Parametric vs non parametric

Parametric test

Parametric tests makes assumptions about the population parameters such as the data is normally distributed, variables involved must be measured in interval or ratio scale etc.

Non Parametric test

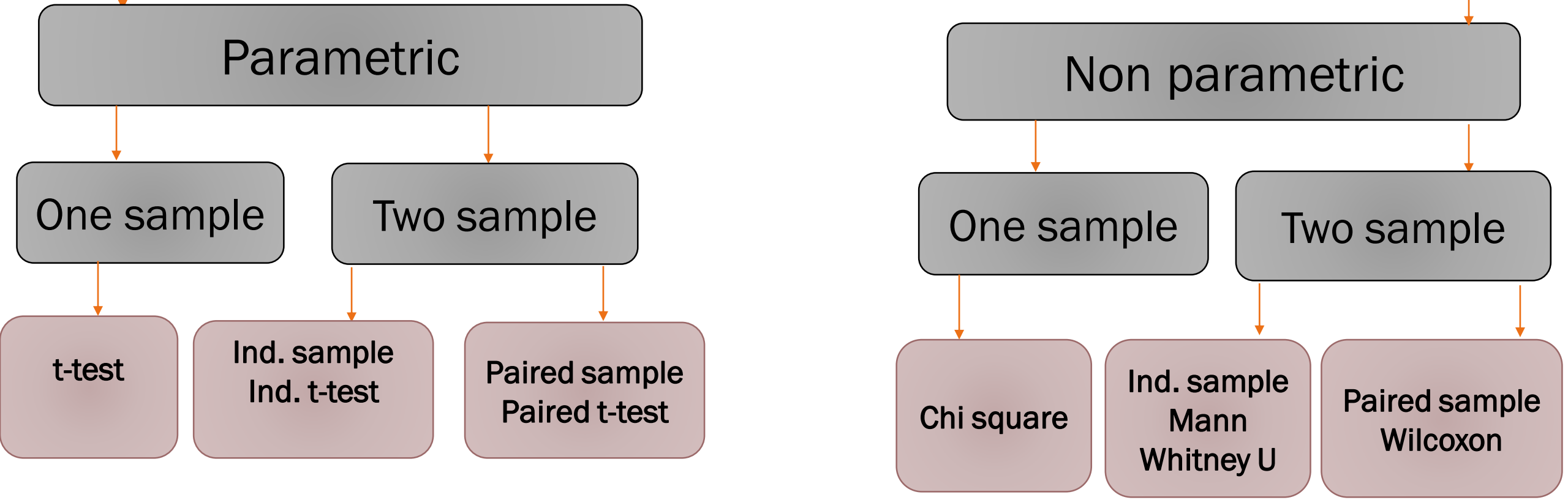
Non parametric tests are distribution free tests. It is based on the rank of the observations. There is no assumptions about the population parameters.

Parametric vs non parametric tests

PARAMETRIC TEST	NON PARAMETRIC TEST
Absolute value based test	Rank based test
Sample size > 30	Sample size < 30
Make assumptions	No assumptions about the population are made
Information about the population are known	No information about the population
Data should be normally distributed	Distribution free test
Applicable for interval and ratio scale	Applicable for nominal and ordinal scale
powerful	Less powerful

	Parametric test	Non parametric
Benefits	Can draw more conclusions	Simplicity, less affected by outliers
Central tendency	Mean	Median
correlation	Pearson	Spearman's Rho, Kendall's Tau
2 group test	t-test	Mann Whitney U
More than 2 group	one way ANOVA	Kruskal Wallis ANOVA
Paired	Paired t-test	Wilcoxon

Parametric vs non parametric



Assumptions of Parametric test

- The populations are normally distributed
- The selected population is representative of general population
- Data is in interval or ratio scale
- Observations must be independent
- Population must have same variance

Assumptions of Non Parametric test

- Data doesn’t follow any specific distribution.
- Data measured on any scale
- No assumptions about the populations are made
- Sample size is quite small
- Data can be expressed in the form of ranks.
- The nature of the population from which sample are drawn is not known to be normal.
- The variables are expressed in nominal form.

ADVANTAGES AND DISADVANTAGES OF NON PARAMETRIC STATISTICS

Advantages

- Can be used in small data
- Makes fewer assumptions about the data which are more relevant .
- More appropriate for research investigation.
- Are measured in nominal scale
- Samples made up of observations from several different populations at times can be handled.
- Much easier to learn and apply.
- More direct interpretation.

Disadvantages

- If all assumptions of parametric test are fulfilled then using non parametric may be wasteful.
- Less powerful
- Tables necessary to implement non parametric test are widely scattered and appear in different formats.

SCALE OF MEASUREMENT

Nominal :

categories which do not have natural order.
E.g. . Gender, eye color, types of building etc.,

Ordinal :

Categories which have a natural order but are not numerical.
E.g. Likert scale(strongly agree, agree , undecided, disagree, strongly disagree)

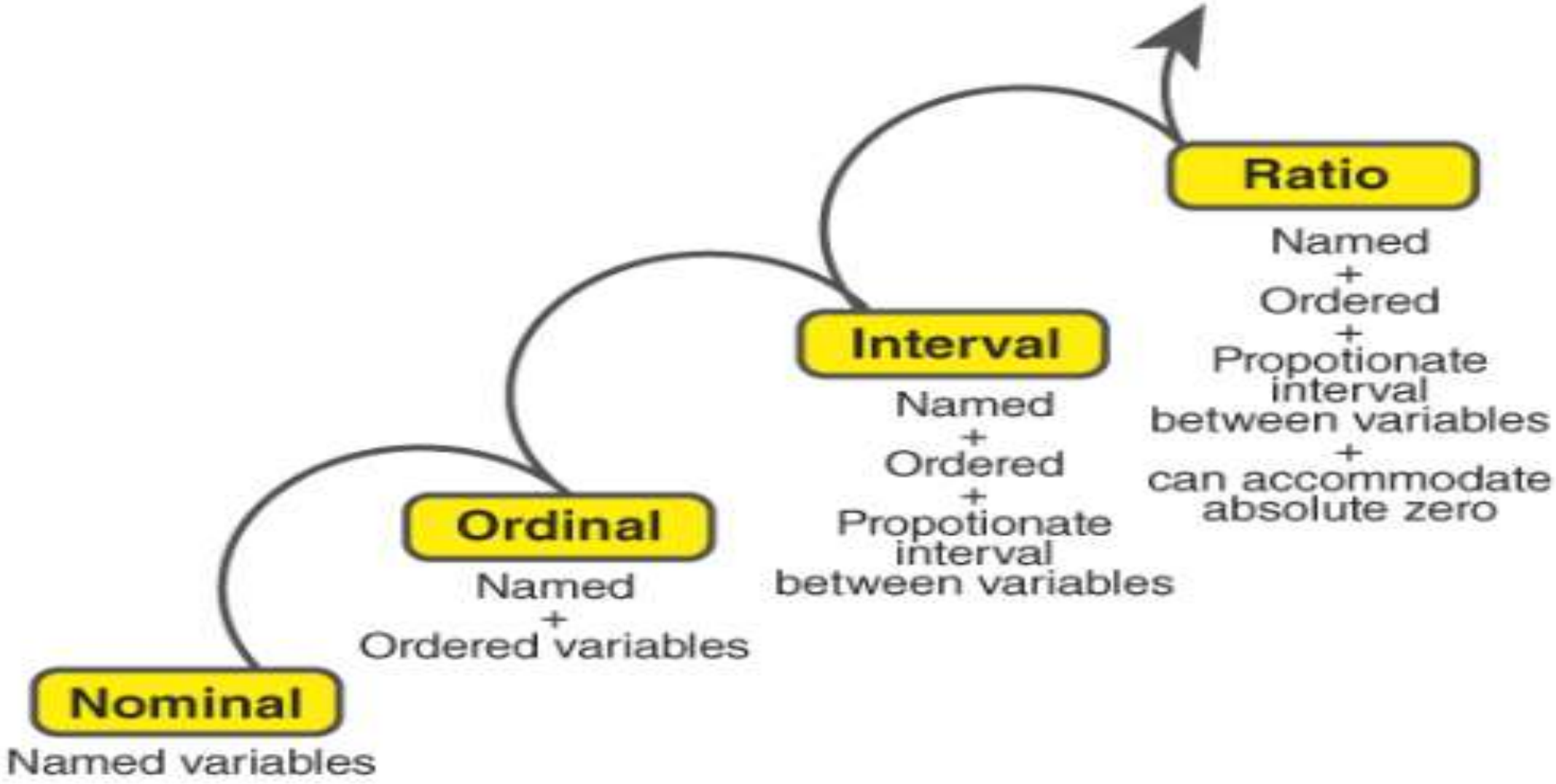
Interval :

Numerical data ordered against a constant scale.
E.g., date, temperature, length, weight, frequency etc.,

Ratio :

Has an absolute zero and permits comparisons such as being twice as high or one-half as much etc.,
e.g., age, sales figure, income, years of experience etc.,

