4. The probability of a success does not change from trial to trial where

   p = probability of success

   q = probability of failure = 1-p.

   p + q =1

If the researcher has an idea about what the probability of success can be and if the null hypothesis states that there are equal chance of both possibilities occurring, we use binomial test to see how much observed results differ from expected results. If there are more than two categories of possible outcomes, we use multinomial test. This test is similar to chi square with a little difference. If the sample size is large, we use chi square test, but if the sample size is small, we use binomial test.

Assumptions of binomial test

1 Observations are sampled at random from a binary population.

2 Each observation is independent (does not affect the value of any other observations sampled)

3 The probability of any sample observation being classified into one of the two categories is fixed for the population.

4 With small sample sizes such as $n \leq 25$, the exact binomial probability can be evaluated.

With larger sample sizes, especially when P is close to 0.5, the binomial approximation to the normal distribution with a continuity correction (because the normal distribution is continuous but the binomial distribution is discrete) can be used. In this case the normal variate *Z* is used to evaluate the probability of the observed outcome.

The shape of Binomial Distribution depends on the value of p and n. If p=q=0.5, the distribution will be symmetrical, no matter what the value of n is. If the p is not equal to q, the distribution will be asymmetrical. For a given particular n, the more the difference between p and q, the greater the skewness of the distribution will be.

Computation: The problem taken is that a six sided dice is rolled 12 times. Out of these 12 trials, what is the probability of getting the number 4 five times.

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x \, q^{n-x}$$

Where

n= the number of trials or the numbers being sampled

x= the number of successes desired

p= probability of getting a success in one trial

q= 1-p= the probability of getting a failure in one trial

$$\mathbf{P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x \, q^{n-x}}$$

n= 12 , x =5 , p = $^1/_6$, q= $^5/_6$ $\binom{n}{x} = nCr = \frac{n!}{(n-r)!r!}$

$= \binom{12}{5} = 12 \text{ C } 5 = \frac{12!}{(12-5)!5!} = \frac{12*11*10*9*8*7!}{7!\,5*4*3*2*1}$

(7! In numerator and denominator cut each other, 12 in numerator cuts 4*3 in denominator, similarly, 5*2 in denominator cuts 10 in numerator. So, we are left with 11*9*8 in numerator and only one in denominator.)

$\binom{n}{x} = 792$

$P(5) = 792 \, (^1/_6)^5 \, (^5/_6)^{12-5} = 792 \, (^1/_6)^5 \, (^5/_6)^7$

$= 0.028425$

If we multiply 0.028425 with 100% we get

$= 2.84\%$

So, the probability of getting 4 on dice face five times if we roll six sided dice 12 times is 2.84%

6.7 PROPORTIONS TEST

A test of proportion will assess whether or not a sample from a population represents the true proportion from the entire population. For example, suppose we want to find out the proportion of males within a given total population of adults. A test of proportion will assess whether or not a sample from a population represents the true proportion from the entire population.

Computation of proportion test using the critical value approval

Suppose a book publisher believes that there are 70% students in a particular university that have opted for psychology as their major subject. The

salesman does not agree with it and believes this percentage to be different. He conducts a survey of 200 students in that university and finds that 130 students have opted for psychology as their major subject. Now let us form a null and alternate hypothesis. At a 95% confidence level, let us see if there is enough evidence to reject the null hypotheses.

Ho: p = o.70 , Qo = 0.30

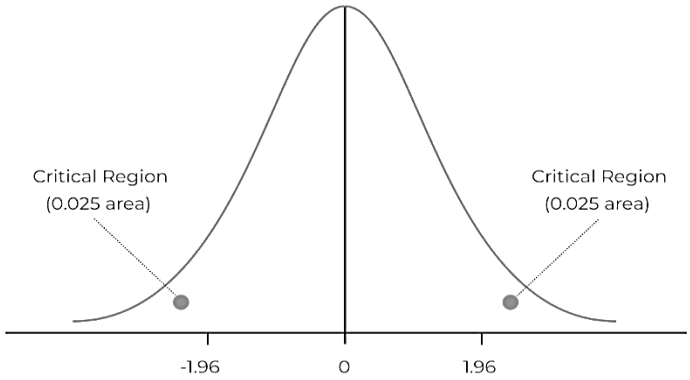
Ha: p ≠ 0.70

N = 200

X = 130

$$\hat{p} = \frac{x}{n} = \frac{130}{200} = \frac{13}{20} = .65$$

Since alternative hypothesis says that p is not equal to .70,the proportion can be more than 70 or less than 70 so we need to use two tailed test.



With 95% confidence level, the critical z value is -1.96 on the left hand side of the mean and + 1.96 on the right hand side of the mean.

Formula for proportion test

$$z = \frac{\hat{p} - \rho_0}{\sqrt{\frac{\rho_0 q_0}{n}}}$$

$$= \frac{.65 - .70}{\sqrt{\frac{.70 * .30}{200}}} = -1.54$$

Since computed z value of -1.54 is less than table critical z value of -1.96, we cannot reject the null hypothesis that out of 200 students 70% have taken psychology as their major subject.

- $\hat{p}$ is the sample proportion
- p_0 is the hypothesized probability
- q_0 is 1- p_o
- n is the sample size.

When |Zcalc| (absolute value of the calculated test statistic) is smaller than Zcrit (critical value), we fail to reject the null hypothesis and claim that there is no statistically significant difference between the population proportion and the hypothesized proportion.

The above example is of one proportion test. There can be a two-proportion test. A 2-proportion test helps you determine whether two population proportions are significantly different. The null hypothesis for two proportion test will be

H_o: p_1-p_2 = 0

or we can say H_o =p_1=p_2

If we assume p1=p2, then sampling distribution of $\hat{p}_1$ - $\hat{p}_2$ will be approximately normal and will have zero mean. The formula to compute 2 proportion test is

$$z^* = \frac{(\hat{p}1 - \hat{p}2) - 0}{\sqrt{\sqrt{\hat{p}^*}(1 - \hat{p}^*)\,(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Conditions to be met for calculating a confidence interval for a proportion are

1. The sample should be randomly selected

2. There should be only two options or categories

3. The sample size should be at least 10, five members in each category.

6.8 MULTIPLE REGRESSION

In previous unit we discussed linear regression where we had one predictor(X) and one dependent variable (Y). However, in real life, there might be many factors that may jointly influence the dependent variable. For example, job satisfaction of an employee may be influenced by salary, relationship with boss, work itself, work environment and company policies. So, when there are more than one independent variable that can predict change in one dependent variable, we need to use multiple regression. In multiple linear regression analysis, each IV provides its own contribution in predicting the DV. We can say that the multiple linear regression (R^2) explains the relationship between **one continuous dependent variable** (*y*) and **two or more independent variables** (*x*1, *x*2, *x*3… etc)**.**

In multiple regression, while finding out the correlation between a set of IVs and a DV, it is necessary to account for the potential correlation between the IVs themselves. Otherwise, the total correlation will be greater than 100%. So, we need to compute partial correlations between *one* IV and the DV that is not shared with any of the other IVs. It represents the extent to which an IV and the DV are associated with one another, while controlling for the effects of *all* other IVs on both the DV and the IV.

Now let us see how the multiple regression should be computed.

In liner regression the equation used was $\hat{y} = a + bX$, for multiple regression, as the number of IVs are more, the equation is changed to

$\hat{y} = b_o + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

where X_1 and X_2 are the predictors and b_0 is the intercept. b_1 is the value of the estimated coefficient for the first IV (X_1), b_2 is the value of the estimated coefficient for the second IV (X_2), through *K* IVs.

The assumptions that were mentioned in previous unit about linear regression apply to multivariate regression too. Apart from that, for multiple regression, other conditions are that

a) the independent variables in the model should not be strongly correlated with one another. IVs that are strongly correlated with one another are basically measuring the same thing. Thus, the values of the estimated coefficients for strongly correlated IVs would be interchangeable, making it nearly impossible to determine their unique contribution to explaining the variance in the DV.

b) The sample size should be large enough to detect results at the desired effect size.

c) There should be no outliers.

6.9 LOGISTIC REGRESSION

Logistic regression is another common regression method. There are two types of logistic regression – binomial logistic regression and multinomial logistic regression. Binomial logistic regression is a variant of linear multiple regression. But it uses DV as a dichotomous variable, with only two values. For example, job satisfaction of employees in relation to their gender and permanency status of the job. In such a case gender and permanency status are dichotomous variables. The predictor variables may be any type of variable including scores. Like other forms of regression, logistic regression generates *B*-weights (or slope) and a constant. However, these are used to calculate something known as the logit rather than scores. The logit is the natural logarithm of odds for the category. The percentage predicted in each category of the dependent variable can be calculated from this and compared with the actual percentage. The final regression calculation provides information about the significant predictors among those being employed.

Though we can analyze dichotomous dependent variable through discriminant analysis, but that has certain limitations. For example, discriminant analysis can produce a probability value that may lie beyond the range of 0 to 1, which is an impossible probability and secondly it depends upon certain restrictive normality assumptions of independent variables, that are many times not realistic. Logistic regression, on the other hand, does not produce probabilities beyond 0 and 1, and requires no such

restrictive assumptions on the independent variables, which can be categorical or continuous.

Another advantage of logistic regression is that it can be applied in situations where there are three or more levels of the dependent variable. In simple linear regression, we use standard F and t statistics to find our the significance of the relationship and the contribution of each independent variable. But in logistic regression, we use many chi square tests instead of F and t test.

Uses of Binomial logistic regression:

Binomial logistic regression is used to -

1. Determine a small group of variables which characterise the two different groups or categories of cases.

2. Identify which other variables are ineffective in differentiating these two groups or categories of cases.

3. Make actual predictions about which of the two groups a particular individual is likely to be a member given that individual's pattern on the other variables.

If the dependent variable has three or more nominal categories, then multinomial logistic regression should be used. Multinomial logistic regression is a form of binomial logistic regression or multiple regression in which a number of predictors are used to predict values of a single nominal dependent or criterion variable. In other words, Multinomial logistic regression uses nominal or category variables as the criterion or dependent variable. The independent or predictor variables may be score variables or nominal (dichotomised) variables. The concept of dummy variable is crucial in multinomial logistic regression. A dummy variable is a way of dichotomising a nominal category variable with three or more different values. A new variable is computed for each category (just one!) and participants coded as having that characteristic or not. The code for belonging to the category is normally 1 and the code for belonging to any of the other categories is normally 0.

Uses of multinomial logistic regression:

It can help to

1. Identify a small number of variables which effectively distinguish between groups or categories of the dependent variable.

2. Identify the other variables which are ineffective in terms of distinguishing between groups or categories of the dependent variable.

3. Make actual predictions of which group an individual will be a member (i.e. what category of the dependent variable) on the basis of their known values on the predictor variables.

Multinomial logistic regression is preferred over discriminant analysis because it has less unattainable assumptions about the characteristics of data and yet there findings are more or less same.

6.10 SUMMARY

Kendal tau is a non parametric measure of correlation that uses rank ordered data. It finds out the degree of agreement or disagreement between two judges or parties. The range of kendall's tau is same as Pearson's Product Moment Correlation, -1 to +1 . Another non parametric correlation test is Spearman's rho that is to be used when the conditions of parametric tests are not satisfied. Spearman's rho can test the significance of monotonic relationship, i.e., non linear relationships. Nominal data is basically categorical data. The central tendency of nominal data can be found out through mode and variability with the formula of index of qualitative variations. The association between various categories can be measured through chi square, Crame'r's V test and Cohen's w test. Chi square measures the significance of association between two categorical variables. Binomial test is used for categorical data in Binomial distribution. It is used when there are repeated measures overs various trials. Proportion test looks at whether a sample from population represents the true proportion from the entire population. Multiple regression is used when there are more than one independent variables and one dependent variable. Logistic regression is another regression method. There are two types of logistic regression methods- binomial logistic regression and multinomial logistic regression.

6.11QUESTIONS

1. Discuss in detail kendall's tau and Spearman's rho
2. What is meant by binomial test? What are its uses?
3. What is chi square test? What are its assumptions and applications?
4. What is a multiple regression equation?
5. What is the difference between multiple regression and logistic regression? Discuss in detail logistic regression.

6.12 REFERENCES

1. Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.
2. Mangal S.K, (2010) Statistics in Psychology and Education. PHI Learning Private Ltd. New Delhi.

❄❄❄❄❄❄❄

7

FACTOR ANALYSIS AND SOFTWARE PACKAGES – I

Unit Structure:

7.0 OBJECTIVES

After studying this unit, you should be able to:

- Understand factor analysis and various concepts associated with it.
- Explain the assumptions, advantages, disadvantage and application of factor analysis.

- Discuss the different types of methods used to extract the factor from the data set
- Explain rotation and the various rotation methods available
- Classify factor analysis into two types: exploratory and confirmatory and discuss it in detail.

7.1 INTRODUCTION

Large amounts of data are frequently collected by researchers. Occasionally, they speculatively add extra questions to a survey for no apparent reason. With so many variables, it becomes difficult to make sense of the data's complexity. When using questionnaires, it is natural to look for patterns in the correlations between questions. However, due to the sheer number of interrelationships, this is difficult. Take a look at the following brief questionnaire (Table 1):

Table 1:- Sample Questionnaire

Item 1: It is possible to bend spoons by rubbing them.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 2: I have had 'out of body' experiences.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 3: Satanism is a true religion.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 4: Tarot cards reveal coming events.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 5: Speaking in tongues is a peak religious experience.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 6: The world was saved by visiting space beings.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 7: Most people are reincarnated.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 8: Astrology is a science, not an art.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 9: Animals have souls.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly
Item 10: Talking to plants helps them to grow.				
Agree strongly	Agree	Neither	Disagree	Disagree strongly

Source: - Howitt, D. & Cramer, D. (2011). Factor analysis: Simplifying complex data. Introduction to Statistics in Psychology (5th Ed.). Pearson.

Agree strongly could be scored as 1, agree scored as 2, neither as 3, disagree as 4 and disagree strongly as 5. This converts the words into numerical values. Correlating the answers to each of these ten questions with each of the others for 300 respondents yields a large correlation matrix (a table of all possible correlations between all of the possible pairs of questions). Ten questions result in 102 or 100 correlations. Despite the fact that the correlation matrix is symmetrical along the diagonal from top left to bottom right, there are 45 different correlations to investigate. A matrix of this type could resemble the one in Table 2:

Table 2: Correlation matrix of 10 items

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Item 1	1.00	0.50	0.72	0.30	0.32	0.20	0.70	0.30	0.30	0.10
Item 2	0.50	1.00	0.40	0.51	0.60	0.14	0.17	0.55	0.23	0.55
Item 3	0.72	0.40	1.00	0.55	0.64	0.23	0.12	0.17	0.22	0.67
Item 4	0.30	0.51	0.55	1.00	0.84	0.69	0.47	0.44	0.56	0.35
Item 5	0.32	0.60	0.64	0.84	1.00	0.14	0.77	0.65	0.48	0.34
Item 6	0.20	0.14	0.23	0.69	0.14	1.00	0.58	0.72	0.33	0.17
Item 7	0.70	0.17	0.12	0.47	0.77	0.58	1.00	0.64	0.43	0.76
Item 8	0.30	0.55	0.17	0.44	0.65	0.72	0.64	1.00	0.27	0.43
Item 9	0.30	0.23	0.22	0.56	0.48	0.33	0.43	0.27	1.00	0.12
Item 10	0.10	0.55	0.67	0.35	0.34	0.17	0.76	0.43	0.12	1.00

Source: - Howitt, D. & Cramer, D. (2011). Factor analysis: Simplifying complex data. Introduction to Statistics in Psychology (5th Ed.). Pearson.

Overall interpretation is challenging due to the volume of information, which makes it difficult to grasp completely. Large matrices are simply too complex for our brains to process. This is where factor analysis can be useful. It is a method that aids in getting around the difficulty of correlation matrices. In essence, it takes a correlation matrix and produces a much smaller set of "supervariables" that define the correlation matrix's major trends. In comparison to the original matrix, these supervariables or factors are typically much simpler to comprehend.

Let us begin with a brief overview of the history of factor analysis-

7.2 BRIEF HISTORY OF FACTOR ANALYSIS

Factor analysis is not a recent technique – it dates back to shortly after the First World War. Factor analysis is not a recent technique; it first appeared soon after World War One. It was initially developed by psychologists for a very specific use in the field of mental testing. There are many psychological assessments available for various types of intellectual ability. The first psychologist to discuss common factor analysis was Charles Spearman, who did so in his 1904 paper. It was focused on single-factor models and provided few specifics regarding his methodologies. He discovered that the scores of schoolchildren on a wide range of seemingly unrelated subjects were positively correlated, leading him to propose that a single general mental ability, or g, underpins and shapes human cognitive performance. Louis Thurstone presented the first development of common factor analysis with multiple factors in two papers in the early 1930s, which were summarised in his 1935 book, The Vector of Mind. Thurstone introduced several crucial concepts for factor analysis, such as communality, uniqueness, and rotation. He advocated for "simple structure" and developed rotational methods that could be used to achieve it. Stephenson, a Spearman student, distinguishes between R factor analysis—which is focused on the study of inter-individual differences—and Q factor analysis—which is focused on subjective intra-individual differences—in Q methodology. Raymond Cattell was a strong supporter of factor analysis and psychometrics, and he explained intelligence using Thurstone's multi-

factor theory. In addition, Cattell created the "scree" test and similarity coefficients.

As we can see, the goal of factor analysis was to identify which types of mental abilities are distinctive and which tend to go together. It is employed in the creation of psychological tests and questionnaires due to its more widespread usefulness. It is frequently used in personality, attitude, intelligence, and aptitude tests because it aids in determining which items from the tests and measures to retain. It is possible to get "purer" measurements of psychological variables by using factors. The fact that some theorists have used it extensively is not surprising. Raymond Cattell and Hans Eysenck's (Cramer, 1992) personality theories rely heavily on factor analysis. The development of high-speed electronic computers has made the technique relatively common, as it no longer requires months of hand calculations.

Furthermore, in this topic, we will discuss Factor analysis and its various concepts, application in the field of psychology, advantages and disadvantages, methods of extraction, and methods of rotation. Let's start by discussing what factor analysis is.

7.3 WHAT IS FACTOR ANALYSIS?

Factor analysis is a technique for reducing a large number of variables to a smaller number of factors. This method extracts the maximum common variance from all variables and converts it into a single score. This score can be used for further analysis as an index of all variables. Factor analysis is a component of the general linear model (GLM). There are several methods available, but principal component analysis is the most commonly used. When examining variable relationships for complex concepts like socioeconomic status, dietary habits, or psychological scales, factor analysis is a useful tool. It enables researchers to look into ideas they are unable to directly measure. It achieves this by estimating a small number of interpretable underlying factors using a large number of variables.

Basic Factor Model **-** The basic factor analysis problem uses several observable variables to explain how they relate to one another in a way similar to a regression equation. In the common factor model, the common factors serve as predictors of the observed X variables in a regression equation. Equation 1 shows the fundamental factor model.

Equation 1: - Fundamental factor model.

$$\frac{P\text{'s outcomes}}{P\text{'s inputs}} = \frac{O\text{'s outcomes}}{O\text{'s inputs}}$$

In this equation, X represents the observed variables, L represents the factor loadings or regression weights, f represents the common factors, and u represents the residuals. The goal is to explain the interrelationships between the X variables using the common factors, f, and residual error

terms, which is referred to as uniqueness. X's variation is divided into common and specific components. However, unlike regression, the predictors, f, are unknown.

For the sake of illustration, let's assume that several managers are asked to rate the importance of six personality traits for effective employee performance. The characteristics evaluated are organised, systematic, careless, creative, intellectual, and imaginative. The hypothetical correlation matrix for these variables is shown in Table 3. The goal of factor analysis is to find fewer than six underlying latent factors that can adequately explain the relationships between these variables. Organized, systematic, and careless are all associated with one another, but they are not associated with creative, intellectual, and imaginative. Similarly, while creative, intellectual, and imaginative are all related, they are not related to organised, systematic, or careless. There are two sets of correlations, each one indicating a different underlying factor.

Table 3: Hypothetical Correlation Matrix of Observed Variables

Variable	*1*	*2*	*3*	*4*	*5*	*6*
1. Organized	1.00					
2. Systematic	.72	1.00				
3. Careless	–.63	–.55	1.00			
4. Creative	.00	.00	.00	1.00		
5. Intellectual	.00	.00	.00	.56	1.00	
6. Imaginative	.00	.00	.00	.48	.42	1.00

NOTE: These are hypothetical correlations based on imaginary supervisory ratings of "how important are these characteristics in an employee."

The factor pattern matrix for these variables and their corresponding latent factors are shown in Table 4. The correlations between the observed variables and the factors are represented by the pattern coefficients in Columns 2 and 3 of Table 4. Different pattern or loading matrices will display various correlation types (e.g., Pearson correlations, partial correlations).

Table 4: Factor Pattern Matrix with Communalities and Uniqueness for Hypothetical Data

Observed Variable	*Factor 1 Conscientiousness*	*Factor 2 Intellect*	*Communality* (h^2)	*Uniqueness* ($1 - h^2$)
1. Organized	.90	.00	.81	.19
2. Systematic	.80	.00	.64	.36
3. Careless	–.70	.00	.49	.51
4. Creative	.00	.80	.64	.36
5. Intellectual	.00	.70	.49	.51
6. Imaginative	.00	.60	.36	.64
Sum of squared loadings	1.94	1.49		
Proportion of variance	1.94/6 = .32	1.49/6 = .25	.32 + .25 = .57	1 – .57 = .43

NOTE: Variance in each observed variable is 1; therefore, coefficients can be interpreted as standardized (i.e., in *z*-score form). Data are fictional for illustrative purposes only.

The variance that the variable shares with the factors it represents is indicated by the term "communality" (h2) in column 4 of Table 4. In regression, communality is analogous to a squared multiple correlation. The variance that is specific to a variable and is not considered by the factor is

represented by the uniqueness in column 5. The residual variance in the observed variable after the factors have been considered is analogous to uniqueness. The two unobservable factors represented by the six fictitious variables account for 36% to 81% of the variance in the observed variables.

In addition to describing the variance linked to the observed variables, the factor analysis solution also describes the variance in the factors themselves. The variance of each factor is expressed as the sum of the squared loadings (SSL) for a set of variables. These SSLs are known as eigen values before the factor loadings have undergone any transformation. The variance in the observed variables should ideally be accounted for by a small number of factors. Table 4 shows that the factors account for 32% and 25% of the variance in the observed variables, for a total of 57% of the variance. The factors do not account for the remaining 43% of the variance in the observed variables.

Next, we move on to the assumptions of Factor analysis-

7.4 ASSUMPTIONS OF FACTOR ANALYSIS

Factor analysis has several assumptions. These include:

1. **No outlier:** Assume that there are no outliers in data.
2. **Adequate sample size:** The case must be greater than the factor.
3. **No perfect multicollinearity:** Factor analysis is an interdependency technique. There should not be perfect multicollinearity between the variables.
4. **Homoscedasticity**: Since factor analysis is a linear function of measured variables, it does not require homoscedasticity between the variables.
5. **Linearity**: Factor analysis is also based on linearity assumption. Non-linear variables can also be used. After transfer, however, it changes into linear variable.
6. **Interval Data**: Interval data are assumed.

7.5 PURPOSE OF FACTOR ANALYSIS

Factor analysis serves three primary purposes -

- First, it is helpful for measuring constructs that are difficult to see in the natural world. For instance, although intelligence cannot be heard, seen, smelled, tasted, or touched, it can be inferred from the evaluation of observable variables, such as performance on specific ability tests. By analysing the responses to particular questions, factor analysis is also useful in the creation of scales to measure attitudes or other similar latent constructs.

- Second, factor analysis can be used to condense a large number of observations into a smaller number of factors. In the English language, for example, there are thousands of personality descriptors. Researchers have been able to reduce the number of distinct factors required to describe the structure of personality using factor analysis.

- Third, factor analysis can be used to provide evidence of construct validity (e.g., factorial, convergent, and discriminant validity). For instance, if a set of observable variables are conceptually related to one another, factor analysis should show these conceptual connections while also showing that the same variables are largely unrelated to variables from other latent factors.

All three of these applications of factor analysis can be used in the development and testing of psychological theories. Now that we have an idea of what are the assumptions and purpose of factor analysis, let us get in more detail to some key topics in factor analysis.

7.6 ADVANTAGES AND DISADVANTAGES OF FACTOR ANALYSIS

Advantages of Factor Analysis:

- Attributes can be both objective and subjective.
- It can be used to find hidden constraints or dimensions that may or may not be visible through direct analysis.
- It doesn't require a lot of skill, is reasonably priced, and produces precise results.
- The naming and employing of dimensions are flexible.

Disadvantages of Factor Analysis:

- The usefulness is reliant on the researcher's capacity to create a comprehensive and accurate set of product attributes. The value of a procedure is diminished if crucial attributes are overlooked.
- Naming the factors can be challenging. Multiple characteristics may be highly correlated without any obvious causes.
- A meaningful pattern cannot be produced by factor analysis if the observed variables are wholly unrelated.
- Only theory can provide the researcher with information about what the various factors actually mean.

7.7 KEY CONCEPTS AND TERMS IN FACTOR ANALYSIS

7.7.1. What is a factor?

A "factor" is a collection of observed variables with similar response patterns; they are linked to a hidden variable (referred to as a confounding

variable) that is not directly measured. The factor loadings, or the amount of variation in the data that each factor can explain, are used to order the factors. The fundamental idea behind factor analysis is that because several observed variables are connected to a single latent variable, they all exhibit similar patterns of behaviour (i.e. not directly measured). People may, for example, respond similarly to questions about income, education, and occupation, all of which are related to the latent variable socioeconomic status.

There is always one less factor than there are variables in a factor analysis. The factors are always listed in order of how much variation they explain, with each factor accounting for a specific portion of the overall variance in the observed variables. An indicator of how much of the common variance of the observed variables a factor explains is its eigenvalue. Any factor with an eigenvalue ≥1 explains more variance than a single observed variable. Therefore, if the socioeconomic status factor had an eigenvalue of 2.3, it would account for 2.3 of the variance in the three variables. The factor could then be applied to other analyses because it captures the majority of the variance in those three variables. Generally, the factors that explain the least amount of variation are eliminated.

7.7.2. What are Factor loadings?

Factor loading is essentially the correlation between the factor and the variable. Factor loading displays the variance on that specific factor that is explained by the variable. According to the SEM approach, a factor loading of 0.7 or higher indicates that the factor extracts enough variance from the variable. The relationship between each variable and the underlying factor is expressed by the factor loadings. Here is an illustration (Table 5) of the output of a simple factor analysis that looked at wealth indicators, with only six input variables and two resulting factors.

Table 5: Illustration of the output of a simple factor analysis

Variables	Factor 1	Factor 2
Income	0.65	0.11
Education	0.59	0.25
Occupation	0.48	0.19
House value	0.38	0.60
Number of public parks in neighbourhood	0.13	0.57
Number of violent crimes per year in neighbourhood	0.23	0.55

The variable with the strongest association to the underlying latent variable. Factor 1, is income, with a factor loading of 0.65. Since factor loadings can be thought of as standardised regression coefficients, it is also possible to state that Factor 1 and Variable Income have a correlation of 0.65. This is regarded as a strong association for a factor analysis in most research fields.

Education and occupation are two additional factors that are linked to Factor 1. We might refer to it as "Individual socioeconomic status" based on the variables that load heavily onto Factor 1. However, the other factor, Factor 2, has high factor loadings for house value, the number of public parks, and the quantity of violent crimes per year. They appear to indicate the overall wealth of the neighbourhood, so we might call Factor 2 as "Neighbourhood socioeconomic status." Notably, Factor 1's loading of 0.38 indicates that the variable house value is also only tangentially significant. This makes sense because a person's income ought to be correlated with the value of his or her home.

Example - Let's say a psychologist believes there are two types of intelligence, "verbal intelligence" and "mathematical intelligence," neither of which can be directly observed. [note 1] The psychologist looks for evidence for the hypothesis in the test results from each of the 1000 students across 10 different academic fields. If each student is chosen at random from a large population, then the ten scores assigned to each student are random variables. The psychologist's hypothesis might be that for each of the 10 academic fields, the average score across all students who have some common value for their verbal and mathematical "intelligences" is some constant times their level of verbal intelligence plus another constant times their level of mathematical intelligence, or that it is a linear combination of those two "factors." The hypothesis postulates that the numbers for a given subject, by which the two types of intelligence are multiplied to obtain the expected score, are the same for all intelligence level pairs and are referred to as "**factor loading**" for this subject. The assumption might, for instance, be that the predicted average student's aptitude for the subject of astronomy is

{10 × the student's verbal intelligence} +
{6 × the student's mathematical intelligence}.