LEVELS OF MEASUREMENT



Scale of measurement

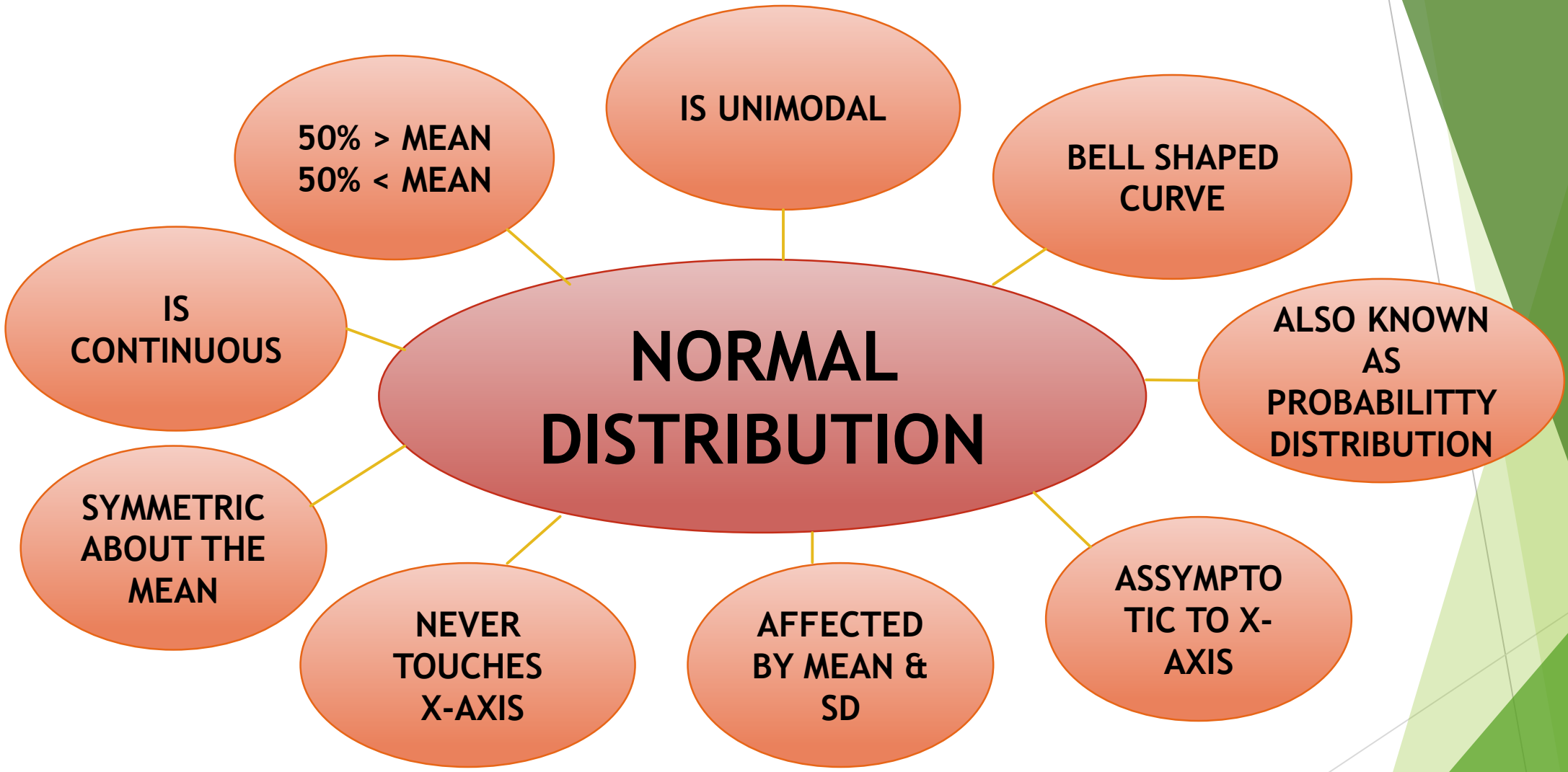
Nominal & ordinal

- Non parametric
- Qualitative data
- Discrete variable
- Categorically scaled

Interval & ratio

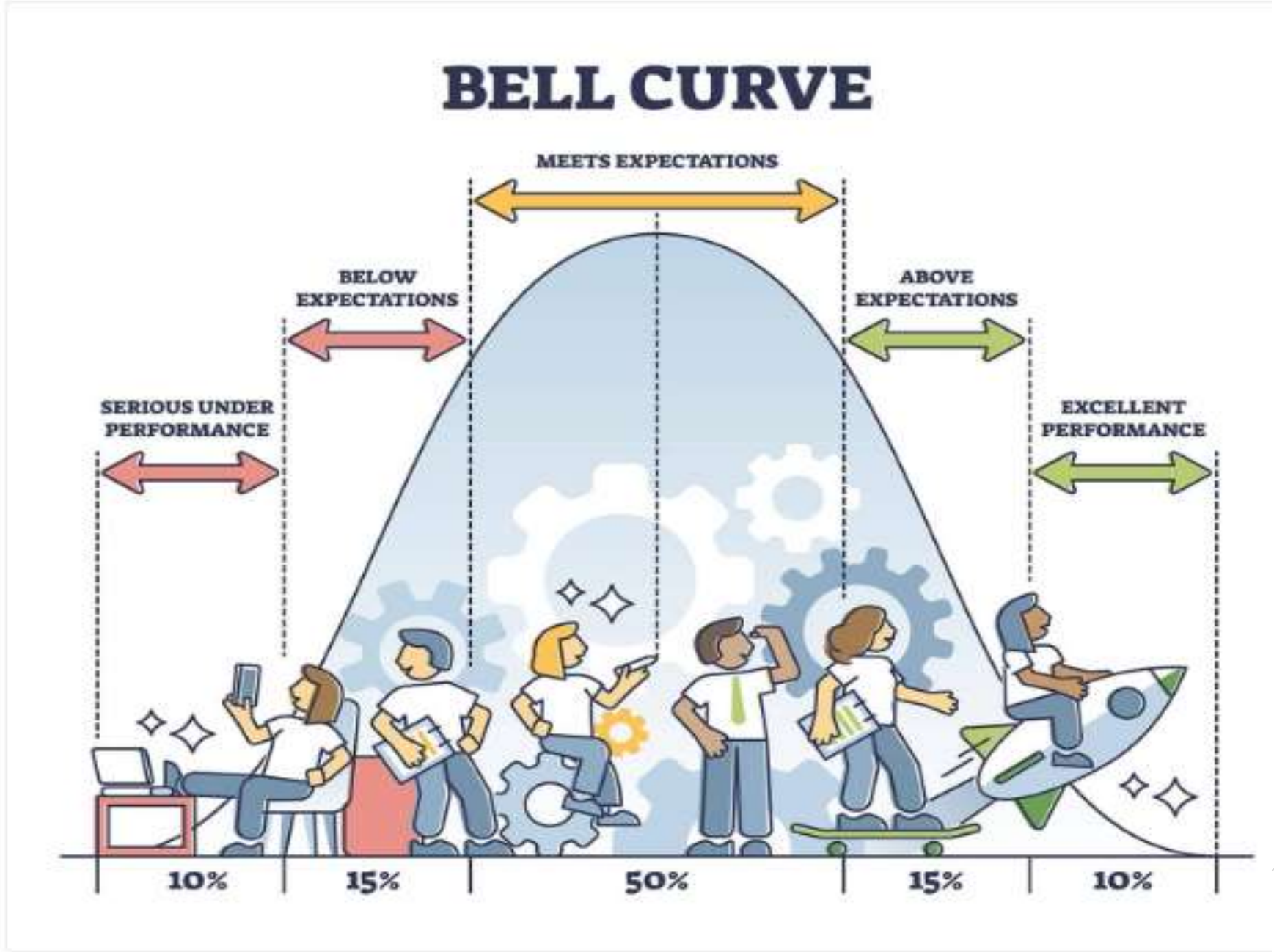
- Parametric
- Quantitative data
- Continuous variable
- Continuously scaled

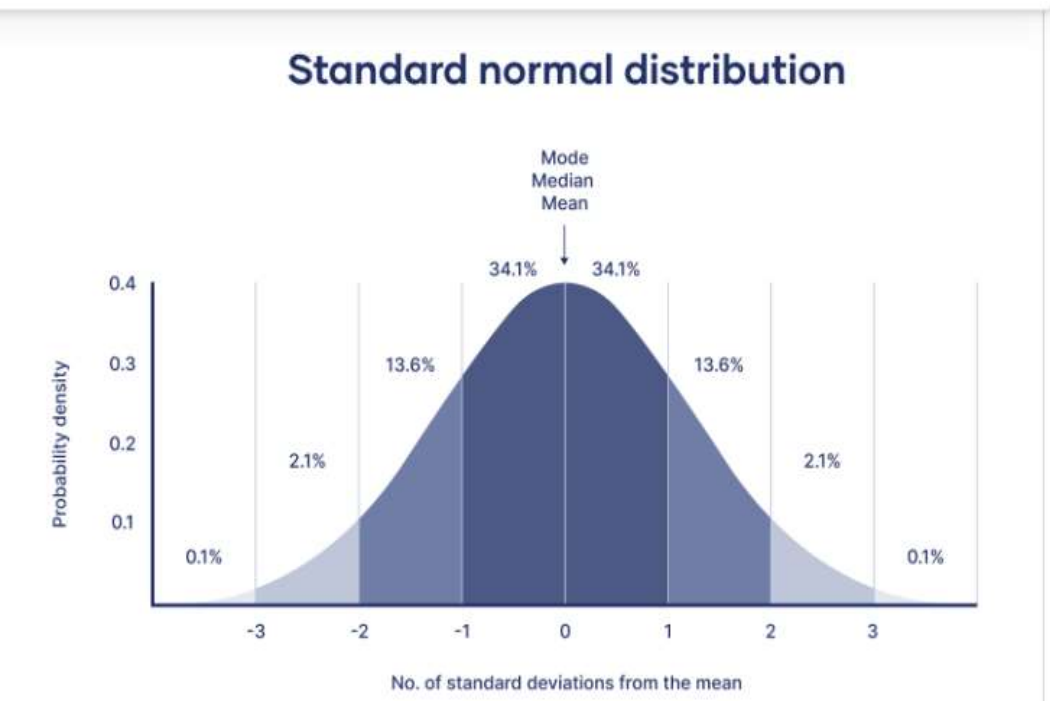
NORMAL PROBABILITY CURVE & DIVERGENCE FROM NORMALITY


NORMAL DISTRIBUTION
IS UNIMODAL
50% > MEAN
50% < MEAN
BELL SHAPED CURVE
IS CONTINUOUS
ALSO KNOWN AS PROBABILITTY DISTRIBUTION
SYMMETRIC ABOUT THE MEAN
NEVER TOUCHES X-AXIS
AFFECTED BY MEAN & SD
ASSYMPTOTIC TO X-AXIS

What Is a Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The normal distribution appears as a "bell curve" when graphed.


BELL CURVE
MEETS EXPECTATIONS
BELOW EXPECTATIONS
ABOVE EXPECTATIONS
SERIOUS UNDER PERFORMANCE
EXCELLENT PERFORMANCE
10%
15%
50%
15%
10%



The term Normal Distribution is defined as a function that represents the distribution of many random variables as a symmetrical bell shaped graph. It is the most probable continuous distribution.

The total area the normal curve logically represents the sum of all probabilities for a random variable. Hence, the area under the normal curve is one. Also the standard normal curve represents a normal curve with mean 0 and standard deviation 1.