The factor loadings associated with astronomy are 10 and 6. Other academic subjects' factor loadings may differ. Because individual aptitudes vary from average aptitudes (predicted above) and due to measurement error itself, two students who are assumed to have the same levels of verbal and mathematical intelligence may have different measured aptitudes in astronomy. Such variations make up the "error," a statistical term that refers to the degree to which an individual deviates from what is typical of or predicted by his or her levels of intelligence. The 10 scores of each of the 1000 students, or a total of 10,000 numbers, would be the observable information used in factor analysis. Each student's factor loadings and levels for each of the two types of intelligence must be inferred from the data.

7.7.3. Eigenvalues

Eigenvalues are also known as characteristic roots. A factor eigen value is the sum of the square of its factor loading. Eigenvalues show the percentage of variance explained by a specific factor out of the total variance. The commonality column tells us how much of the total variance is explained

by the first factor. For example, if our first factor explains 68% of the total variance, the other factor will explain 32% of the variance.

7.7.4. Communality (h^2)

Community, represented by h^2, displays the proportion of each variable that is explained by the underlying factor as a whole. When communality is high, little of the variable remains after the factors' respective contributions have been taken into account. The calculation for each variable is as follows:

h^2 of the *i*th variable = (*i*th factor loading of factor A) ^{2} +
(*i*th factor loading of factor B) ^{2} + …

7.7.5. Factor score

The term "factor score" can also refer to the "component score." This score includes every row and every column, serving as an index of all the variables and allowing for further analysis. We can standardise this score by multiplying it by a common term. Any analysis we perform will be based on this factor score and assume that all variables will behave and move as factor scores.

7.7.6. Criteria for determining the number of factors

According to the Kaiser Criterion, Eigenvalues is a good criterion for determining a factor. If Eigenvalues is greater than one, it should be considered as a factor; if Eigenvalues is less than one, it should not. It should be greater than 0.7 in accordance with the variance extraction rule. If variance is less than 0.7, it shouldn't be considered.

7.8 PERFORMING FACTOR ANALYSIS

The objective of a factor analysis as a data analyst is to lower the number of variables needed to explain and interpret the findings. Two steps can be taken to complete this:

- Factor Extraction and
- Factor Rotation

Factor extraction entails deciding on the type of model to use as well as the number of factors to extract. After the factors have been extracted, factor rotation occurs with the goal of achieving simple structure in order to improve interpretability.

7.8.1. Types of factoring:

There are different types of methods used to extract the factor from the data set:

i. ***Principal component analysis (PCA):*** This is the most commonly used method among researchers. PCA begins by extracting the

maximum variance and storing it in the first factor. Following that, it removes the variance explained by the first factors and begins calculating the maximum variance for the second factor. This procedure is repeated until all of the factors have been chosen. All of the factors chosen explain the most residual variance in the entire set of standardised response scores. Using a covariance matrix, principal component analysis involves locating the variables with the greatest variance. The differences and correlations between a set of variables are shown visually in a covariance matrix. Each variable is compared by being given a score between 0 and 1, with 0 denoting no relationship and 1 denoting a relationship. It might be possible to combine two variables into one factor if their scores are close to one. An illustration of a covariance matrix is given below:

VariableA**B**C**D** A1.000.250.450.95 B0.251.000.550.65 C0.450.551.000.75 D**0.950.650.751.00**

In the preceding example, the variable "A" has the highest correlation with "D," so a statistician might include both in the same factor. In contrast, because the variables "A" and "B" have a low correlation, the statistician may exclude them from the same factor.

ii. ***Common factor analysis***: The second most popular method among researchers, it extracts common variance and organises it into factors. This method does not account for the individual variance of all variables. SEM makes use of this technique. In common factor analysis, correlations between variables are calculated rather than variables with the greatest possible variance between them, as in principal component analysis. The most highly correlated variables are found using covariance matrices, and these variables are then put together into a factor using common factor analysis. For instance, a study on the similarities between twins might discover a connection between genetics and physical characteristics.

iii. ***Image factoring:*** The correlation matrix serves as the foundation for this method. In image factoring, the OLS Regression method is used to predict the factor. In order to create precise measures of covariance between variables, image factoring uses image theory. An examination of the relationship between the human cognitive process and decision-making skills is done using the behavioural theory known as image theory. It is especially useful in psychological or social studies because this method of factor analysis combines psychological ideas with statistical correlations. For example, a statistician might use image factoring to assess the variables involved in human career decisions.

iv. ***Maximum likelihood method:*** This method also uses a correlation metric, but it factors using the maximum likelihood approach and has the advantage to analyze statistical models with different characters on the same basis.

v. ***Least-squares method:*** The least-squares method involves minimising the sum of squared correlational differences. The sum of squared differences is a statistical measure that compares observed variables to predicted values. During factor analysis, you can use two types of least-squares methods: weighted and unweighted. The weighted least-squares method involves weighing correlations by the inverse of their uniqueness so that variables with a high level of uniqueness have a greater weight. This weight can then be considered when determining factors. In contrast, the unweighted least-squares method does not take into account the weight of variable uniqueness. When many of the observed variables are similar, for instance, you might use the unweighted least-squares approach.

vi. ***Principal axis factoring:*** Principal axis factoring entails the construction of numerous covariance matrices, each of which can enhance the precision of the matrix that follows. To use this factor analysis method, first generate an initial covariance matrix. Next, identify the variables in the matrix that might be factors and assign squared correlation coefficients to those variables. You could create a new covariance matrix using this. You can keep doing this until there are barely any differences between the subsequent matrices. If you want to achieve a certain level of accuracy, think about setting a goal. For instance, you could iterate matrices until the values differ by less than 0.05.

vii. ***Other methods of factor analysis:*** Alfa factoring outperforms least squares. Weight square is another regression-based factoring method.

7.8.2. Types of Rotation:

In the context of factor analysis, rotation is analogous to staining a microscope slide. Different rotations reveal different structures in the data, just as various stains on it reveal various structures in the tissue. Even though different rotations produce results that seem to be completely different, all results are treated equally from a statistical perspective, neither being superior to nor inferior to others. The correct rotation must be chosen, though, in order to make sense of the factor analysis results. If the factors are unrelated, an orthogonal rotation is performed; if they are, an oblique rotation is performed. Although the Eigen values will change as a result of rotation, communality for each variable will remain unaffected. The rotation method makes it easier to understand the output. Eigenvalues have no effect on the rotation method, but the rotation method has an effect on the extracted Eigenvalues or percentage of variance.

Rotation techniques can be divided into two categories: ***oblique and orthogonal*** (based on the angle maintained between the X and Y axes). While oblique methods allow the X and Y axes to assume an angle other than 90^o, orthogonal rotations produce factors that are uncorrelated (i.e., maintain a 90^o angle between axes). Because uncorrelated factors are easier to interpret, researchers have traditionally been advised to use orthogonal rotation. Another argument in favour of orthogonal rotation is that the math

is easier, which made a big difference in the early 20th century when calculations were mostly done by hand or with much less computing power. In most statistical computing packages, orthogonal rotations are typically set to the default.

There doesn't seem to be a strong argument for why contemporary researchers should always use orthogonal rotations. Since behaviour is rarely divided into neatly packaged units that work independently of one another, we typically expect some correlation among factors in the social sciences (and many other sciences, such as the biomedical sciences). Therefore, when factors are correlated, using orthogonal rotation might lead to a less practical solution.

There are various rotation methods available:

i. ***Varimax rotation method*** **-** an orthogonal rotation technique that reduces the number of variables that each factor is heavily loaded with. The interpretation of the factors is made easier by this technique.

ii. ***Quartimax rotation method*** **-** a rotational approach that reduces the number of variables that must be considered to explain each variable. The analysis of the observed variables is made easier by this method.

iii. ***Equimax Method*** **-** a rotation technique that combines the quartimax and varimax methods to simplify the variables and factors, respectively. It is best to reduce the number of variables that heavily influence a factor and the number of factors required to fully explain a variable.

iv. ***Direct oblimin rotation method*** **-** A technique for oblique (nonorthogonal) rotation. Solutions are most oblique when delta is equal to 0 (the default value). The factors become less indirect as delta decreases. To override the default delta of 0, enter a value less than or equal to 0.8.

v. ***Promax rotation method*** **-** an oblique rotation that makes it possible to correlate variables. Large datasets can benefit from this rotation because it can be calculated more quickly than a direct oblimin rotation.

7.9 TYPES OF FACTOR ANALYSIS

Factor analysis is classified into two types: exploratory and confirmatory. The history of exploratory factor analysis techniques is much longer than that of confirmatory factor analysis techniques. Different approaches lead to different applications (e.g., theory development versus theory confirmation).

7.9.1. Exploratory factor analysis (EFA)

It is assumed that any indicator or variable can be linked to any factor. This is the most commonly used factor analysis method among researchers, and

it is not based on any prior theory. The objective of exploratory factor analysis (EFA) is typically to allow the data to reveal the relationships between a set of variables. Although a researcher using an EFA may have a theory explaining how the variables relate to one another, the basic factor model in an EFA has relatively few limitations. For more than a century, this kind of analysis has been helpful in the development and discussion of theories. Exploratory factor analysis is especially useful in the early stages of theory development and scale or test development. First, EFA can be used to reduce data when the relationships between variables are not known in advance. When using EFA, a researcher employs inductive reasoning by starting with a number of observations and drawing theory from them. In the example given above (Table 3 & 4) of high-performing employee characteristics, six personality variables were eventually reduced to two. Thus, instead of the six initial variables, only two variables need to be discussed in subsequent analyses. Data reduction is particularly helpful in addressing the issues associated with multi-collinearity (excessively high correlations) among a group of predictors. A second advantage of EFA is the ability to detect a general factor. One general factor often emerges along with several specific factors when several different cognitive ability tests are factor analysed. For instance, all aptitude tests have some correlation with the general factor of intelligence, or g, when used to measure intelligence.

Finally, because it enables the researcher to establish the dimensionality of the test and identify cross-loadings, EFA is especially helpful in scale or test development (correlations of variables with more than one factor). In general, cross-loadings are undesirable. It is advantageous when developing scales to have items that are specific to one factor. The three variables that represent conscientiousness in the previous example do not cross-load onto intellect or vice versa.

7.9.2. Confirmatory factor analysis (CFA):

Confirmatory factor analysis (CFA) is used to determine the factor and factor loading of measured variables, as well as to confirm what is expected based on the fundamental or pre-established theory. Confirmatory factor analysis (CFA) is used to test theories that are derived from a set of data. Equation 1's fundamental factor model is still valid, but certain limitations are imposed because of the specific theoretical model being examined. In the previous example, for instance, one could use CFA to impose constraints on the factor pattern to forbid cross-loadings. CFA is a more recent statistical advancement having been developed in the 1960s than EFA (developed in 1904).

In a deductive reasoning process, confirmatory factor analysis is particularly helpful. When using CFA, precise hypothesis testing is possible. A researcher might, for instance, discuss the statistical importance of specific factor loadings. In the previous example, one could determine with statistical certainty how closely the observed variable, imagination, is correlated with the latent factor, intellect, given the low correlation.

A set of data can be analysed using CFA to determine whether two factors or just one factor (or any other numeric combination) underlies the data. Unlike CFA, where models can be explicitly compared through statistical testing under the null hypothesis, EFA relies on rules of thumb and intuition, which can mislead researchers. CFA can also be used to determine whether various components of the basic factor model within a specific data set are equivalent. For example, one could hypothesise that all of the variables observed for intelligence are equally related to intelligence. By placing restrictions on the loadings in the basic factor model using CFA, it is possible to determine whether these relationships are equivalent (i.e., L in Equation 1).

It's crucial to check whether results of a factor analysis are consistent across demographic categories. Confirmatory factor analysis enables tests of invariance, or the similarity of factor structure, loadings, and uniqueness, across various groups of people (such as ethnic, gendered, and cultural ones). If responses from supervisors of manufacturing workers were contrasted with responses from supervisors of service workers, a researcher might be curious to know whether the same hypothetical factor structure would emerge. It's possible that the observed variables for the two groups do not relate to the latent factors in the same way. For instance, for service workers as opposed to manufacturing workers, the observed indicator systematic may be less related to the factor conscientiousness. The hypothesis that the correlations from the two groups are the same or different can be tested in a CFA that is evaluating the equivalence of factor loadings.

Compared to EFA, confirmatory factor analysis has more flexibility in control. When using CFA, some factors can be designated as oblique (correlated with one another), while others can be designated as orthogonal (uncorrelated with one another). The factors are interpreted as either oblique or orthogonal within a single EFA, but not as a combination of the two. Additionally, subject to theory, CFA enables the researcher to flexibly impose additional constraints (e.g., allowing correlated uniqueness). EFA has the advantage that no such theoretical restrictions or requirements are required. Therefore, EFA might be a better option if none are available.

CFA assumes that each factor is linked to a subset of measured variables. It typically employs two approaches:

i. ***The traditional method:*** Instead of using common factor analysis, the traditional factor method is based on the principal factor analysis method. The researcher can learn more about insight factor loading using the conventional method.

ii. ***The SEM approach:*** A different method of factor analysis that can be carried out in SEM is CFA. In SEM, we only add the arrow that must observe the variable that represents the covariance between each pair of latent variables and remove all straight arrows from the latent variable. In addition, we will leave the straight arrows error-free and disturbance terms to the corresponding variables. If the SEM's

standardised error term is less than the absolute value of two, the factor is thought to be well-explained; if it is greater than two, there is still some unexplained variance that can be accounted for by a factor. To evaluate how well the model fits, chi-square and a number of other goodness-of-fit indices are used. SEM approach would be discussed in detail in the next chapter.

Lastly, let's discuss some applications of factor analysis in the field of psychology.

7.10 APPLICATIONS IN PSYCHOLOGY

Factor analysis is used to identify "factors" that explain a wide range of test results. For instance, intelligence studies have shown that people who perform well on verbal aptitude tests also perform well on other verbal aptitude tests. This was clarified by researchers by using factor analysis to isolate one factor, popularly known as verbal intelligence, which measures how well someone can use verbal skills to solve problems. In psychology, factor analysis is most commonly associated with intelligence research. However, it has also been employed to identify variables in a variety of areas, including personality, attitudes, beliefs, and so forth. It is related to psychometrics because it can determine whether an instrument is valid by determining whether it actually measures the postulated factors. Factor analysis is a technique that is frequently used in cross-cultural research. It is used to extract cultural dimensions.

7.11 SUMMARY

- Factor analysis is used largely when the researcher has substantial numbers of variables seemingly measuring similar things. It has proven particularly useful with questionnaires.