Properties of NPC – (given in the beginning)

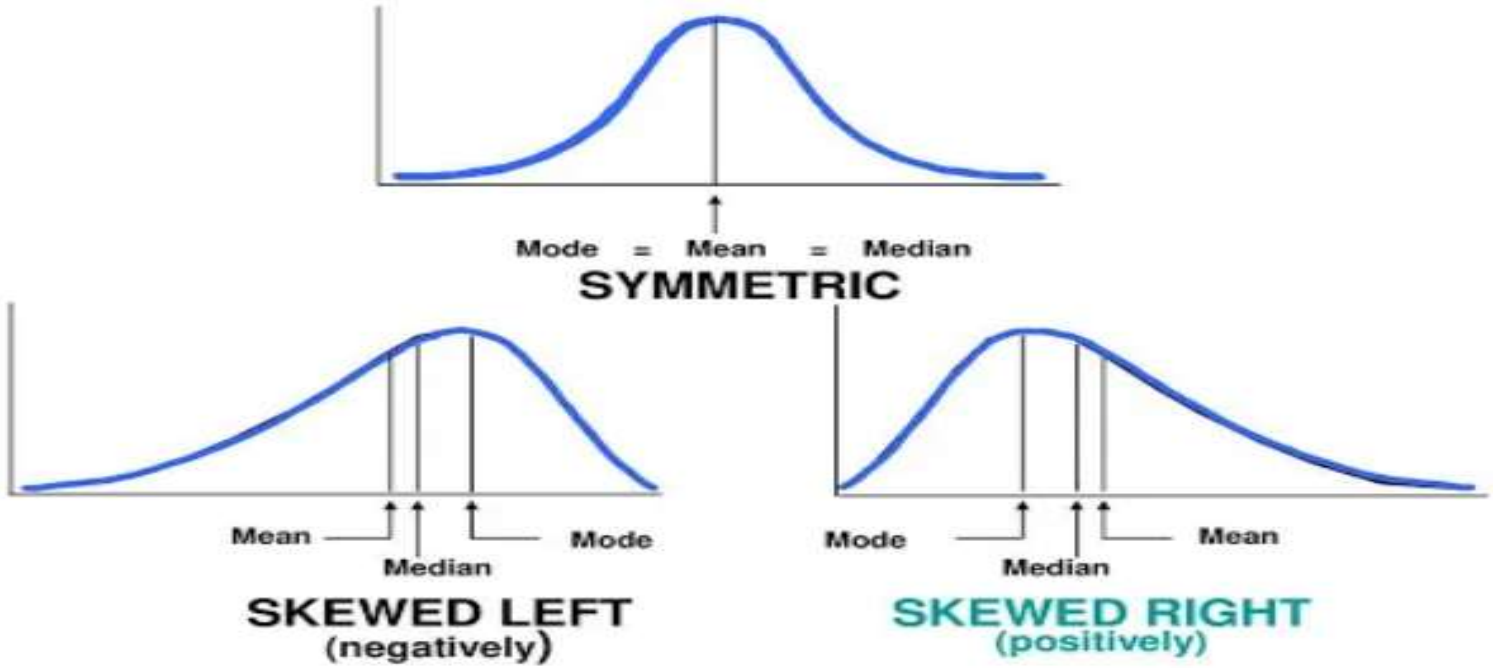
- Normal probability curve is symmetrical
- The normal curve is Unimodal
- Normal Curve is asymptotic to X-axis
- Mean- Median- Mode
- Area under curve is 1
- The maximum Ordinator occurs at the center
- Normal curve is a mathematical model in behavioral science specially in mental measurement.

DIVERGENCE IN NORMALITY

Statistically two numerical measures of shape - skewness & kurtosis - can be used to test for normality. If either of these values is not close to zero, then the data set is not normally distributed.

SKEWNESS

A distribution is normal when the mean, median and mode coincide together and there is a perfect balance between the right & left halves of the figure. But when the mean, median and mode falls at different points a shift in center of gravity occur to one side, it is said to be skewed.

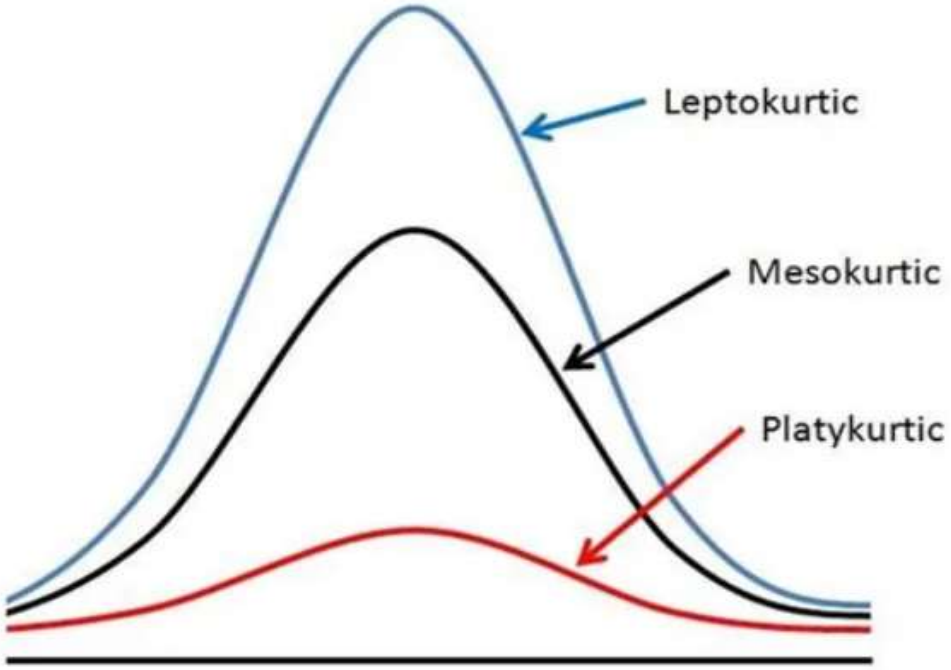


Skewness is the degree of asymmetry of the distribution .

Skewness and variability are usually related , the more the skewness the greater the variability.

KURTOSIS

Normal Probability Curve is moderately peaked . If any frequency curve is more peaked or flatter than the Normal Probability Curve , we can say it diverges from Normality, kurtosis measure such divergence.



Lepto Kurtic : If a distribution is more peaked than Normal Distribution it is said to be leptokurtic . It implies a thin distribution.

Meso Kurtic : Normal Curve

Platy Kurtic : if a distribution is flatter than the normal distribution it is known as platykurtic distribution.

Causes of divergence

Selection of the sample : If the sample size is small or the sample is biased one, skewness is possible in the distribution of scores obtained.

Unsuitable or poorly made test : If the measuring tool is poorly made, the asymmetry is likely to occur.

The trait being measured is Non-Normal : A real lack of normality in the trait that is measured , produces Skewness & Kurtosis.

Errors in the construction & Administration of the tests : A poorly constructed test causes asymmetry in the normal distribution . Unclear instructions, errors in scoring , lack of motivation etc., may cause skewness in the distribution.

MPC 06 : STATISTICS IN PSYCHOLOGY

INTRODUCTION

FOCUS POINTS IN NUMERICALS

- **ONE WAY ANOVA**
- **CHI SQUARE TEST**
- **MANN WHITNEY U TEST**
- **KENDALLS TAU CORRELATION**
- **PEARSON PRODUCT MOMENT CORRELATION**
- **SPEARMANS RANK ORDER CORRELATION**
- **KRUSKAL WALLIS ANOVA**
- **REGRESSION**
- **T TEST**
- **WILCOXON MATCHED PAIR SIGNED RANK TEST**

FOCUS POINT THEORY

1. PARAMETRIC V/S NON PARAMETRIC STATISTICS
2. DESCRIPTIVE V/S INFERENTIAL STATISTICS.
3. NORMAL PROBABILITY CURVE & DIVERGENCE FROM NORMALITY
4. HYPOTHESIS TESTING
5. TYPES OF CORRELATIONS AND DEFINITIONS
6. LEVEL OF SIGNIFICANCE
7. SCALE OF MESUREMENTS
8. ERRORS – TYPE I & TYPE II.
9. STANDARD ERROR
10. TABULATION & DIAGRAMATIC REPRESENTATION OF DATA
11. MEASURE OF CENTRAL TENDENCY WITH EXAMPLES

MANN WHITNEY U TEST

- Rank based non- parametric test
- Used to determine if there are any differences between two groups
- Smallest number get rank 1 & largest gets the nth rank
- Rank the data as a whole.
- Smaller among U & U' is taken as MANN WHITNEY U .

$$u = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_x$$

$$u' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum Ry$$

n_1- total number of terms in group 1

n_2- total number of terms in group 2

R_x -rank of sample 1

Ry- rank of sample 2

PROCESS

HYPOTHESIS

- Ho – **null hypothesis** – the sum of ranking in the two groups doe not differ.
- H_A – **Alternative hypothesis** – there is a difference in the sum of ranking in the groups

INTERPRETATION

U calculated > U critical

Accept Null Hypothesis (Ho)

U calculated < U critical

Reject Null Hypothesis (Accept H_A)

TABULATION

Sl no	X	Rx	Y	Ry
1				
2				
.				
.				
.				
n				
		$= \sum Rx$		$= \sum Ry$

VERIFICATION

$n_1 n_2$ = U + U'

Compute Mann Whitney U for the following data

Group A – 21 30 26 32 41 35 15

Group B - 12 16 17 19 20 24 31

PRACTICE WORK

Q . Compute Mann Whitney U for the following data

Rx	-	7	10	9	12	14	13	2
Group A	–	21	30	26	32	41	35	15
Group B	-	12	16	17	19	20	24	31
Ry	-	1	3	4	5	6	8	11

Ans – (rank the data as a whole)

HYPOTHESIS

- Ho – **null hypothesis** – the sum of ranking in the two groups doe not differ.
- H_A – **Alternative hypothesis** – there is a difference in the sum of ranking in the groups.

TABULATION

Sl no	X	Rx	Y	Ry
1	21	7	12	1
2	30	10	16	3
3	26	9	17	4
4	32	12	19	5
5	41	14	20	6
6	35	13	24	8
7	15	2	31	11
		$= \sum Rx$ =67		$= \sum Ry$ =38

CALCULATION

$$u = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_x$$

$$u' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum Ry$$

U = (7x7) + $\frac{(7+8)}{2}$ - 67

= 49 + 28 – 67

U = 10

U′ = (7x7) + $\frac{(7+8)}{2}$ - 38

= 49 + 28 – 38

U′ = 39

Smaller value among U & U′ is taken as Mann Whitney U

Mann Whitney U = 10

verification : $n_1 n_2$ = U + U′

7x7 = 10+39

49 = 49

LHS = RHS

Hope y'all Understood

How about a practice test ?