- It examines the pattern of correlations between the variables and calculates new variables (factors) which account for the correlations. In other words, it reduces data involving a number of variables down to a smaller number of factors which encompass the original variables.

- Factors are simply variables. The correlations of factors with the original variables are known as factor loadings, although they are merely correlation coefficients. Hence, they range from −1.0 through 0.0 to +1.0. It is usual to identify the nature of each factor by examining the original variables which correlate highly with it. Normally each factor is identified by a meaningful name.

- Because the process is one of reducing the original variables down to the smallest number of factors, it is important not to have too many factors. The scree plot may be used to identify those factors which are likely to be significantly different from a chance factor.

- Factors are mathematically defined to have the maximum sum of squared factor loadings at every stage. They may be more easily interpreted if they are rotated. This maximises the numbers of large factor loadings and small factor loadings while minimizing the number of moderate factor loadings, making interpretation easier.

- Factor scores provide a way of treating factors like any other variable. They are similar to standard or *z*-scores in that they have symmetrical numbers of positive and negative values and their mean is 0.00. They can be used to compare groups in terms of their mean factor scores.

7.12 QUESTIONS

1. What is factor analysis and explain the types of factor analysis?
2. What are the applications of factor analysis?
3. Discuss the assumptions and purpose of factor analysis.
4. What are the advantages and disadvantages of Factor Analysis?
5. Write a note on the different types of methods used to extract factors from the data set.
6. Explain rotation and its various techniques.

7.13 REFERENCES

1. Daniel, W. W. (1995). *Biostatistics*. (6th Ed.). N.Y.: John Wiely.
2. Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. NY: Sage. 3
3. Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 143-164). Hoboken, NJ: Wiley.
4. Howitt, D. & Cramer, D. (2011). Factor analysis: Simplifying complex data. Introduction to Statistics in Psychology (5^{th} Ed.). Pearson.

❄❄❄❄❄❄❄

8

FACTOR ANALYSIS AND SOFTWARE PACKAGES – II

Unit Structure:

8.0 OBJECTIVES

After studying this unit, you should be able to:

- Understand Structural equation modelling (SEM) and its assumptions.
- Explain the steps to SEM.
- Discuss the theoretical and practical limitations of SEM.
- Describe R syntax and its pros and con.
- Demonstrate how to download, install, create, execute, save, and run R.

- Explain the three components of R programme - variables, comments, and keywords.
- Illustrate graphs in R
- Explain how to conduct Descriptive, Basic and Multivariate Statistics In R
- Explain R GUI, Other Software

8.1 INTRODUCTION TO STRUCTURAL EQUATION MODELLING (SEM)

General Purpose and Description

Structural equation modelling (SEM) is a set of statistical techniques that allows the examination of a set of relationships between one or more IVs, either continuous or discrete, and one or more DVs, either continuous or discrete. IVs and DVs can both be considered factors or measured variables. Structural equation modelling is also known as *causal modeling, causal analysis, simultaneous equation modeling, analysis of covariance structures, path analysis, and confirmatory factor analysis.* The latter two are actually subsets of SEM.

Multiple regression analyses of various factors can be used to answer questions using SEM. SEM is created by combining exploratory factor analysis (EFA) and multiple regression analysis. At the most fundamental level, a researcher proposes a correlation between one measured variable (let's say, graduate school success) and other measured variables (let's say, undergraduate GPA, gender, and average daily caffeine consumption). Figure 8.1's diagram depicts a multiple regression using a straightforward model. The measured variables are all represented as boxes with lines and arrows connecting them, showing that GPA, gender, and caffeine (the IVs) are the three factors that predict success in graduate school (the DV). An IV correlation is shown by a line with two arrows. An imperfect prediction is indicated by the presence of a residual.

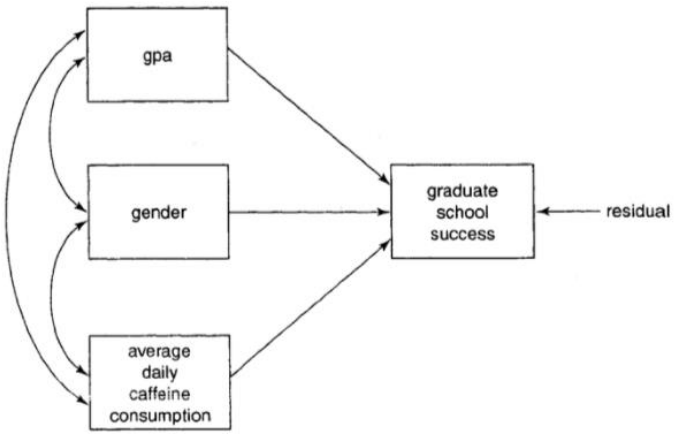


Figure 8.1: Path diagram of multiple regression

Source:- Tabachnick, B.G & Fidell, L.S.(2007). Structural Equation Modeling. *Using Multivariate Statistics (5th Ed.).* Pearson. 676-780.

Figure 8.2 depicts a more complex model of graduate school success. Graduate School Success is a latent variable (a factor) in this model, which is assessed indirectly by looking at three measured variables: the number of publications, academic performance, and faculty evaluations. In turn, gender (a measured variable) and undergraduate success (a second factor measured by undergraduate GPA, faculty recommendations, and GRE scores) are predictive factors for graduate school success (three additional measured variables). For the sake of readability, names of factors are written in initial capital letters, while names of measured variables are written in lowercase letters.

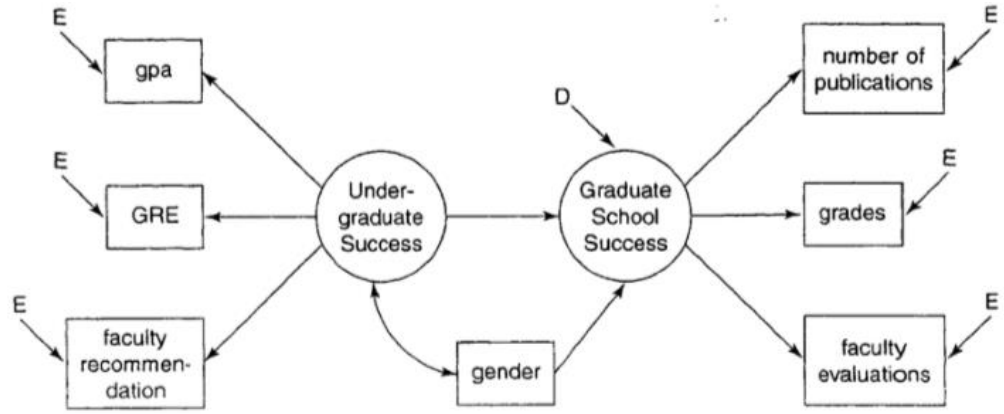


Figure 8.2: - Path diagram of a structural model.

Source:- Tabachnick, B.G & Fidell, L.S.(2007). Structural Equation Modeling. *Using Multivariate Statistics (5th Ed.).* Pearson. 676-780.

Path diagrams are shown in Figures 8.1 and 8.2. These diagrams are essential in SEM because they enable the researcher to depict the hypothesised set of relationships—the model. The diagrams aid in clarifying a researcher's ideas about variable relationships, and they can be directly translated into the equations required for the analysis.

SEM diagrams are created using a variety of conventions. Squares or rectangles are used to represent measured variables, also known as *observed variables, indicators, or manifest variables.* Factors are also known as *latent variables, constructs, or unobserved variables* because they have two or more indicators. In path diagrams, factors are represented by circles or ovals. Lines indicate relationships between variables; the absence of a line connecting variables indicates that no direct relationship has been hypothesised. Lines can have one or two arrows. A line with one arrow represents a proposed direct relationship between two variables, with the variable with the arrow pointing to it being the DV. A line with an arrow at both ends denotes an unanalyzed relationship, which is simply a covariance between two variables with no implied direction of effect.

In Figure 8.2, Graduate School Success is a latent variable (factor) that is predicted by gender (a measured variable) and Undergraduate Success (a factor). Take note of the line with an arrow at both ends that connects Undergraduate Success and Gender. This line with an arrow at both ends suggests a relationship between the variables but makes no predictions

about the direction of effect. Take note of the arrows connecting the Graduate School Success construct (factor) to its indicators as well: The construct *predicts* the variables that are measured. The implication is that the number of publications, grades, and faculty evaluations of graduate students is driven or created by Graduate School Success. Because measuring this construct directly is impossible, we do the next best thing and measure several indicators of success. We hope that by measuring a variety of observable indicators, we will be able to tap into graduate students' true level of success. This follows the same logic as factor analysis.

Figure 8.2 shows the DVs for GPA, GRE, faculty recommendations, graduate school success, publications, grades, and faculty evaluations. They are all indicated by one-way arrows. Undergraduate Success and Gender are IVs in the model. They are not pointed at by any one-way arrows. All of the DVs, both observed and unobserved, have arrows with the letters "E" or "D" pointing in their direction. Ds (disturbances) point to latent variables while Es (errors) point to measured variables (factors). There is always some residual or error, just as there is in multiple regression. In SEM, the diagram containing these paths includes the residual not predicted by the IV(s).

The portion of the model that connects the factors and measured variables is sometimes referred to as the *measurement model*. The *measurement model* in this instance is made up of the two constructs (factors), Undergraduate Success and Graduate School Success, as well as their respective indicators. The term "*structural model*" refers to the proposed connections between the constructs, in this case, the single pathway between Undergraduate Success and Graduate School Success.

Note that neither of the models up to this point has included hypotheses about means or mean differences, only about how variables relate to one another (covariances). The SEM framework can also be used to test mean differences related to group membership.

The researcher prefers the SEM method because it estimates multiple and interrelated dependences in a single analysis. This analysis employs two types of variables: endogenous variables and exogenous variables. Endogenous variables are the same as dependent variables and have the same value as the independent variable.

Theory: This is a set of relationships that provide consistent and comprehensive explanations for actual phenomena. Models are classified into two types:

- *Model of measurement:* The theory is represented by the measurement model, which specifies how measured variables are combined to represent the theory.
- *Structural model:* Represents the theory that demonstrates how constructs are related to one another. Because it tests the proposed casual relationships, structural equation modelling is also known as casual modelling.

Next, let us look at the assumptions.

8.2 ASSUMPTIONS OF SEM

Assumptions are made as follows:

- **Multivariate normal distribution:** For multivariate normal distribution, the maximum likelihood method is used and assumed. Small differences in multivariate normality can result in large differences in the chi-square test.
- **Linearity**: Endogenous and exogenous variables are assumed to have a linear relationship.
- **Outlier**: Outliers should be avoided in data. Outliers have an impact on the model's significance.
- **Sequence**: Endogenous and exogenous variables must have a cause and effect relationship, and the cause must occur before the event.
- **Non-spurious relationship**: The observed covariance must be true.
- **Model identification**: Equations must be greater than the estimated parameters, or models will be over- or under-identified. Models that have been identified are not considered.
- **Sample size:** The majority of researchers prefer a sample size of 200 to 400 with 10 to 15 indicators. As a general rule, there are 10 to 20 times as many cases as variables.
- **Uncorrelated error terms**: Error terms are assumed to be uncorrelated with one another.
- **Data**: Interval data is used.

8.3 STEPS SEM

- **Defining individual constructs**: The constructs must first be defined theoretically. To assess the item, run a pretest. CFA is used to perform a confirmatory test on the measurement model.
- **Developing the overall measurement model:** The measurement model is also referred to as a path analysis. The relationships between exogenous and endogenous variables make up path analysis. An arrow is used to demonstrate this. The measurement model bases its predictions on the idea of unidimensionality. According to measurement theory, the error term is uncorrelated within the measured variables and latent constructs are what drive the measured variable. An arrow is drawn in a measurement model from the measured variable to the constructs.
- **Design the study to produce the empirical results:** The researcher must specify the model in this step. The study should be planned by the researcher to reduce the possibility of an identification issue.

Methods like order condition and rank condition are used to reduce the identification issue.

- **Assessing the measurement model validity:** The term "CFA" also refers to evaluating/ assessing the measurement model. A researcher uses CFA to contrast the theoretical measurement with the reality model. The validity of the constructs must be connected to the CFA's outcome.

- **Specifying the structural model:** In this phase, structural connections between constructs are drawn. No arrow can enter an exogenous construct in the structural model. A proposed structural relationship between two constructs is represented by a single-headed arrow. This demonstrates the causal connection. Each proposed relationship employs a single degree of latitude. The model may or may not be recursive.

- **Examine the structural model validity:** The structural model validity is examined in the final step by the researcher. If the chi-square test result is insignificant and at least one incremental fit index (such as CFI, GFI, TLI, AGFI, etc.) and one badness of fit index (such as RMR, RMSEA, SRMR, etc.) meet the predetermined criteria, the model is said to be well-fit.

8.4 LIMITATIONS TO STRUCTURAL EQUATION MODELING

8.4.1 Theoretical Issues

Unlike exploratory factor analysis, SEM is a confirmatory method. It is most frequently used to test a theory, even if it is just a personal theory. Without prior knowledge of, or hypotheses regarding, potential relationships among variables, one cannot perform SEM. Perhaps the biggest distinction between SEM and other methods—and one of its strongest points—is this. Planning a SEM analysis that is theory-driven is crucial.

SEM is a confirmatory technique, but after a model has been estimated, there are ways to test a variety of different models (models that test particular hypotheses or might provide better fit). However, if numerous iterations of a model are tested in an effort to identify the model that fits the data the best, the researcher has switched to exploratory data analysis, and appropriate precautions must be taken to avoid inflated Type I error levels. Finding the best model is acceptable as long as significance levels are carefully considered and cross-validation with a different sample is carried out whenever practical.

Using SEM for exploratory work without the required controls has contributed to SEM's negative reputation in some circles. It might also be because structural equation modelling is sometimes referred to as causal

modelling. The application of SEM is not causal in the sense of implying causality. It is a design issue, not a statistical one, to attribute causality.

Unfortunately, SEM is frequently considered to be only appropriate for correlational or nonexperimental designs. This is far too restrictive. Both experimental and nonexperimental designs can use SEM, just like regression. In fact, there are a few benefits to using SEM for experiment analysis: It is possible to test mediational processes and include information about the effectiveness of the manipulations in the analysis (Feldman, Ullman, & Dunkel-Schetter, 1998).