Compute Mann Whitney U test for the following data

Data A – 37 62 71 65 66 45

Data B – 42 61 70 63 72 47

HAPPY LEARNING

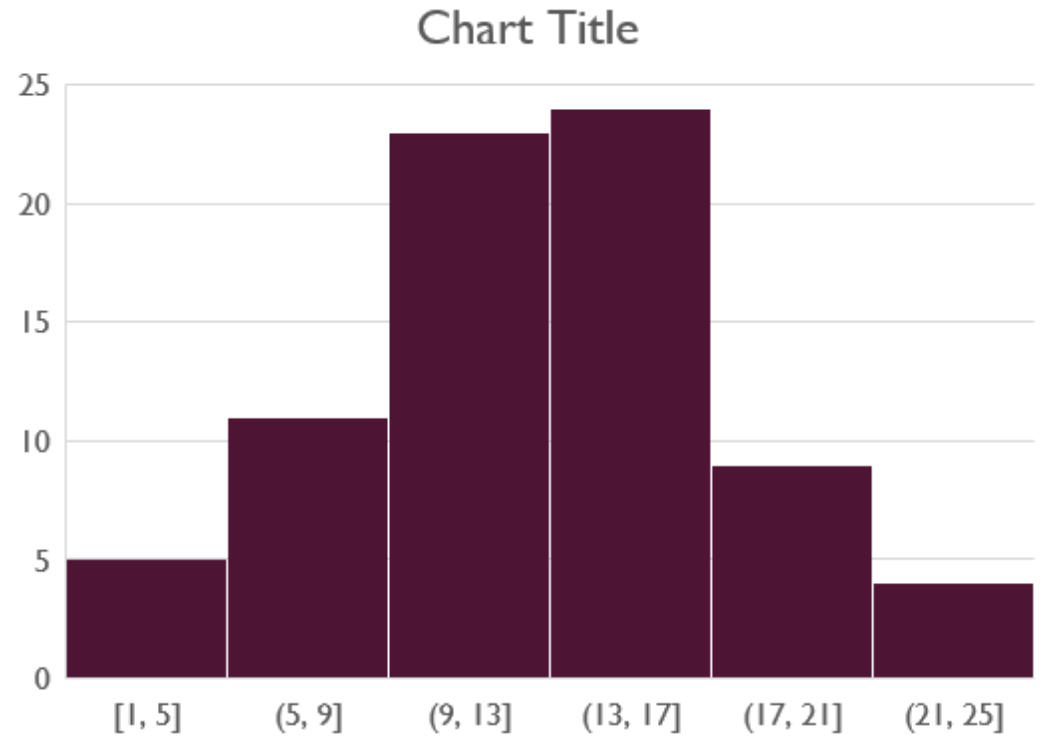
GRAPHICAL & DIAGRAMATIC PRESENTATION OF DATA

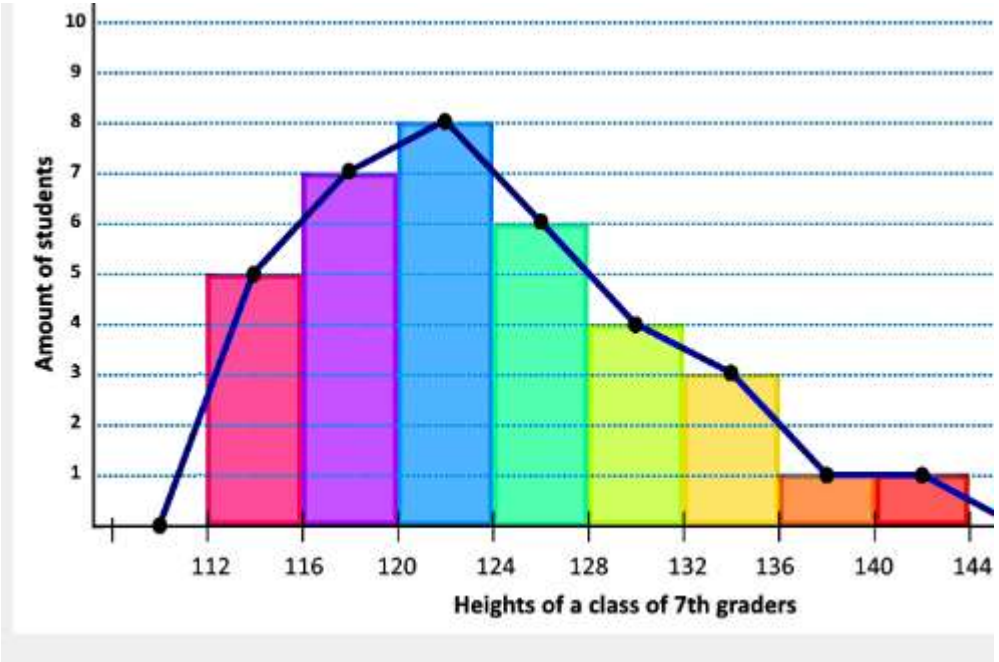
GRAPHICAL PRESENTATION OF DATA

Here frequencies are plotted on a pictorial platform formed of horizontal and vertical lines known as graph.

HISTOGRAM

Consist of series of rectangles with width equals class interval, on horizontal axis & corresponding frequency on the vertical axis.



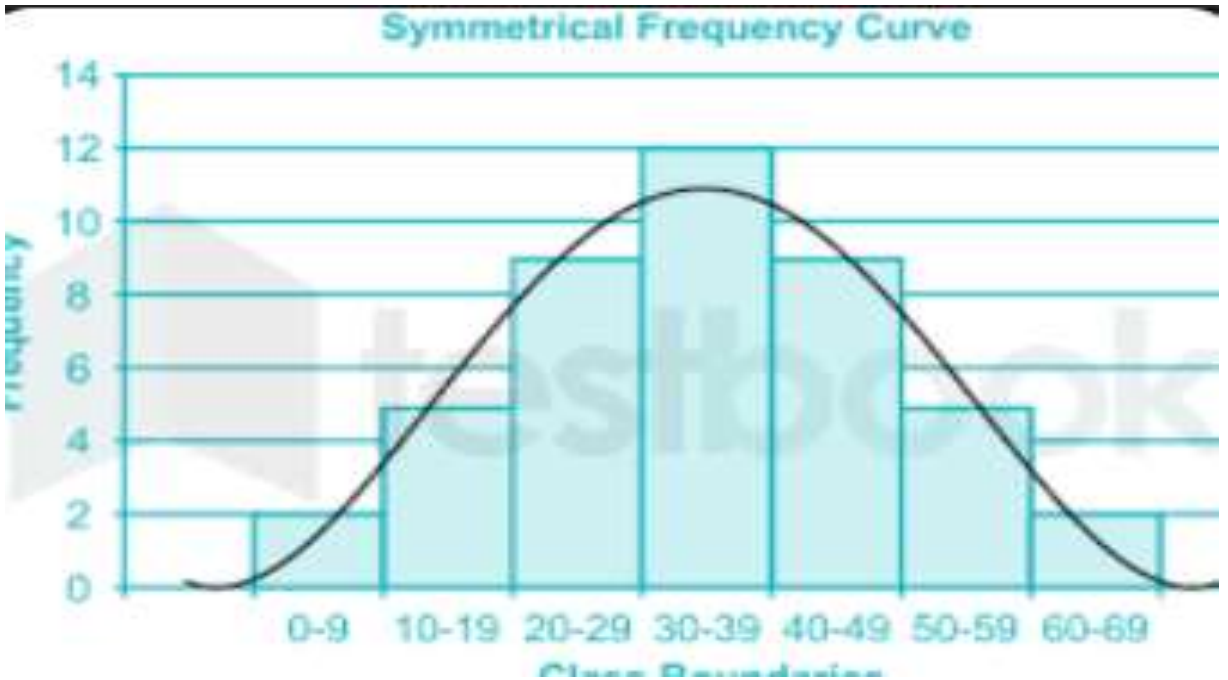


FREQUENCY POLYGON

Instead of rectangle in histogram, if the points of succession class marks are being connected then it becomes frequency polygon.it mainly used for comparison of 2 or more distribution.

FREQUENCY CURVE

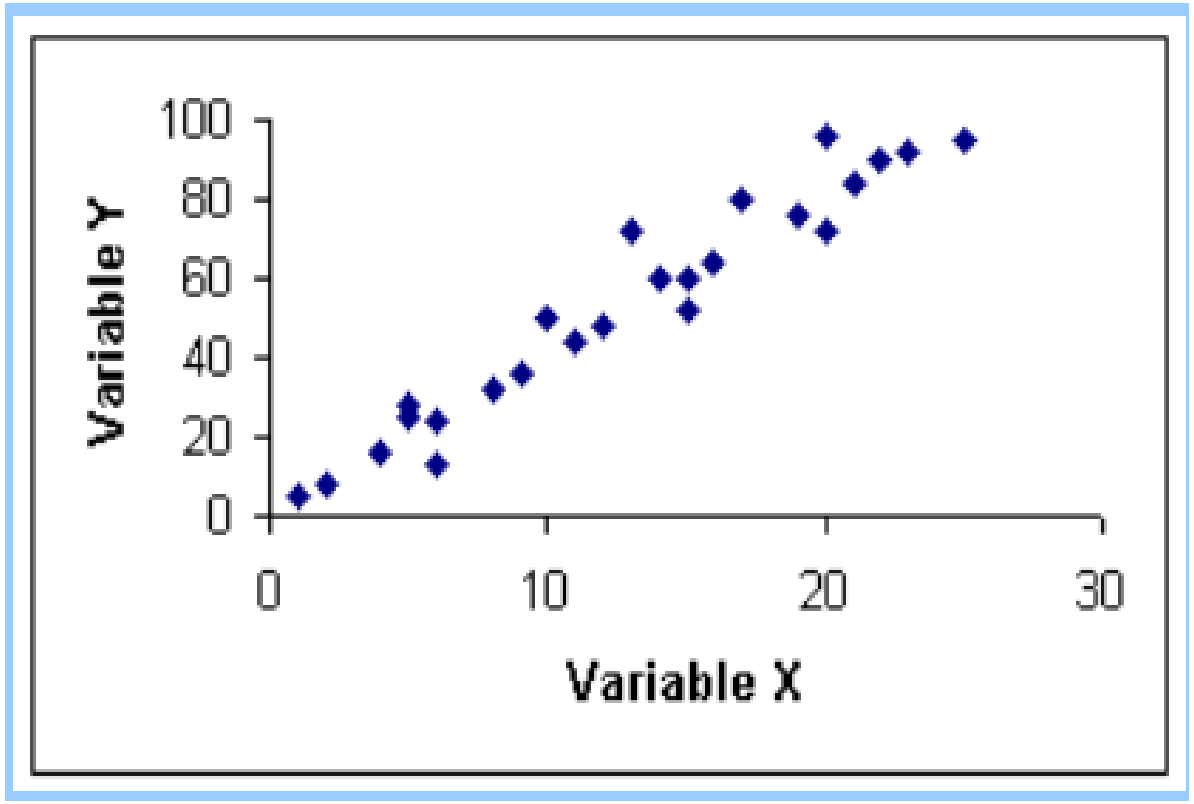
A smooth free hand curve drawn through frequency polygon.