8.4.2 Practical Issues

i. ***Sample Size and Missing Data*** **-** When covariances are estimated from small samples, they are less stable than correlations. Based on covariances, SEM is used. The sensitivity of parameter estimates and chi-square tests of fit to sample size is also well known. SEM is a large sample method, just like factor analysis. According to Velicer and Fava (1998), a good factor model in exploratory factor analysis depends on the size of the factor loadings, the number of variables, and the sample size. SEM models can be used in place of this. Less participants may be needed for models with strong expected parameter estimates and trustworthy variables. New test statistics have been developed that enable the estimation of models with as few as 60 participants, despite the fact that SEM is a large sample technique (Bentler & Yuan, 1999). MacCallum, Browne, and Sugawara (1996) present tables of the minimal sample sizes required for tests of the goodness of the tit in order to estimate adequate sample size for power calculations. These tables use the model degrees of freedom and effect size to estimate sample sizes.

ii. ***Multivariate Normality and Outliers*** **-** The majority of SEM's estimation methods presuppose multivariate normality. Screen the measured variables for outliers, both univariate and multivariate, as well as the skewness and kurtosis of the measured variables, in order to ascertain the size and shape of nonnormally distributed data. Regardless of whether they are DVs or IVs, all measured variables are combined to check for outliers. (Some SEM software programmes check for skewness, kurtosis, and multivariate outliers.) Transformations can be tried if significant skewness is discovered, but frequently variables remain highly skewed or highly kurtotic even after transformation. Some variables, like those related to drug use, are not predicted to have a normal distribution in the population. An estimation method that addresses the nonnormality can be chosen if transformations fail to restore normality or if a variable is not anticipated to be normally distributed in the population.

iii. ***Linearity*** **-** SEM methods only look at linear relationships between variables. Although it is challenging to evaluate linearity among latent variables, scatterplots can be used to evaluate linear relationships between pairs of measured variables. When multiple regression is

used, the measured variables are raised to powers in order to account for any hypothesised nonlinear relationships among the measured variables. For instance, the square of average daily caffeine consumption is used if the relationship between graduate school success and average daily caffeine consumption is quadratic (a little caffeine is not enough, a few cups are good, but more than a few are harmful).

iv. ***Absence of Multicollinearity and Singularity*** **-** In SEM, matrices must be inverted. Therefore, the required matrices cannot be inverted if variables are perfect linear combinations of one another or are extremely highly correlated. Examine the covariance matrix's determinants if at all possible. A problem with multicollinearity or singularity may be indicated by an extremely small determinant. Typically, if the covariance matrix is singular, SEM programmes terminate and issue alerts. Check your data set if you receive a message like that. Inadvertently including linear combinations of variables is a common occurrence. Just eliminate the variable that is the singularity's cause. Create composite variables and use them in the analysis if true singularity is discovered.

v. ***Residuals*** **-** The residuals following model estimation ought to be small and centred at zero. The residual covariances' frequency distribution ought to be symmetrical. In the context of SEM, residuals are residual covariances rather than residual scores. SEM software offers residuals diagnostics: A poorly fitting model may be indicated by nonsymmetrically distributed residuals in the frequency distribution; the model is estimating some covariances well and others poorly. Even when the model fits reasonably well and the residuals seem to be symmetrically distributed and centred around zero, it occasionally happens that one or two residuals remain quite large. It is frequently beneficial to look at the Lagrange Multiplier (LM) test and think about including paths in the model when large residuals are discovered.

8.5 R: SYNTAX

R is a statistical computing and graphics system. It offers a programming language, high level graphics, interfaces to other languages, and debugging tools, among other things. For statistical computing and graphics, R is a free software environment. It is "open source," which means that unlike commercial software companies that protectively hide away the code on which their software is based, the individuals who developed R allow everyone access to their code. Anyone, anywhere can make contributions to the software thanks to the open source movement. As a result, R's capabilities grow dynamically as contributions come from all over the world. R exemplifies everything that is admirable about the World Wide Web.

Since its creation in the 1980s and widespread adoption in the statistical community, the R language is a dialect of S. John M. Chambers, the

project's lead designer, received the 1998 ACM Software Systems Award for S.

Although the language's syntax has a passing resemblance to C, its semantics are more closely related to those of the FPL (functional programming language) family and Lisp and APL. It specifically permits "computing on the language," which enables the creation of functions that accept expressions as input and are frequently useful for statistical modelling and graphics..

When using R interactively and running straightforward expressions from the command line, you can get quite far. Others will want to write their own functions either ad hoc to systematise repetitive work or with the perspective of writing add-on packages for new functionality. Some users may never need to go beyond that level.

The language's design includes a number of fine points and common pitfalls that may surprise the user. The majority of these arise from deeper consistency considerations. Additionally, there are many helpful idioms and shortcuts that the user can use to express fairly complex operations clearly. Once one is familiar with the underlying concepts, many of these come naturally. There may be several ways to complete a task in some circumstances, but some of the methods will depend on how the language is implemented while others operate at a higher level of abstraction. In these situations, we will specify the preferred usage.

R is basically just a base package with a fair amount of functionality. Once R has been downloaded and installed on your computer, you can begin creating graphs and data analysis. The beauty of R is that it can be enhanced by downloading packages that give the software extra functionality. A package can be created by anyone with a sufficiently large brain, some time, and effort. These packages are kept in a central repository known as the **CRAN** (Comprehensive **R** Archive Network) along with the software itself. Anyone with an Internet connection can use the CRAN to download and install packages, which they can then use with their own copy of R after being stored there. R is essentially a huge international family of kind-hearted, selfless individuals who are all working toward the common objective of creating a powerful data analysis tool that is accessible to all users for no cost. It's sort of like "give *p*s a chance," a statistical manifestation of The Beatles' idealistic vision of world peace, love, and humanity.

8.5.1 Pros and cons of R

R is a versatile and dynamic environment, and it is free, which are its two main benefits. There are many things you can do with it that are not possible with commercially available packages thanks to its open source format and statisticians' ability to contribute packages to the CRAN. Additionally, it is a tool that is rapidly expanding and is able to adapt quickly to new developments in data analysis. R is a very potent tool as a result of these benefits.

R's main drawback is its simplicity of use. R's guiding principle is to work with a command line rather than a graphical user interface (GUl). This simply means typing commands as opposed to pointing, clicking, and dragging objects with a mouse. R's written commands are, in my opinion, a much more effective way to work once you have mastered a few fairly simple things. At first, this might seem strange and like a rather "retro" method of working.

8.5.2 Base R vs. R packages

There are “default” packages that come with R. Some of these include:-

- `as.character`
- `print`
- `setwd`

And there are R packages developed and shared by others. Some R packages include:

- `tidyverse`
- `stargazer`
- `foreign`

8.5.3 Downloading and installing R

R is a free programming language and environment that is supported by the R Foundation for Statistical Computing. It is available under the GNU licence. The language's potent statistical and data interpretation capabilities are its most well-known features.

Installing the R environment on your computer and using an IDE (Integrated development environment) to run the language are prerequisites for using the R language (can also be run using CMD on Windows or Terminal on Linux).

8.5.3.1 Why use R Studio?

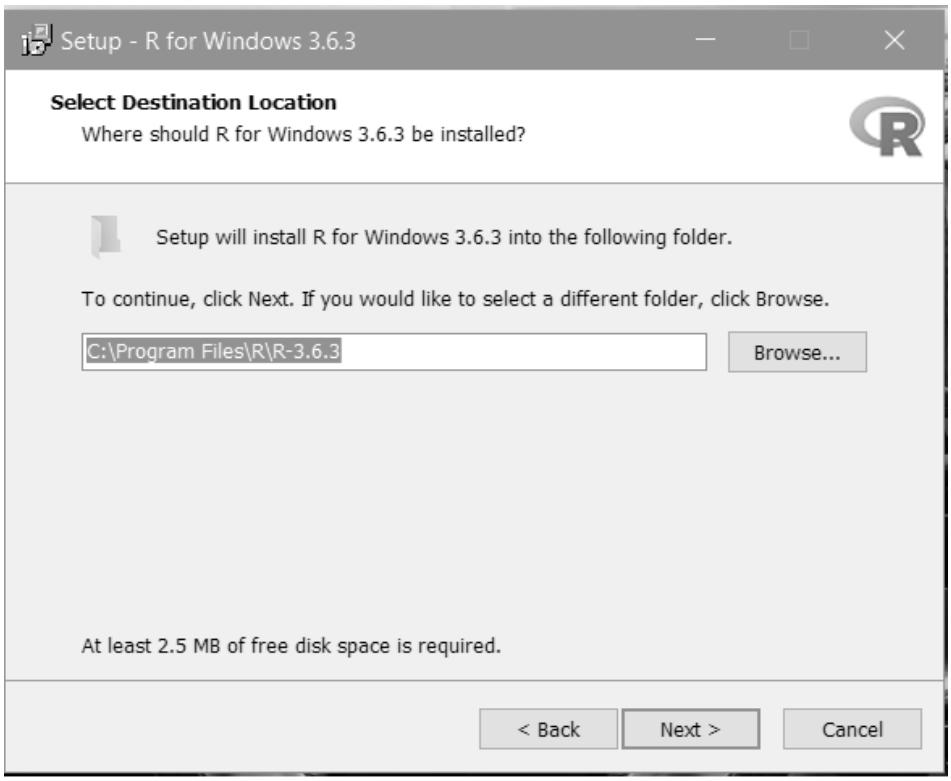
- It is a potent IDE designed specifically for the R programming language.
- Offers literate programming tools, essentially enabling the incorporation of R outputs, reports, text, and images into Word documents, HTML files, and reports.
- We can include interactive content in reports and presentations by using Shiny, an open-source R package.

8.5.3.2 Installing R Studio on Window

The steps below should be followed in order to install R Studio on Windows.

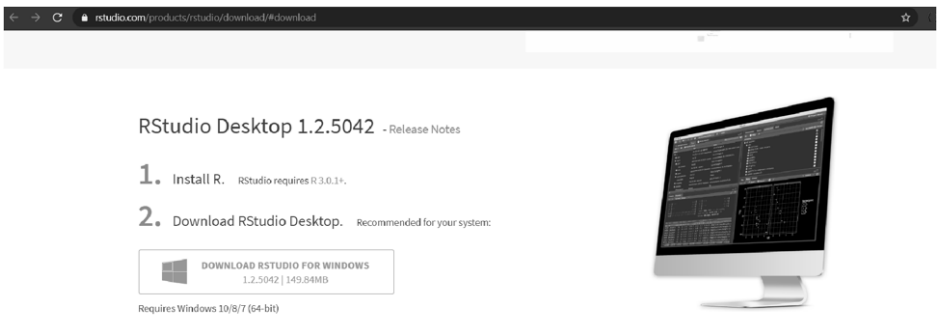
Step 1: First, create a R environment on your local machine. The same can be downloaded from ***r-project.org***

Step 2: Download R for the Windows operating system and double-click it to install it.



Step 3: Download R Studio from their official page.

Note: It is free of cost (under AGPL licensing).

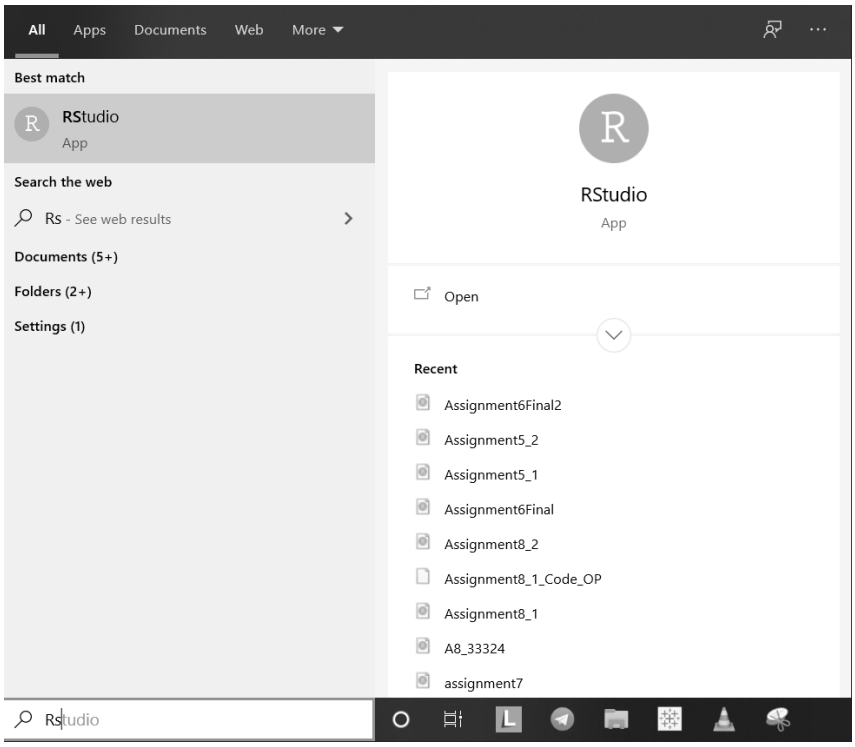


Step 4: Once the download is complete, look in your Downloads folder for a file with the name "RStudio-1.x.xxxx.exe."

Step 5: Double-click the installer to complete the software installation.

Step 6: Verify the installation of R Studio.

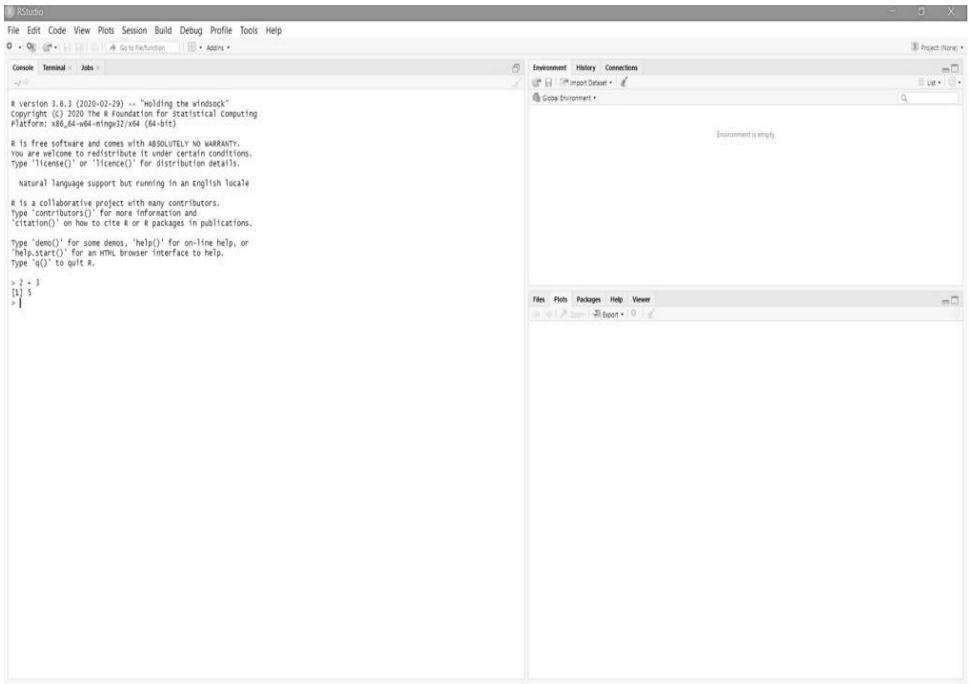
- Search for RStudio in the Window search bar on Taskbar.