SCATTER PLOTTER

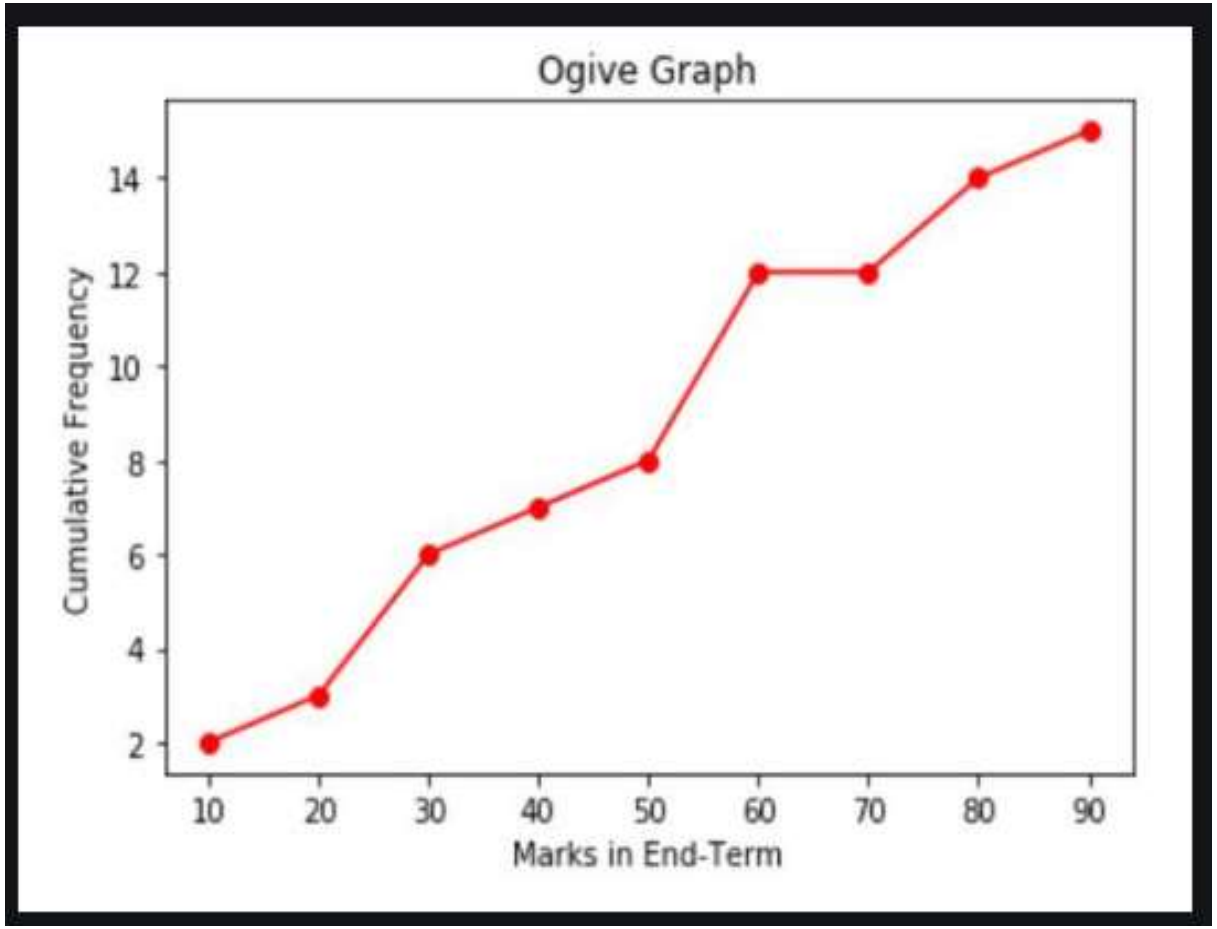
Graphical presentation of the relationship between 2 quantitative variable. Independent variable in x-axis dependent variable in y-axis.

Used for plotting continuous data.



OGIVE

Graph obtained by adding the frequencies up to the given value (cumulative frequency). These values are then listed in cumulative frequency table. The curve obtained by plotting their value is called ogive.

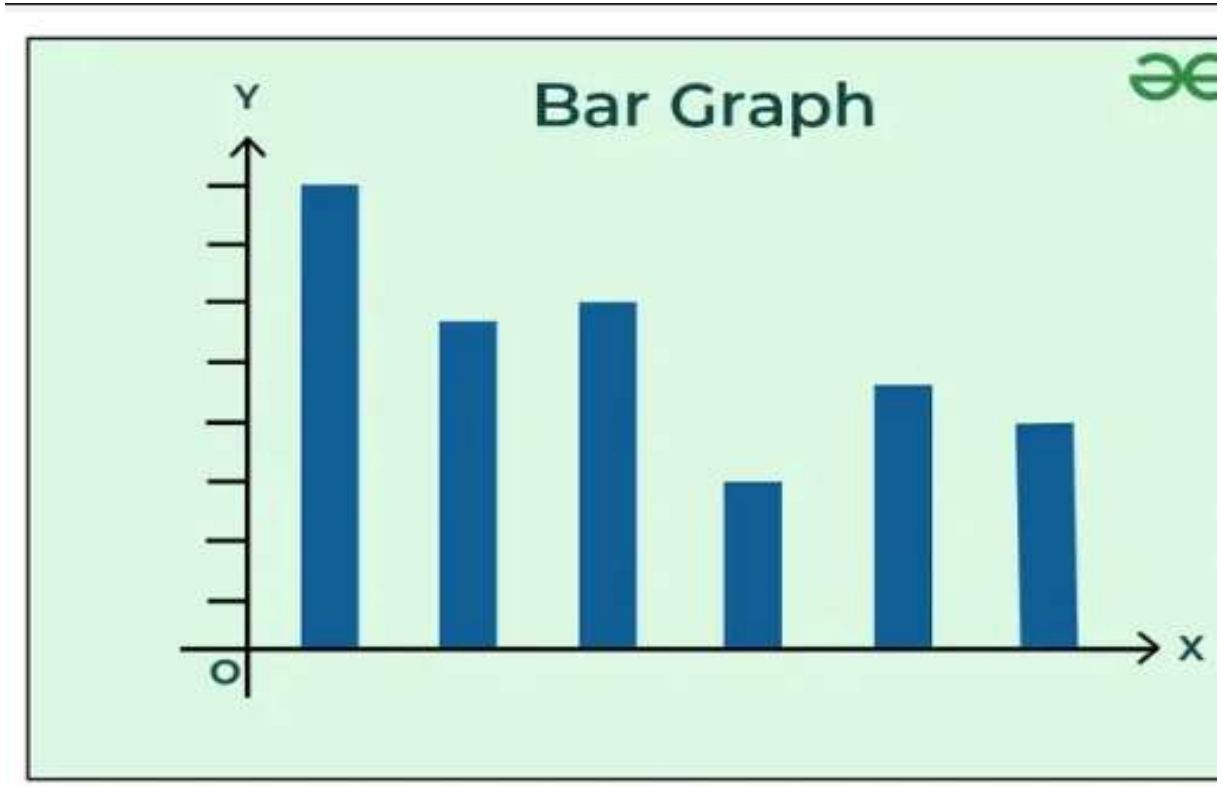


DIAGRAMMATIC PRESENTATION OF DATA

Visual form for the presentation of statistical data . Used to present the data in visual form.

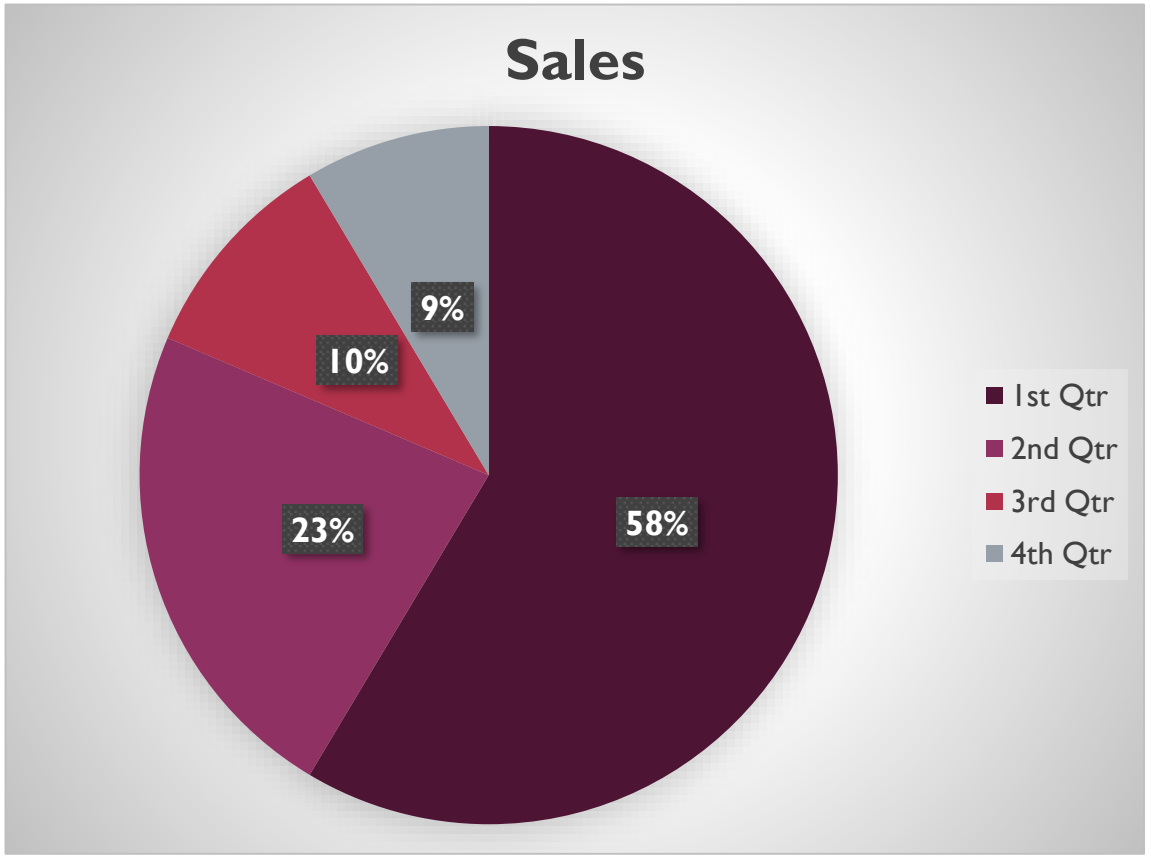
BAR DIAGRAM

Variable on horizontal axis and frequency on the vertical axis. Most useful for categorical data. The bars are separated by small gap.



PIE CHART

Used for categorical data. Circle divided into component sectors. Each sector will be proportional to frequency of the variable. Also known as angular diagram.



LEVEL OF SIGNIFICANCE

" Level of significance is the probability of rejecting the null hypothesis when it is true."

- Probability of committing a type I error.
- Typical values are 1%(0.01 level).
  5%(0.05 level).
- Selected by researcher in the beginning.
- Provide critical values of a test.
- Region of rejection in sampling distribution.
- Denoted by α
- It is a measurement of statistical significance.

LEVEL OF SIGNIFICANCE OR LEVEL OF CONFIDENCE

Confidence

- **99%**
- **95%**

Risk of committing type I error

- **0.01**
- **0.05**

It is the measure of trustworthiness.

DEGREE OF FREEDOM

Degree of freedom is the degree of freedom to "VARY".

Number of degree of freedom depends on number of restrictions placed upon the scores. With each restriction reducing one (df).

Mostly df = N - 1

STANDARD ERROR $SE = \frac{\sigma}{\sqrt{n}}$

The standard error of the mean, or simply standard error, indicates how different the population mean is likely to be from a sample mean would vary if you were to repeat a study using new samples from within a single population.

SE – Standard Error
σ – standard deviation
n- no. of elements.

A high standard error shows that sample means are widely spread around the population mean – your sample may not closely represent your population. A low standard error shows that sample means are closely distributed around the population mean – your sample is representative of your population.

Greater the sample size lesser the standard error

RANK ORDER CORRELATIONS

KENDALL'S TAU CORRELATION

KENDALL'S TAU

- Non-parametric test
- Measures the strength of dependence between 2 variables
- Lowest of each get rank 1
- Consider as the non-parametric alternative of Pearson's Product Moment correlation

STEPS

- Rank the data (Rx & Ry)
- Arrange Ry according to Rx ascending
- Compute TAU
- $\tau = \frac{C-D}{C+D}$ **or** $\tau = \frac{2s}{n(n-1)}$

INTERPRETATION

- **Ho => $\tau = 0$, there is no correlation between X and Y**
- **H_A => $\tau \neq 0$, there is a correlation between X and Y**

  & if $\tau > 0$ then we can say that there exist a positive correlation between the variables

  if $\tau < 0$ there exist a negative correlation between the variables

Range of τ is from -1 to +1

If the value is

0 which means no correlation

+1 means perfect positive correlation

-1 perfect negative correlation

± 0.7 to ± 0.9 strong positive/negative correlation

± 0.4 to ± 0.7 moderate positive/ negative correlation

± 0.1 to ± 0.4 weak positive/negative correlation

Compute Kendall's Tau for the following data

X –	5	3	7	6	2
Y -	7	8	9	6	4

ANSWER

Ho =>Null hypothesis $\tau = 0$, there is no correlation between X and Y

H_A =>Alternative hypothesis $\tau \neq 0$, there is a correlation between X and Y

Compute Kendall's Tau for the following data

Rx-	3	2	5	4	1
X –	5	3	7	6	2
Y -	7	8	9	6	4
Ry-	3	4	5	2	1

Separate the ranks and pair it

Rx –	3	2	5	4	1
Ry -	3	4	5	2	1

Arrange Ry according to Rx ascending

Rx - 1 2 3 4 5

Ry - 1 4 3 2 5

Compute τ C= no of +ve sign , D = no of – ve sign

Ry - 1	4	3	2	5	Total
1	+	+	+	+	= +4
	4	-	-	+	= -1
		3	-	+	= 0
			2	+	= +1
				5	----------
					S = +4

$$\tau = \frac{C - D}{C + D}$$

$$\tau = \frac{7-3}{7+3}$$

$$\tau = \frac{4}{10}$$

$$\tau = +0.4$$

or

$$\tau = \frac{2s}{n(n-1)}$$

$$\tau = \frac{2*4}{5(5-1)}$$

$$\tau = \frac{8}{20}$$

$$\tau = +0.4$$

INTERPRETATION

Since $\tau \neq 0$ we reject Ho and accept H_A

τ = +0.4 which means there exist a moderate positive correlation between X & Y

SPEARMAN'S RHO

SPEARMAN'S RHO (ρ)

- Data is in rank order format
- Rank order data present the rank of the individuals or subjects
- Non-parametric version of Pearson Product Moment Correlation
- Measure the strength of association between two ranked variable

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

TABULATION

Si no	X	Y	Rx	Ry	D=Rx-Ry	D^2
I						
2						
3						
4						
5						
.						
.						
.						
n						
						$\sum D^2$

Compute Spearman's Rho for the following data

Data A – 16, 19, 18, 10, 12, 13, 17, 9, 7, 5

Data B - 15, 17, 16, 9, 10, 12, 19, 8, 6, 4

ANSWER

Ho =>Null hypothesis ρ= 0, there is no correlation between X and Y

H_A =>Alternative hypothesis $\rho \neq 0$, there is a correlation between X and Y

Compute Spearman's Rho for the following data

Rx – 7 10 9 4 5 6 8 3 2 1

Data A – 16, 19, 18, 10, 12, 13, 17, 9, 7, 5

Data B - 15, 17, 16, 9, 10, 12, 19, 8, 6, 4

Ry- 7 9 8 4 5 6 10 3 2 1

Si no	X	Y	Rx	Ry	D=Rx-Ry	D^2
1	16	15	7	7	0	0
2	19	17	10	9	1	1
3	18	16	9	8	1	1
4	10	9	4	4	0	0
5	12	10	5	5	0	0
6	13	12	6	6	0	0
7	17	19	8	10	-2	4
8	9	8	3	3	0	0
9	7	6	2	2	0	0
10	5	4	1	1	0	0
						$\sum D^2$=6

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

$$\rho = 1 - \frac{6 * 6}{10(100 - 1)}$$

$$\rho = 1 - \frac{36}{990}$$

$$\rho = 1 - 0.0367$$

$$\rho = 0.963$$

Interpretation

$\rho \neq 0$ here we reject Ho and accept H_A

$\rho = 0.963$

We can say that there exist a strong positive correlation between the variables

Spearman's Rho with tied ranks

$$\rho = \frac{\sum R_x \cdot R_y - \frac{\sum R_x \sum Ry}{n}}{\sqrt{\left(\sum R_x^2 - \frac{(\sum R_x)^2}{n}\right)\left(\sum R_y^2 - \frac{\left(\sum R_y\right)^2}{n}\right)}}$$

Si no	**X**	**Y**	R_x	R_y	R_x^2	R_y^2	R_x*R_y
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
			$\sum R_x$=	$\sum R_y$=	$\sum R_x^2$=	$\sum R_y^2$=	$\sum R_x * R_y$=

$$\rho = \frac{\sum R_\chi \cdot R_y - \frac{\sum R_x \sum Ry}{n}}{\sqrt{\left(\sum R_x^2 - \frac{(\sum R_x)^2}{n}\right)\left(\sum R_y^2 - \frac{\left(\sum R_y\right)^2}{n}\right)}}$$

Ho – there is no correlation between the data $\rho = 0$

HA – there exist a correlation between the data $\rho \neq 0$

Compute Spearman's Rho for the following data.

Data 1 – 44, 45, 45, 34, 43, 23, 54, 34, 67, 45

Data 2 – 12, 21, 32, 12, 12, 15, 26, 12, 16, 12

ANSWER

Ho – there is no correlation between the data $\boldsymbol{\rho = 0}$

HA – there exist a correlation between the data $\boldsymbol{\rho \neq 0}$

Si no	**X**	**Y**	R_x	R_y	R_x^2	R_y^2	R_x*R_y
1	44	12					
2	45	21					
3	45	32					
4	34	12					
5	43	12					
6	23	15					
7	54	26					
8	34	12					
9	67	16					
10	45	12					
			$\sum R_x$=	$\sum R_y$=	$\sum R_x^2$=	$\sum R_y^2$=	$\sum R_x * R_y$=

Si no	**X**	**Y**	R_x	R_y	R_x^2	R_y^2	R_x*R_y
1	44	12	5	3	25	9	15
2	45	21	7	8	49	64	56
3	45	32	7	10	49	100	70
4	34	12	2.5	3	6.25	9	7.5
5	43	12	4	3	16	9	12
6	23	15	1	6	1	36	6
7	54	26	9	9	81	81	81
8	34	12	2.5	3	6.25	9	7.5
9	67	16	10	7	100	49	70
10	45	12	7	3	49	9	21
			$\sum R_x$=55	$\sum R_y$=55	$\sum R_x^2$=382.5	$\sum R_y^2$=375	$\sum R_x * R_y$=346

$$\rho = \frac{\sum R_\chi \cdot R_y - \frac{\sum R_x \sum Ry}{n}}{\sqrt{\left(\sum R_x^2 - \frac{(\sum R_x)^2}{n}\right)\left(\sum R_y^2 - \frac{\left(\sum R_y\right)^2}{n}\right)}}$$

$$= \frac{346 - \frac{55 * 55}{10}}{\sqrt{\left(382.5 - \frac{55 * 55}{10}\right)\left(375 - \frac{55 * 55}{10}\right)}}$$

$$= \frac{346 - 302.5}{\sqrt{(382.5 - 302.5)(375 - 302.5)}}$$

$$= \frac{43.5}{\sqrt{(80)(72.5)}}$$

$$= \frac{43.5}{\sqrt{5800}}$$

$$\rho = \frac{43.5}{76.158} = 0.571$$

WORK 2

Si no	X	Y	R_x	R_y	R_x^2	R_y^2	R_x*R_y
1	34	43	2.5	6.5	6.25	42.25	16.25
2	45	45	7.5	9	56.25	81	67.5
3	54	54	10	10	100	100	100
4	34	34	2.5	3	6.25	9	7.5
5	23	34	1	3	1	9	3
6	43	43	4.5	6.5	20.25	42.25	29.25
7	45	43	7.5	6.5	56.25	42.25	48.75
8	45	23	7.5	1	56.25	1	7.5
9	43	34	4.5	3	20.25	9	13.5
10	45	43	7.5	6.5	56.25	42.25	48.75
			$\sum R_x$=55	$\sum R_y$=55	$\sum R_x^2$=379	$\sum R_y^2$=378	$\sum R_x * R_y$=342

$$\rho = \frac{\sum R_\chi \cdot R_y - \frac{\sum R_x \sum Ry}{n}}{\sqrt{\left(\sum R_x^2 - \frac{(\sum R_x)^2}{n}\right)\left(\sum R_y^2 - \frac{\left(\sum R_y\right)^2}{n}\right)}}$$

$$= \frac{342 - \frac{55 * 55}{10}}{\sqrt{\left(379 - \frac{55 * 55}{10}\right)\left(378 - \frac{55 * 55}{10}\right)}}$$

$$= \frac{342 - 302.5}{\sqrt{(379 - 302.5)(378 - 302.5)}}$$

$$= \frac{39.5}{\sqrt{(76.5)(75.5)}}$$

$$= \frac{39.5}{76}$$

$\rho = 0.52$here we reject Ho accept HA . There is a moderate positive correlation between the data.

PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT(r)

- **Parametric test**
- **Measures the strength and direction of association that exists b/w 2 variables**
- **Range -1 to +1**
- **Value of r = 0 means there is no correlation between the variable**
- **Value of r < 0 means there is a negative correlation**
- **Value of r > 0 means there is a positive correlation**

HYPOTHESIS

- **Ho - Null Hypothesis – r = 0 there is no correlation between the variables**
- **HA – Alternative Hypothesis – r ≠ 0 there exist a correlation between the variables**

TABULATION

Si no	X	Y	x^2	y^2	X.Y
	Σx	Σy	Σx^2	Σy^2	ΣX.Y

Pearson's Product Moment Correlation Coefficient (r)

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

Compute Pearson's Product Moment Correlation Coefficient for the following data

X – 5, 5, 4, 4, 4, 3, 2, 2, 6, 5

Y – 6, 6, 4, 4, 4, 5, 5, 5, 4, 7

Hypothesis

- **Ho - Null Hypothesis – r = 0 there is no correlation between the variables**
- **H_A – Alternative Hypothesis – r ≠ 0 there exist a correlation between the variables**

Si no	X	Y	x^2	y^2	X.Y
1	5	6	25	36	30
2	5	6	25	36	30
3	4	4	16	16	16
4	4	4	16	16	16
5	4	4	16	16	16
6	3	5	9	25	15
7	2	5	4	25	10
8	2	5	4	25	10
9	6	4	36	16	24
10	5	7	25	49	35
	Σx =40	Σy =50	Σx^2 =176	Σy^2 =260	ΣX.Y =202

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

$$r = \frac{202 - \frac{40 * 50}{10}}{\sqrt{\left(176 - \frac{40^2}{10}\right)\left(260 - \frac{50^2}{10}\right)}}$$

$$r = \frac{202 - 200}{\sqrt{(176 - 160)(260 - 250)}}$$

$$\frac{2}{\sqrt{16 * 10}}$$

$$\frac{2}{12.649}$$

r = 0.158

Interpretation

r ≠ 0 hence we reject Ho and accept HA

r = 0.158 there exist a weak positive correlation between the given data

Compute Pearson's Product Moment Correlation Coefficient for the following data

X – 2, 3, 4, 7, 8, 9, 2, 3, 4, 8

Y – 10, 7, 8, 2, 3, 1, 10, 10, 7, 2

Si no	X	Y	x^2	y^2	X.Y
1	2	10	4	100	20
2	3	7	9	49	21
3	4	8	16	64	32
4	7	2	49	4	14
5	8	3	64	9	24
6	9	1	81	1	9
7	2	10	4	100	20
8	3	10	9	100	30
9	4	7	16	49	28
10	8	2	64	4	16
	Σx =50	Σy =60	Σx^2 =316	Σy^2 =480	ΣX.Y =214

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

$$r = \frac{214 - \frac{50 * 60}{10}}{\sqrt{\left(316 - \frac{50^2}{10}\right)\left(480 - \frac{60^2}{10}\right)}}$$

$$r = \frac{214 - 300}{\sqrt{(316 - 250)(480 - 360)}}$$

$$\frac{-86}{\sqrt{66 * 120}}$$

$$\frac{-86}{88.99}$$

r = -0.966

Interpretation

r ≠ 0 hence we reject Ho and accept HA

r = -0.966 there exist a strong negative correlation between the given data

Compute Pearson's Product Moment Correlation Coefficient for the following data

X – 10, 12, 13, 11, 9, 8, 12, 5, 10, 10

Y – 12, 13, 5, 12, 13, 15, 10, 10, 10, 10

ONE WAY ANOVA

- **ANALYSIS OF VARIANCE (ANOVA)**

- **Also called ‘F’ statistics or ‘F’ test**
- **Used to compare the means of 3 or more populations**
- **One independent variable is studied. Hence called ONE WAY ANOVA**

1. DEFINE HYPOTHESIS

Ho – Null Hypothesis – All means are equal
Or
There is NO significant difference in all population

H_A – at least one mean is different
Or
There is significant difference in all populations

2. DEFINE VARIABLE

K = ---------- (No of groups)
n= ---------- (No of samples in one group)
N =---------- (total No of observations)

3. DEGREE OF FREEDOM

dfb – (K -1) (degree of freedom between variable)

dfw – (N – K) (degree of freedom within variable)

dft – (N – 1) (total degree of freedom)

Si no	GROUP A		GROUP B		GROUP C	
	x_A	x_{A^2}	x_B	x_{B^2}	x_C	x_{C^2}
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
	$\sum x_A =$	$\sum x_{A^2} =$	$\sum x_B =$	$\sum x_{B^2} =$	$\sum x_C =$	$\sum x_{C^2} =$
	$(\sum x_A)^2 =$		$(\sum x_B)^2 =$		$(\sum x_C)^2 =$	

SUM OF SQUARES BETWEEN

$$SS_B = \frac{\sum(\Sigma x)^2}{n} - \frac{(\sum(\Sigma x))^2}{N}$$

$$\mathbf{SSB=}\frac{(\Sigma x_A)^2}{n} + \frac{(\Sigma x_B)^2}{n} + \frac{(\Sigma x_C)^2}{n} - \frac{\left((\Sigma x_A + \Sigma x_B + \Sigma x_C)\right)^2}{N}$$

SUM OF SQUARES WITHIN

$$SSw = \left(\sum X_A^2 - \frac{(\Sigma x_A)^2}{n}\right) + \left(\sum X_B^2 - \frac{(\Sigma x_B)^2}{n}\right) + \left(\sum X_C^2 - \frac{(\Sigma x_C)^2}{n}\right)$$

MEAN SQUARE BETWEEN

$$\text{MSB} = \frac{ss_B}{\mathrm{d}f_B}$$

MEAN SQUARE WITHIN

$$\text{MSW} = \frac{ssw}{\mathrm{d}fw}$$

$$\text{F ratio} = \frac{MSB}{MSW}$$

SUMMARY TABLE OF ANOVA

Source of variance	Sum of squares	Degree of freedom	Mean square	F ratio
Between variable	SSB=	dfB=	MSB=	
Within variable	SSW=	dfW=	MSW=	
total	SST=	dfT=		

INTERPETATION

Critical Value @ 0.05 level of significance =

Critical value @ 0.01 level of significance =

F calculated < F critical

Accept Ho and reject HA

F calculated > F critical

Reject Ho and accept HA

Compute ANOVA for the following data

Group A – 2, 3, 4, 5, 2, 3, 4, 3, 2, 1

Group B- 3, 4, 3, 2, 6, 5, 4, 3, 3, 2

Group C -4, 5, 4, 3, 3, 3, 3, 3, 3, 3

Ho – there is no significant difference in the means of the populations

HA – there exist significant difference between the means of the population

Si no	GROUP A		GROUP B		GROUP C	
	x_A	x_{A^2}	x_B	x_{B^2}	x_C	x_{C^2}
1	2	4	3	9	4	16
2	3	9	4	16	5	25
3	4	16	3	9	4	16
4	5	25	2	4	3	9
5	2	4	6	36	3	9
6	3	9	5	25	3	9
7	4	16	4	16	3	9
8	3	9	3	9	3	9
9	2	4	3	9	3	9
10	1	1	2	4	3	9
	$\sum x_A = 29$	$\sum x_{A^2} = 97$	$\sum x_B = 35$	$\sum x_{B^2} = 137$	$\sum x_C$=34	$\sum x_{C^2} = 120$
	$(\sum x_A)^2$=841		$(\sum x_B)^2$=1225		$(\sum x_C)^2$=1156	

Define variables

K = 3
n=10
N=30

Degree of freedom

dfB = 2

dfW= 27

dfT=29

$$SS_B = \frac{\sum(\Sigma x)^2}{n} - \frac{(\sum(\Sigma x))^2}{N}$$

$$\text{SSB}=\frac{29^2}{10}+\frac{35^2}{10}+\frac{34^2}{10}-\frac{\left((29+35+34)\right)^2}{30}=\frac{841}{10}+\frac{1225}{10}+\frac{1156}{10}-\frac{98^2}{30}$$

$=322.2 - 320.13 = 2.067$ $\qquad SS_B = 2.067$

$$SSw = \left(\sum X_A^2 - \frac{(\Sigma x_A)^2}{n}\right) + \left(\sum X_B^2 - \frac{(\Sigma x_B)^2}{n}\right) + \left(\sum X_C^2 - \frac{(\Sigma x_C)^2}{n}\right)$$

$$=\left(97-\frac{29^2}{10}\right)+\left(137-\frac{35^2}{10}\right)+\left(120-\frac{34^2}{10}\right)$$

$=(97-84.1) + (137 - 122.5) + (120 - 115.6)$

$= 12.9 + 14.5 + 4.4$

$SSw = 31.8$

$$\text{MSB} = \frac{SS_B}{df_B} = 2.067/2 = 1.0335$$

$$\text{MSW} = \frac{SS_w}{df_w} = 31.8/27 = 1.178$$

$$\text{F ratio} = \frac{MSB}{MSW} = 1.0335/1.178 = 0.877$$

Source of variance	Sum of squares	Degree of freedom	Mean square	F ratio
Between variable	SSB=2.067	dfB=2	MSB= 1.0335	0.877
Within variable	SSW=31.8	dfW=27	MSW= 1.178	
total	SST=33.867	dfT=29		

INTERPETATION

Critical Value @ 0.05 level of significance = 3.35

Critical value @ 0.01 level of significance = 5.49

@0.05 level of significance

F calculated	<	F critical
0.877	<	3.35

@ 0.01 level of significance

F calculated	<	F critical
0.877	<	5.49

Hence we accept Ho and reject HA .We can say that there is no significant difference between the means of all the populations.

Compute ANOVA for the following data

Group A – 4, 6, 7, 9, 2, 4, 7, 8, 9, 2

Group B-6, 7, 10, 12, 13, 14, 15, 17, 18, 3

Group C -10, 11, 12, 13, 14, 15, 17, 10, 2, 4

Ho – there is no significant difference in the means of the populations

HA – there exist significant difference between the means of the population

Si no	GROUP A		GROUP B		GROUP C	
	x_A	x_{A^2}	x_B	x_{B^2}	x_C	x_{C^2}
1	4	16	6	36	10	100
2	6	36	7	49	11	121
3	7	49	10	100	12	144
4	9	81	12	144	13	169
5	2	4	13	169	14	196
6	4	16	14	196	15	225
7	7	49	15	225	17	289
8	8	64	17	289	10	100
9	9	81	18	324	2	4
10	2	4	3	9	4	16
	$\sum x_A$ =58	$\sum x_{A^2}$ =400	$\sum x_B$ =115	$\sum x_{B^2}$ =1541	$\sum x_C$=108	$\sum x_{C^2}$ = 1364
	$(\sum x_A)^2$=3364		$(\sum x_B)^2$=13225		$(\sum x_C)^2$=11664	

Define variables

K = 3
n=10
N=30

Degree of freedom

dfB = 2

dfW= 27

dfT=29

$$SS_B = \frac{\sum(\Sigma x)^2}{n} - \frac{(\sum(\Sigma x))^2}{N}$$

$$\mathbf{SSB=}\frac{(\Sigma x_A)^2}{n} + \frac{(\Sigma x_B)^2}{n} + \frac{(\Sigma x_C)^2}{n} - \frac{\left((\Sigma x_A + \Sigma x_B + \Sigma x_C)\right)^2}{N}$$

$$3364/10+13225/10+11664/10 - \frac{\left((58+115+108)\right)^2}{30}$$

$$336.4+1322.5+1166.4 - \frac{\left((281)\right)^2}{30}$$

=2825.3 – 2632.03 = SSB =193.27

$$SSw = \left(\sum X_A^2 - \frac{(\Sigma x_A)^2}{n} \right) + \left(\sum X_B^2 - \frac{(\Sigma x_B)^2}{n} \right) + \left(\sum X_C^2 - \frac{(\Sigma x_C)^2}{n} \right)$$