• Launch the programme.

• Enter the code listed below into the console.

```
Input  : print('Hello world!')
Output : [1] "Hello world!"
```



Step 7: Your installation is successful.

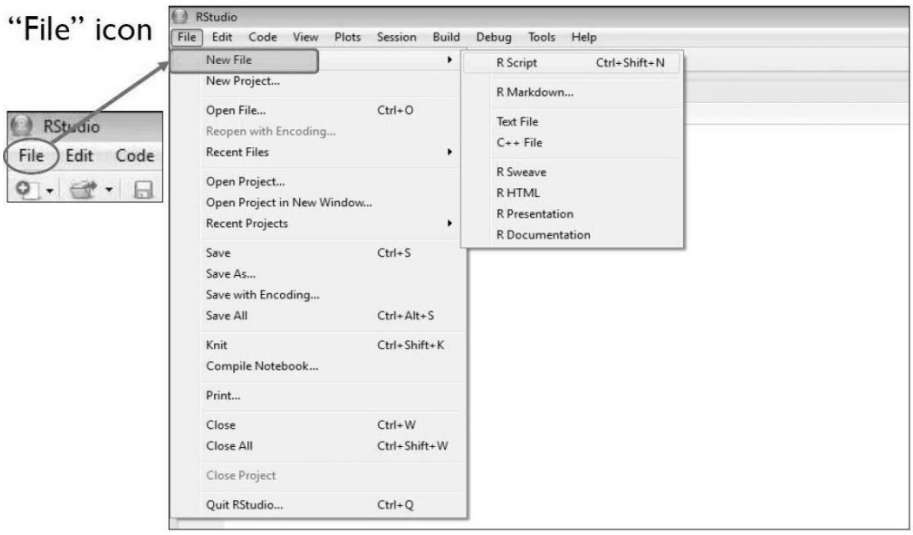
8.5.3.3 Creation and Execution of R File in R Studio

An integrated development environment (IDE) for R is called R Studio. The IDE is a graphical user interface where you can write your quotes, view the results, and also see the variables created during programming. R is available as client and server versions of Open Source software.

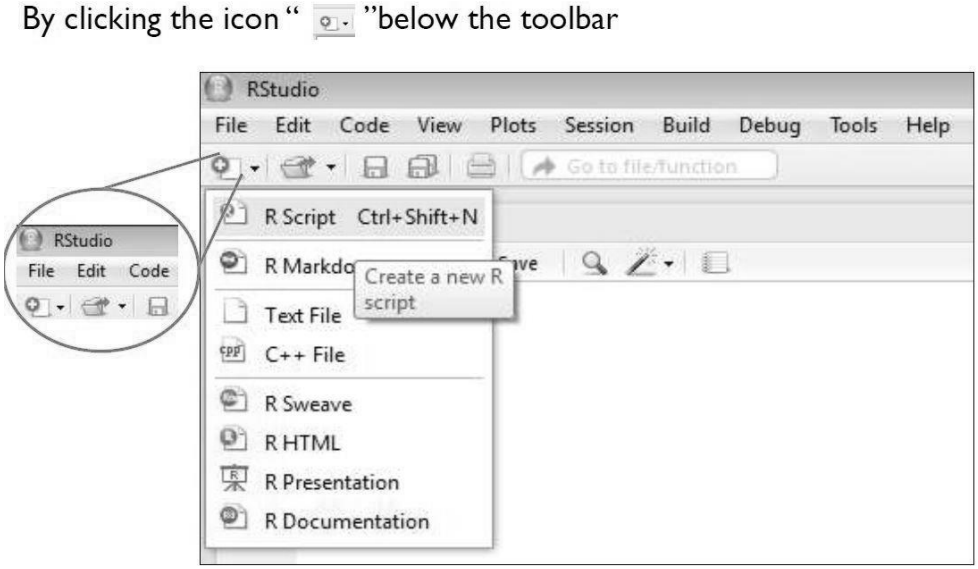
Creating an R file

In R studio, there are two ways to create a R file:

- You can select a new file and a R script from the drop-down menu that appears when you click the File tab, which will result in the opening of a new file.



- To open a new R script file, click the plus button, which is located directly beneath the file tab.

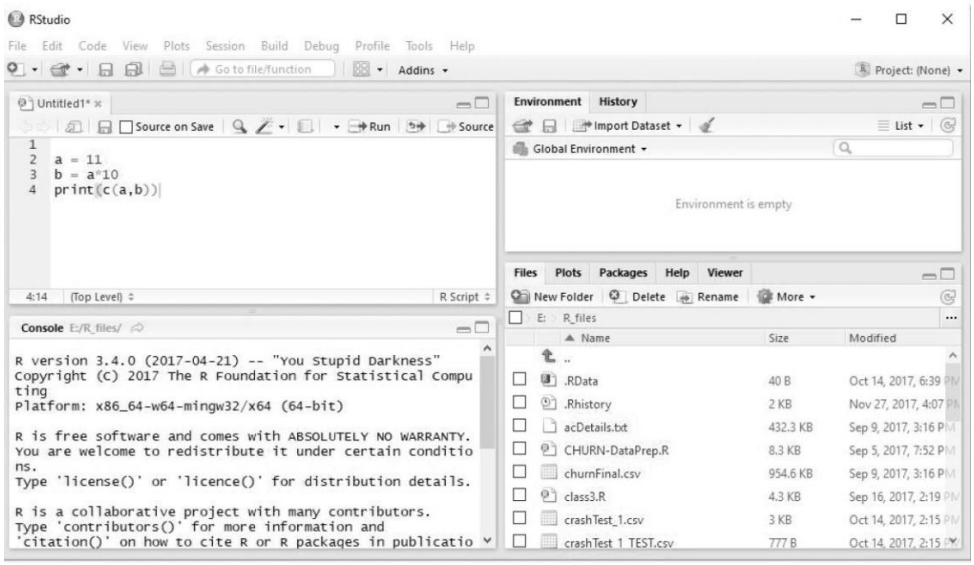


This is how a R Studio with the script file open appears once a R script file is opened.

There are three panels: console, environment/history, and file/plots. A new window, which is currently being opened as a script file, can be found on top left. You are now prepared to use R Studio to create a script file or programme.

Writing Scripts in an R File

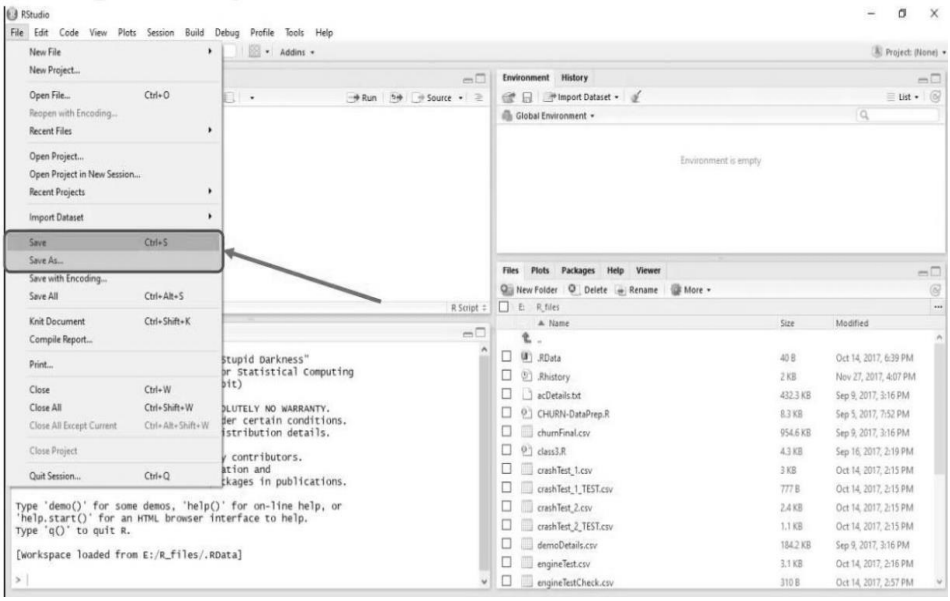
Here is an illustration of writing scripts to an R file:



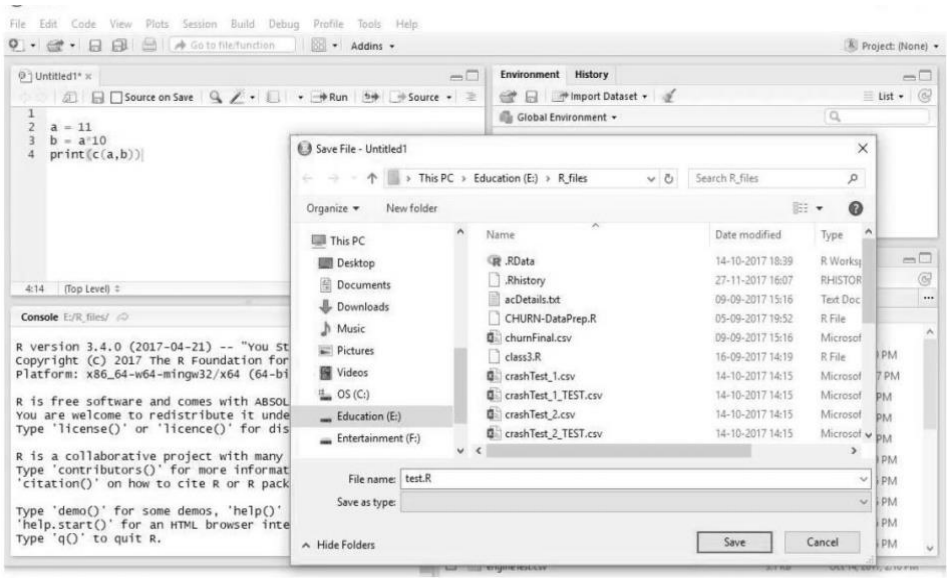
The first line of the code in the above example assigns the value 11 to the variable "a," and the second command, "b," is "a" multiplied by 10. Here, the code multiplies the value of a by 10 and assigns the result to the variable b. The third statement, print(c(a, b)), concatenates these two variables and prints the outcome. And thus, this is how a script file in R is created. It is necessary to save a script file after writing it before running it.

Saving an R File

Let's look at saving a R file. If you select the file tab from the file menu, you can choose between the save and save as buttons. If you click the save button when you want to save a file, it will automatically save the file with an untitled x. Therefore, depending on how many R scripts you have already opened, this x could be either 1 or 2.



Alternatively, it's a good idea to use the Save as button, which is located just below the Save button, to rename the script file as you please. Let's assume that we chose to save the file. A window similar to this will appear, allowing you to rename the script file to **test.R.** After renaming, you can save the script file by selecting the Save button.

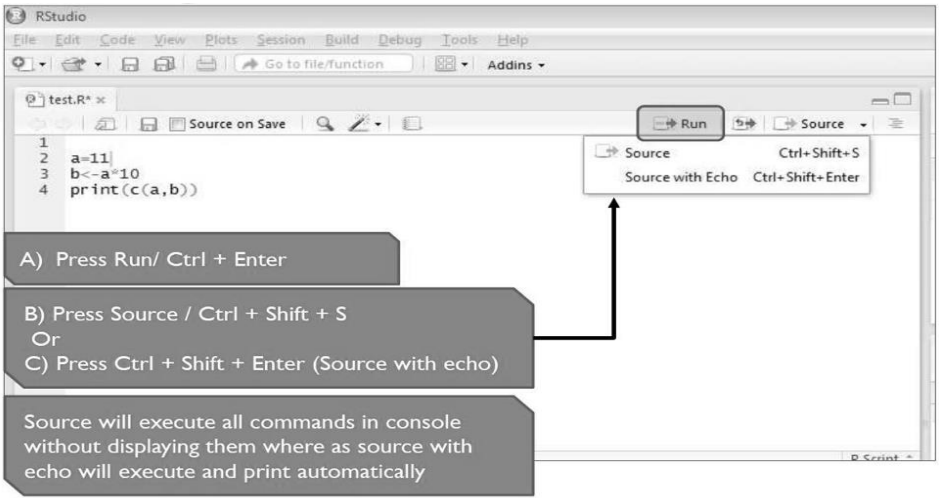


We have now learned how to open a R script, write some code inside it, and save the script file.

Executing the R file is the next step.

Execution of an R file

The commands that are available in the R file can be executed in a variety of ways.

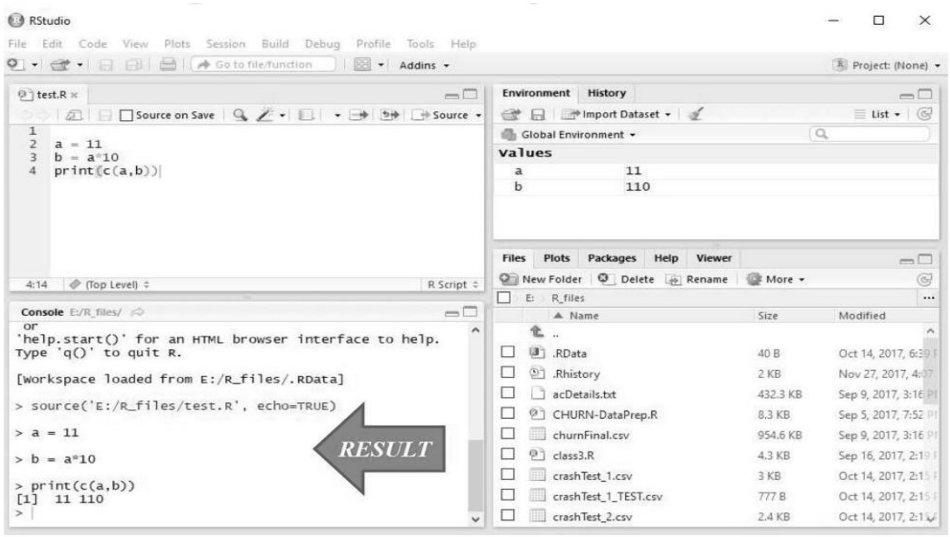


- **Using the run command:** You can use the GUI to run the "run" command by clicking the corresponding run button there, or you can use the shortcut key control + enter. The line where the cursor is located will be executed.

- **Using the command source:** You can use the source button on the GUI to run the "source" command, or you can press the shortcut key control + shift + S. This will run the entire R file and only print the results that you requested.

- **Using the source with echo command: You can use the source with echo button on the GUI or the shortcut key control + shift + enter to run this "source with echo" command. Along with the output you are printing, the commands will also be printed.**

This is an illustration of how to run a R file using the source and echo.



The output print(c(a, b)) with the values can be seen in the console along with the commands a = 11 and b = a*10 that were printed.

A is therefore 11 and b is 11 times 10, which equals 110. The output will be displayed in the console in this manner.

The environment panel also displays the values of a and b.

Run command over Source command:

- Run can be used to run the R code that has been chosen.
- To execute the entire file, use Source and Source with echo.
- The benefit of using Run is that when something does not behave as you would expect, you can troubleshoot or debug the programme.
- Using the run command has the drawback of cluttering and unnecessarily clogging the console.

8.5.3.4 Basic Syntax in R Programming

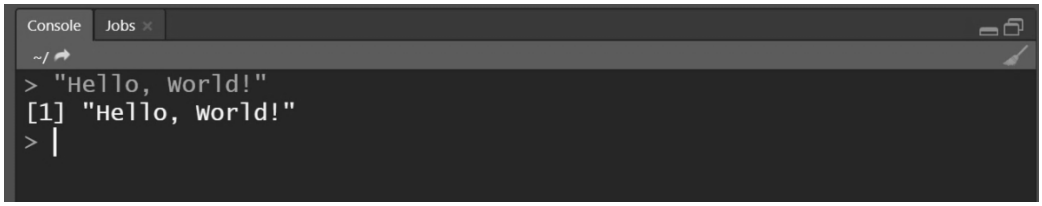
Statistical computing and data analysis are the most common uses of R, which has over 10,000 free packages available in the CRAN repository.

If you want to use R's robust features, you must understand its specific syntax, which is true of any programming language.

We'll be using RStudio assuming R is already installed on your computer, but you can also use the command prompt by entering the following command:

```
$ R
```

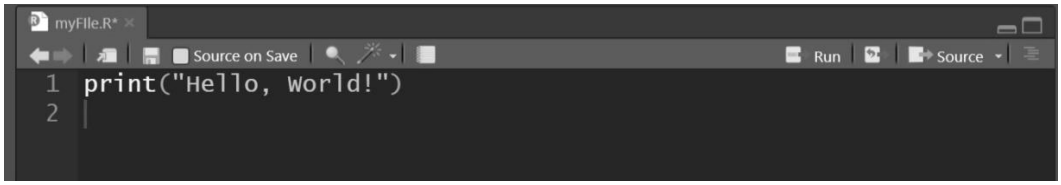
This will start the interpreter, so let's start by creating a simple Hello World programme.



On the console, we can see the words "Hello, World!" printed.

Using print(), which prints to the console, we can now accomplish the same task. Scripts, or RScripts as they are known in R, are typically where we write our code. In order to make one, write the code below in a file and save it as myFile.R. You can then run it in the console by writing:

```
Rscript myFile.R
```



Output:
[1] "Hello, World!"

Syntax of R program

Three components make up a R programme: variables, comments, and keywords. The data is stored in variables, comments are used to make the code more readable, and keywords are reserved words with a particular meaning for the compiler.

Variables in R - We wrote all of our code previously in a **print(),** but we don't have a way to address them to carry out additional operations. This problem can be solved by using **variables**, which are the names given to reserved memory locations in any programming language that can store any type of data.

The assignment can be represented in R in one of three ways:

1. **= (Simple Assignment)**
2. **<- (Leftward Assignment)**
3. **-> (Rightward Assignment)**

Example:

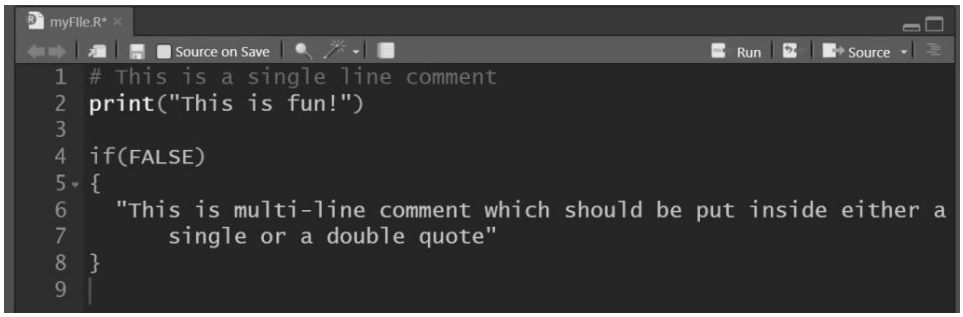
Output:
"Simple Assignment"

"Leftward Assignment!"

"Rightward Assignment"

Comments in R - Comments are a way to make your code easier to read, and since they are only for the user, the interpreter will not read them. R only supports single-line comments, but we can use multiline comments by employing a straightforward trick that is demonstrated below. Use # at the beginning of the statement to create single-line comments.

Example:



Output:
[1] "This is fun!"

Both comments were disregarded by the interpreter, as can be seen from the output above.

Keywords in R - A keyword is a word that a programme reserves because it has a specific meaning; as a result, it cannot be used as a variable name, function name, etc.

Using ***help(reserved)*** or ***?reserved***, we can view these keywords.

Reserved words in R

if	else	while	repeat	for
function	in	next	break	TRUE
FALSE	NULL	Inf	NaN	NA
NA_integer_	NA_real_	NA_complex_	NA_character_	…

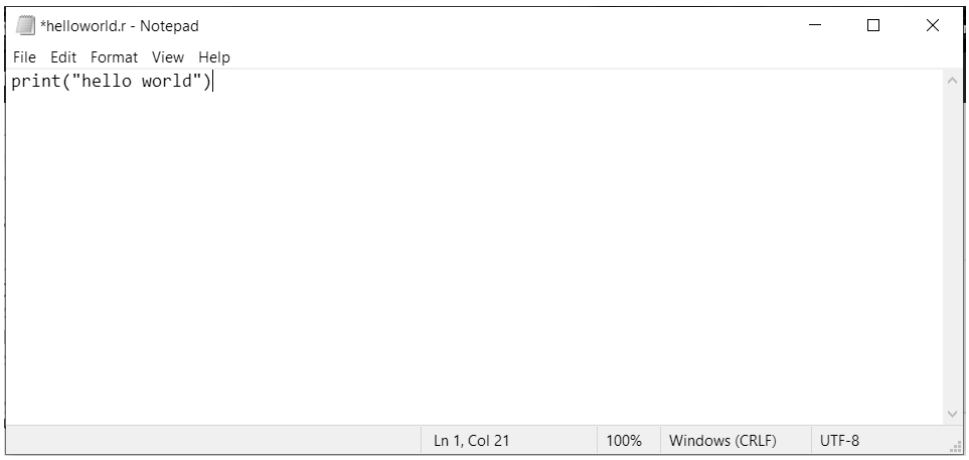
- if, else, repeat, while, function, for, in, next and break are used for control-flow statements and declaring user-defined functions.
- The ones left are used as constants like **TRUE/FALSE** are used as boolean constants.
- **NaN** defines Not a Number value and **NULL** are used to define an Undefined value.
- **Inf** is used for Infinity values.

Because R is case-sensitive, **TRUE** is not the same as **True**.

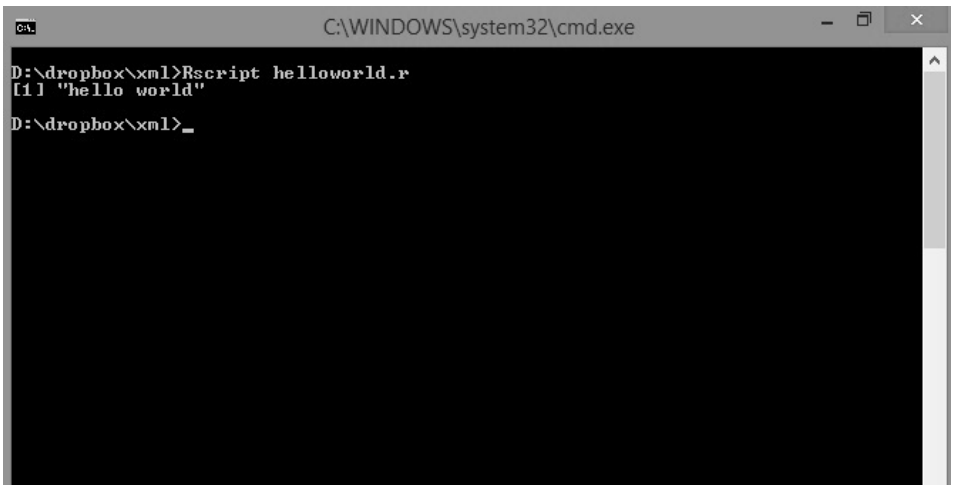
8.5.3.5 Different ways to run a R program

There are many ways to run an R program:

- **Method 1: Using command prompt or terminal** - Write your code in notepad or any text editor, save it as "helloworld.r",



Run it in command prompt or terminal using the command "Rscript helloworld.r".



- **Method 2: Using an online IDE** - There are many online IDE available. We can use that without the need of installing or downloading anything.

- **Method 3: Using IDE like Rstudio, RTVS, StatET** - You can download and install these IDE in your system and can write and run the program there. Rstudio & statET(Eclipse software)is available for Windows, Mac, and Linux. RTVS presently available only on windows.

8.6 DATA MANAGEMENT

Once you have access to your data, you should reformat it so that it is usable. Creating new variables (including recoding and renaming existing

variables), sorting and merging datasets, aggregrating data, reshaping data, and subsetting datasets (including selecting observations that meet criteria, randomly sampling observation, and dropping or keeping variables) are examples of how to do this.

R's built-in operators (arithmetic and logical) and functions (numeric, character, and statistical) are typically used for each of these tasks. Additionally, you might need to write your own functions or use control structures (if-then, for, while, switch) in your programmes. Finally, you might need to convert variables or datasets from one type to another (e.g. numeric to character or matrix to data frame).

8.7 GRAPHS

Before you actually start to analyse your data, graphs are a really helpful way to look at it. You might be perplexed as to why creating graphs is even necessary. Before attempting to interpret the more important statistics, examine your data visually and note how it appears. You can easily create very snazzy-looking graphs using R (and other packages), and you might find yourself passing out from excitement as you colour your graph. Keep in mind that the purpose of the graph is to present information, not to make yourself (or others) squeal with joy at the colour of the graph (dull, perhaps, but true).

Tufte (2001) wrote an excellent book on how data should be presented. He emphasised that graphs should include the following among other things:

- Show the data.
- Induce the reader to think about the data being presented (rather than some other aspect of the graph, like how pink it is).
- Avoid distorting the data.
- Present many numbers with minimum ink.
- Make large data sets (assuming you have one) coherent.
- Encourage the reader to compare different pieces of data.
- Reveal data.

The R language's preferred feature for creating different types of graphs and charts for visualisations is its support for graphs. In order to create the graphs using the input data set for data analytics, the R language supports a rich set of packages and functionalities. The scatter plot, box plot, line, pie, histogram, and bar graphs are the most frequently used graphs in the R programming language. For exploratory data analysis, R graphs support both two-dimensional and three-dimensional plots. To create graphs in the R language, functions like plot(), barplot(), and pie() are used. Advanced graph functionalities are supported by R packages such as ggplot2.

Because of R's powerful graphic capabilities, data analysts frequently use it.

Data Visualization in R Programming Language

There are several well-liked tools for data visualisation, including Tableau, Plotly, R, Google Charts, Infogram, and Kibana. The features, functions, and use cases of the various data visualisation platforms differ. They also call for a different set of skills.

R is a language intended for scientific research, graphical data analysis, and statistical computing. It is typically chosen for data visualisation because it provides flexibility and requires little coding thanks to its packages.

Consider the following *airquality* data set for visualization in R:

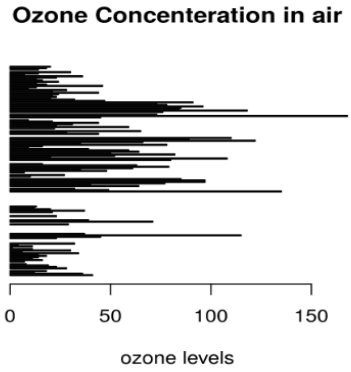
Ozone	Solar R.	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

Bar Plot - There are two different kinds of bar plots: horizontal and vertical. In both, data points are represented by horizontal or vertical bars of varying lengths that are proportional to the value of the data item. For plotting continuous and categorical variables, they are typically utilised. We can obtain horizontal and vertical bar plots, respectively, by setting the horiz parameter to true and false.

Example 1:

```
# Horizontal Bar Plot for
# Ozone concentration in air
barplot (airquality$Ozone,
        main = 'Ozone Concenteration in air',
        xlab = 'ozone levels', horiz = TRUE)
```

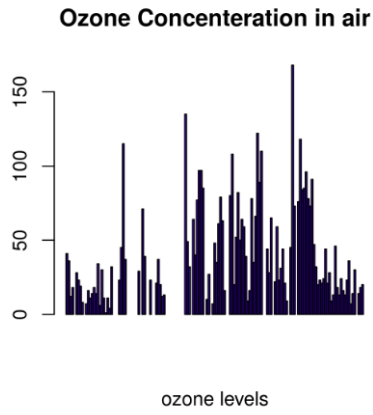
Output:



Example 2:

```
# Vertical Bar Plot for
# Ozone concentration in air
barplot(airquality$Ozone, main = 'Ozone Concenteration in air',
                xlab = 'ozone levels', col ='blue', horiz = FALSE)
```

Output:



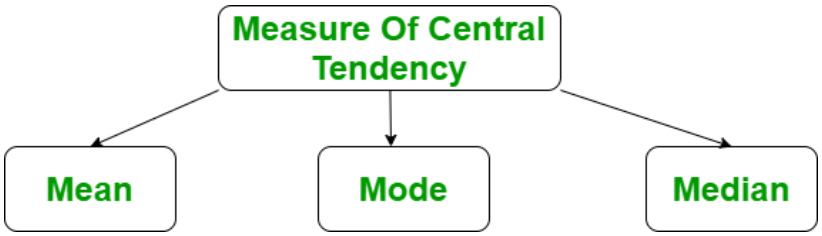
The following situations call for the use of bar plots:

- to conduct a study in which the various data categories in the data set are compared.
- to examine how a variable has changed over the course of several months or years.

Similarly, we can create histogram, box plot, scatter plot, heat map, map visualization, 3D graphs in R.

8.8 DESCRIPTIVE, BASIC AND MULTIVARIATE STATISTICS IN R

Mean, Median and Mode in R Programming- The measure of central tendency in R Language represents the whole set of data by a single value. It gives us the location of central points. There are three main measures of central tendency:



- ***Mean in R Programming Language*** *- It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.*

$$\text{Mean } (\bar{x}) = \frac{\sum x}{n}$$

Where, **n** = number of terms

Example:

```
# R program to illustrate
# Descriptive Analysis

# Import the data using read.csv()
myData =
read.csv("CardioGoodFitness.csv",
                          stringsAsFactors=F)

# Compute the mean value
mean = mean(myData$Age)
print(mean)
```

Output:

```
[1] 28.78889
```

- ***Median in R Programming Language*** *- It is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd then the center element is median and if it is even then the median would be the average of two central elements.*

Odd	Even
$\frac{n+1}{2}$	$\frac{n}{2}, \frac{n}{2}+1$

Where **n** = number of terms

Syntax: *median(x, na.rm = False)*

Where, *X is a vector and na.rm is used to remove missing value*

Example:

```
# R program to illustrate
# Descriptive Analysis

# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                          stringsAsFactors=F)

# Compute the median value
median = median(myData$Age)
print(median)
```

Output:

```
[1] 26
```

- **Mode in R Programming Language -** It is the value that has the highest frequency in the given data set. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency. There is no inbuilt function for finding mode in R, so we can create our own function for finding the mode or we can use the package called modeest.

  Creating user-defined function for finding Mode - R doesn't have a built-in function for locating the mode. So let's develop a user-defined function that returns the data's mode. For this, we'll use the table() method, which generates a table-like categorical representation of the variable names and frequency of the data. The Age column will be sorted in descending order, and the first value from the sorted values will be returned.

  Example: Finding mode by sorting the column of the dataframe

```
# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                        stringsAsFactors=F)
mode = function(){
          return(sort(-table(myData$Age))[1])
}
mode()
```

Output:

```
25: -25
```

Using Modeest Package - We can use the modeest package of the R. This package provides methods to find the mode of the univariate data and the mode of the usual probability distribution.

Example:

```
# R program to illustrate
# Descriptive Analysis

# Import the library
library(modeest)

# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                  stringsAsFactors=F)

# Compute the mode value
mode = mfv(myData$Age)
print(mode)
```

Output:

```
[1] 25
```

Standard Deviation in R Programming Language - *Standard Deviation is the square root of variance. It is a measure of the extent to which data varies from the mean. The mathematical formula for calculating standard deviation is as follows,*

$$StandardDeviation = \sqrt{variance}$$

Example:

Standard Deviation for the above data,

$$StandardDeviation = \sqrt{4} = 2$$

Computing Standard Deviation in R

One can calculate the standard deviation by using **sd()** function in R.

Syntax: *sd(x)*
Parameters:
x: *numeric vector*

Example:-

```
# R program to get
# standard deviation of a list

# Taking a list of elements
list = c(2, 4, 4, 4, 5, 5, 7, 9)

# Calculating standard
# deviation using sd()
print(sd(list))
```

Output:

```
[1] 2.13809
```

R functions for computing descriptive analysis:

Analysis	R Function
Mean	mean()
Median	median()
Mode	mfv() [modeest]
Range of values (minimum and maximum)	range()
Minimum	min()
Maximum	maximum()
Variance	var()
Standrad Deviation	sd()
Sample quantiles	quantile()
Interquartile range	IQR()
Generic function	summary()

Computing ANOVA in R

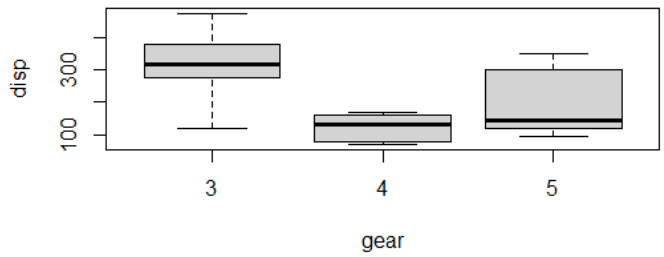
To get started with ANOVA, we need to install and load the **dplyr** package.

One way ANOVA test is performed using mtcars dataset which comes preinstalled with dplyr package between disp attribute, a continuous attribute and gear attribute, a categorical attribute.

Performing One Way ANOVA test in R language

```
# Installing the package
install.packages("dplyr")

# Loading the package
library(dplyr)

# Variance in mean within group and between group
boxplot(mtcars$disp~factor(mtcars$gear),
                        xlab = "gear", ylab = "disp")

# Step 1: Setup Null Hypothesis and Alternate Hypothesis
# H0 = mu = mu01 = mu02(There is no difference
# between average displacement for different gear)
# H1 = Not all means are equal

# Step 2: Calculate test statistics using aov function
mtcars_aov <- aov(mtcars$disp~factor(mtcars$gear))
summary(mtcars_aov)

# Step 3: Calculate F-Critical Value
# For 0.05 Significant value, critical value = alpha = 0.05

# Step 4: Compare test statistics with F-Critical value
# and conclude test p < alpha, Reject Null Hypothesis
```

Output:



The box plot shows the mean values of gear with respect of displacement. Hear categorical variable is gear on which factor function is used and continuous variable is disp.

```
                Df Sum Sq Mean Sq F value   Pr(>F)
factor(mtcars$gear)  2 280221  140110   20.73 2.56e-06 ***
Residuals           29 195964    6757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

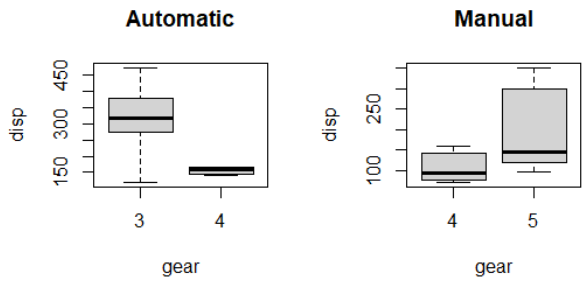
The summary given above shows that the gear attribute is very significant to displacement (Three stars denoting it). Also, the P value is less than 0.05, so proves that gear is significant to displacement i.e related to each other and hence, we reject the Null Hypothesis.

Performing Two Way ANOVA test in R language

Two-way ANOVA test is performed using mtcars dataset which comes preinstalled with dplyr package between disp attribute, a continuous attribute and gear attribute, a categorical attribute, am attribute, a categorical attribute.

```
# Installing the package
install.packages("dplyr")

# Loading the package
library(dplyr)

# Variance in mean within group and between group
boxplot(mtcars$disp~factor(mtcars$gear),
                        xlab = "gear", ylab = "disp")

# Step 1: Setup Null Hypothesis and Alternate Hypothesis
# H0 = mu = mu01 = mu02(There is no difference
# between average displacement for different gear)
# H1 = Not all means are equal

# Step 2: Calculate test statistics using aov function
mtcars_aov <- aov(mtcars$disp~factor(mtcars$gear))
summary(mtcars_aov)

# Step 3: Calculate F-Critical Value
# For 0.05 Significant value, critical value = alpha = 0.05

# Step 4: Compare test statistics with F-Critical value
# and conclude test p < alpha, Reject Null Hypothesis
```

Output:



The box plot given above shows the mean values of gear with respect to displacement. Here, categorical variables are gear and am on which factor function is used and continuous variable is disp.

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
factor(mtcars$gear)  2 280221  140110  20.695 3.03e-06 ***
factor(mtcars$am)    1   6399    6399   0.945    0.339
Residuals           28 189565    6770
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary above demonstrates that the gear attribute is very significant to displacement (Three stars denoting it) and am attribute is not much significant to displacement. P-value of gear is less than 0.05, so it proves that gear is significant to displacement i.e related to each other. P-value of am is greater than 0.05, am is not significant to displacement i.e not related to each other.

8.9 R GUI, OTHER SOFTWARE

R is a programme that runs from the command line. The user types commands into the prompt (> by default), and each command is executed one by one. However, in order to use R, you will need a Graphical User Interface (GUI). There have been several attempts to develop a more graphical user interface, ranging from code editors that work with R to fully developed GUIs that display menus and dialogue boxes to the user. Although many data scientists favour working in the command line, beginners should pick a specific GUI. R has a wide variety of free graphical user interfaces. To save you time searching for them, we have listed a few of them.

Free Graphical User Interfaces for R: - R is the statistical problem-solving software of choice for most developers. You'll also want to use it to address pertinent problems. Any of the GUIs that are offered on the market can be chosen. The fact that the majority of them are free is the most satisfying aspect. Let's review the top 10 free R graphical user interfaces.

1) **RStudio:-** The most well-known IDE and GUI for R is called RStudio. It has a significant advantage over its rivals. Two formats are supported by the free GUI. Most people use the desktop version of Rstudio. The RStudio server is located on a remote server, and you can access it using your browser. RStudio is open-source software because it is licenced under the GNU Affero General Public License.

2) **Rattle:-** A well-liked free graphical user interface for R is called Rattle. The GUI is used effectively for extensive data mining. Anyone with access to GitHub can download the source code for Rattle and contribute new code to the development. It provides robust data mining features by showcasing R software's potential through a GUI. It is currently utilised by numerous governmental and non-governmental organisations throughout the world for statistical and data mining purposes.

3) **StatET for R:-** Eclipse has been developed into StatET. It offers an Eclipse-based integrated development environment and generates modules for R and Java integration. Here, you can find everything a R IDE should have. The GUI offers a variety of sophisticated tools for R coding and package creation, including an integrated R console and the "R" help system.

4) **RKWard:-** In comparison to other free graphical user interfaces, RKWard is simpler. The GUI's main goal is to develop into a user-

friendly, transparent frontend to R. RKWard aims to combine the R-potential language's with the user-friendliness of consumer statistical software. For the KDE desktop environment, it was created. The GUI, however, can function in almost any setting.

5) **JGR:-** JGR, which is pronounced Jaguar, is a uniform and standardised GUI for R. It is free and open-source because it is licenced under the GNU General Public License. The GUI underwent its initial introduction in 2004, and development is still ongoing. It has a limited adaptive R terminal that can be used to replace the standard R GUI in more ways. JGR is well-known among data scientists because of its accommodating R-console.

6) **R Commander:-** A free graphical user interface for R is called R Commander. The software was developed by Prof. John Fox in order to facilitate the teaching of statistics courses and remove barriers associated with software sophistication from the process of learning statistics. You can navigate statistical data analysis using drop-down menus. Beginners will particularly benefit from R Commander because it displays the corresponding code for each data execution.

7) **Deducer:-** A free and open-source GUI for R is called Deducer. It was initially developed as an approachable alternative to programmes like Minitab, SPSS, and JMP. Every operating system supports the GUI, but there is no server version. The 2008 release of Deducer was met with immediate and widespread acclaim. However, installing it is a difficult process. It is suitable for both experts and beginners.

8) **JASP:-** An open-source GUI developed by the University of Amsterdam is called JASP. Users of SPSS should find it easy to use and intuitive. JASP's emphasis on Bayesian analysis is one of its most significant advantages. If you prefer that, JASP could be the one for you. Additionally, it features a strong Machine Learning Module. Compared to the other GUIs discussed here, JASP is a little unique. This is due to the fact that you cannot run your own R code in it, nor can it display the R code that it generates. However, it can carry out every other function you would expect from a R GUI.

9) **Tinn-R:-** Another free graphical user interface available is Tinn-R. Additionally, the GUI is an ASCII/UNICODE generic. It is a simple but effective substitute for the standard R GUI editor. Tinn-R wants to promote education. Additionally, it strives to make using the R environment as simple as possible. By using this GUI, novice users can undoubtedly improve their R learning.

10) **BlueSky Statistics:-** This GUI was created by former SPSS employees and shares many characteristics with SPSS. Before 2018, you had to purchase it. However, it is now open-source. BlueSky makes it impossible to determine whether R is actively engaged. The R code editor can only be accessed by selecting the "Sytax" button. The tidyverse style, which is popular but controversial, is used in BlueSky. Currently, Windows is the only operating system that

supports the GUI. Versions of BlueSky, however, are being created for additional platforms.

Perhaps the most reliable full-featured GUIs that work with Windows, Linux, and MacOS are RStudio and R Commander. These two tools can make using R much simpler.

8.10 SUMMARY

As the saying goes, "Data is the new world currency". However, simply collecting data will not yield a profit. Data utilisation is essential. The appropriate data must be used in the appropriate context. That's what data analytics and statistics are all about. For analysing data, R is a suitable programming language. But to do this, you also need a platform. Therefore, we have also provided a list of 10 free R graphical user interfaces.

8.11 QUESTIONS

1. What exactly is structural equation modelling (SEM)?
2. What are the underlying assumptions of SEM?
3. Describe the Steps for conducting SEM.
4. What are the limitations to structural equation modelling (SEM)?
5. Describe in short R syntax.
6. What are some of the pros and cons of R?
7. What are the steps to download and install R?
8. How can you create and execute R File in R Studio?
9. Illustrate how will you write scripts in R file.
10. What are Variables and Keywords in R?
11. State some ways to run a R program
12. Illustrate by giving example how will you draw a bar graph in R.
13. How would you perform mean / median /mode/ standard deviation / ANOVA in R?
14. Write a note on R GUI and other software.

8.12 REFERENCES

1. Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. NY: Sage.

2. Heumann, C & Shalabh, M. S. (2016). *Introduction to Statistics and Data Analysis - With Exercises, Solutions and Applications in R.* Springer. https://10.1007/978-3-319-46162-5

3. Hasan, M. (2022, November 24th). 10 Best Free Programming Graphical User Interfaces for R. *Ubuntu*. https://www.ubuntupit.com/best-free-graphical-user-interfaces-for-r/

4. Kabacoff, R.I (2017). Graphic User Interfaces. *Quick – R by Datacamp*. https://www.statmethods.net/interface/guis.html

5. Tabachnick, B.G & Fidell, L.S.(2007). Structural Equation Modeling. *Using Multivariate Statistics (5th Ed.).* Pearson. 676-780.

❄❄❄❄❄❄❄