= (400 – 336.4) +(1541 – 1322.5) + (1364 – 1166.4)

= 63.6 + 218.5 + 197.6

SSW = 479.7

MSB $= \frac{SS_B}{df_B} =$ 193.27/2

=96.635

MSW $= \frac{SS_w}{df_w}$ = 479.7/27

= 17.767

F ratio $= \frac{MSB}{MSW}$ = 96.635 /17.767

= 5.439

Source of variance	Sum of squares	Degree of freedom	Mean square	F ratio
Between variable	SSB=193.27	dfB=2	MSB= 96.635	5.439
Within variable	SSW=479.7	dfW=27	MSW=17.767	
total	SST=672.97	dfT=29		

Critical Value @ 0.05 level of significance = 3.35

Critical value @ 0.01 level of significance = 5.49

@0.05 level of significance

F calculated	>	F critical
5.439	>	3.35

Hence we reject Ho and accept Ha , there is significant difference between the groups

@0.01 level of significance CV = 5.49

F calculated	<	F critical
5.439	<	5.49

here we accept Ho .There is no significant difference b/w the variables

CHI SQUARE TEST

CHI SQUARE TEST (χ^2)

- Non- parametric test
- O – observed data , E – expected data
- It compares the expected data with the observed data
- Used to determine the extent to which O matches E
- Or to determine whether there is a significant association between 2 variables (or more)

$$E = \frac{\text{ROW TOTAL x COLUMN TOTAL}}{\text{ALL TOTAL}}$$

HYPOTHESIS

Ho – Null Hypothesis – there is no relationship between the variables

H_A - Alternative Hypothesis – there exist a relationship between the variables

TABULATION - OBSERVED

VARIABLE	RESPONSE		ROW TOTAL
	O1	O2	$\Sigma =$
	O3	O4	$\Sigma =$
COLUMN TOTAL	$\Sigma =$	$\Sigma =$	$\Sigma =$ (ALL TOTAL)

Find expected values of all observations

$$E = \frac{R_T \times C_T}{A_T}$$

χ^2 TABULATION

variable	response	0	E	$(0-E)$	$(0-E)^2$	$\frac{(0-E)^2}{E}$

χ^2 CALCULATION

$$\chi^2 = \sum \frac{(0-E)^2}{E}$$

INTERPRETATION

χ^2calculated < χ^2 critical

Accept Ho & Reject H_A

χ^2calculated > χ^2 critical

Accept HA & Reject H_o

Lets solve a sample problem

Compute chi square for the following data

Gender	Response	
	Yes	No
Male	15	5
Female	10	10

Compute Row Total , Column total & All total

Ans

Gender	Response		RT
	Yes	No	
Male	15	5	=20
Female	10	10	=20
CT	=25	=15	AT =40

Compute E

$$E = \frac{R_T x C_T}{A_T}$$

(Note : Σ CT = Σ RT = AT)

$$E = \frac{R_T x C_T}{A_T}$$

$E1 = \frac{20x25}{40}$ =12.5	$E2 = \frac{20x15}{40}$ =7.5
$E3 = \frac{20x25}{40}$ =12.5	$E4 = \frac{20x15}{40}$ =7.5

0	E	$(0-E)$	$(0-E)^2$	$\frac{(0-E)^2}{E}$
15	12.5	2.5	6.25	6.25/12.5 =0.5
5	7.5	-2.5	6.25	6.25/7.5 =0.833
10	12.5	-2.5	6.25	6.25/12.5 =0.5
10	7.5	2.5	6.25	6.25/7.5 =0.833
		$\chi^2 = \sum \frac{(0-E)^2}{E}$		=2.67

TABULATION

variable	response	0	E	$(0-E)$	$(0-E)^2$	$\frac{(0-E)^2}{E}$
Male	Yes	10	12.5	2.5	6.25	6.25/12.5 = 0.5
	No	5	7.5	-2.5	6.25	6.25/7.5 = 0.833
Female	Yes	10	12.5	-2.5	6.25	6.25/12.5 = 0.5
	No	10	7.5	2.5	6.25	6.25/7.5 = 0.833
				$\chi^2 = \sum \frac{(0-E)^2}{E}$		= 2.67

$$\chi^2 = \sum \frac{(0-E)^2}{E} = 2.67$$

INTERPRETATION

@0.05 level of significance critical value = 3.84

@0.01 level of significance critical value = 6.63

χ^2calculated < χ^2 critical @0.05

2.67 < 3.84

χ^2calculated < χ^2 critical @0.01

2.67 < 6.63

Hence we accept Ho. There is no relationship between the variables.

Compute χ^2 for the following data

	Agree	Undecided	Disagree
Junior manger	10	15	5
Senior manager	10	5	15

	yes	No	Undecided	RT
Male	10	15	5	=30
Female	10	5	15	=30
CT=	=20	=20	=20	=60(AT)

$$E = \frac{R_T \times C_T}{A_T}$$ =20x30/60 = 10

O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
10	10	0	0	0
15	10	5	25	25/10 =2.5
5	10	-5	25	2.5
10	10	0	0	0
5	10	-5	25	2.5
15	10	5	25	2.5

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 10$$

$$\chi^2 = \sum \frac{(0-E)^2}{E}$$

= 10

@0.05 level of significance critical value = 5.99

@0.01 level of significance critical value = 9.21

χ^2calculated > χ^2 critical @0.05

10 > 5.99

χ^2calculated > χ^2 critical @0.01

10 > 9.21

Hence we reject Ho and accept HA. There exist a relationship between the variables.

WILCOXON MATCHED PAIR SIGNED RANK TEST(W)

WILCOXON MATCHED PAIR SIGNED RANK TEST

- ❑ Used to test the median difference
- ❑ Based on rank order difference rather than actual.
- ❑ Alternate to paired T-test
- ❑ Non parametric test
- ❑ Data measured on Ordinal, Ratio, interval scales
- ❑ Used on related data
- ❑ Used when groups will be in equal in size

HYPOTHESIS

Ho – the median difference between the variable is zero
Or
There is no median difference between the variables

HA – the median difference between the variable is not zero
Or
There exist a median difference between the variables.

(note :- rank the differences – ignore zero, ignore +ve and –ve signs while ranking . For tied ranks use mean rank)

Si no	X	Y	X-Y	R	+	-
1						
2						
3						
.						
.						
n						
					$\sum R^{+}=W^{+}$	$\sum R^{-}=W^{-}$

Add the ranks of positive values =W^{+}

Add the ranks of negative values = W^{-}

Verify - $W^{-}+W^{+}=\frac{n[n+1]}{2}$

{note :- n will be the no of ranked data}

Take smaller among W^{-} & W^{+} as Wilcoxon W

Interpretation : W calculated > W critical

accept Ho and reject HA

: W calculated < W critical

accept HA and reject Ho

Find the Wilcoxon signed rank test for the following data

Left ear – 25, 24, 10, 31, 27, 24, 27, 29, 30, 32, 20, 5

Right ear – 32, 30, 7, 36, 20, 32, 26, 33, 32, 32, 30, 32

Si no	X	Y	X-Y	R	+	-
1	25	32				
2	24	30				
3	10	7				
4	31	36				
5	27	20				
6	24	32				
7	27	26				
8	29	33				
9	30	32				
10	32	32				
11	20	30				
12	5	32				
					$\sum R^{+}=W^{+}$ =	$\sum R^{-}=W^{-}$ =

Si no	X	Y	X-Y	R	+	-
1	25	32	-7	7.5		7.5
2	24	30	-6	6		6
3	10	7	3	3	3	
4	31	36	-5	5		5
5	27	20	7	7.5	7.5	
6	24	32	-8	9		9
7	27	26	1	1	1	
8	29	33	-4	4		4
9	30	32	-2	2		2
10	32	32	0	********(IGNORE ZERO)********		
11	20	30	-10	10		10
12	5	32	-27	11		11
					$\sum R^{+}=W^{+}$ =11.5	$\sum R^{-}=W^{-}$ =54.5

Take smaller among W^- & W^+ as Wilcoxon W

Wilcoxon W = 11.5

Critical value of W with n = 11 is 11

W calculated > W critical

11.5 > 11

Hence we accept Ho . There is no rank difference in the median of the given data.

VERIFICATION - $W^- + W^+ = \frac{n[n+1]}{2}$

$$54.5+11.5 = \frac{11(11+1)}{2}$$

$$66 = 66$$

Hence verified.

REGRESSION

REGRESSION ANALYSIS

Regression analysis means the estimation or prediction of the unknown value of one variable from the known value of the other variable.

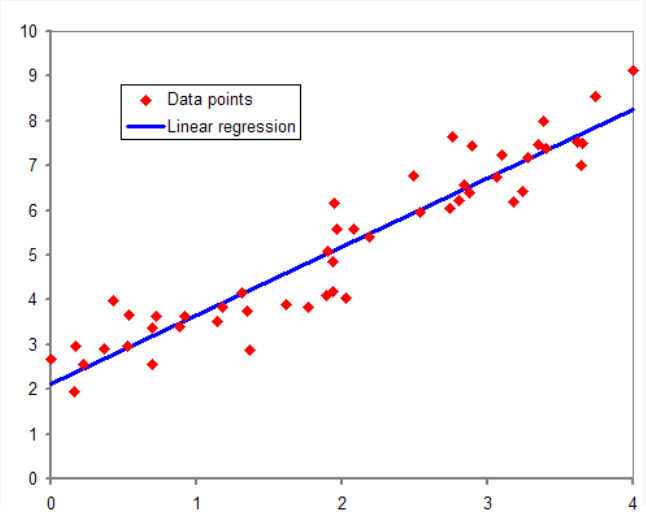
It creates an equation so that values can be predicted within the range framed by the data.

If Y is known then one can predict X

If X is known then one can predict Y

LINEAR REGRESSION

Linear regression analysis is a simple regression type that requires you to create a hypothetical line that best connects all data points. The disadvantage of linear regression is the potential for outliers in the data so it's frequently used for small data pools of information or predictions. This is because some data points may not fit neatly into the regression line.



REGRESSION

Psychologists use Regression to predict a number of human behaviors

Independent variable → predicts → Dependent variable

X → Y

We say X is regressed on Y

FOR EXAMPLE

Stress → leads to → **Health issues**

predicts

regression

Ho – there is no relationship between dependent and independent variable

HA - there exists a relationship between the dependent and independent variable

Sl no	X	Y	X^2	Y^2	XY
1					
2					
.					
.					
N					
Total	Σx	Σy	$\sum x^2$	$\sum y^2$	$\sum xy$

Calculation

Mean $\bar{x} = \frac{\Sigma x}{n}$ $\bar{y} = \frac{\Sigma y}{n}$

$(\Sigma x)^2 =$ $(\Sigma y)^2 =$

Standard deviation

$$sx^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$sy^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

Covariance between X & Y

$$\operatorname{cov} x\, y = \Sigma xy - \frac{\Sigma x \cdot \sum y}{n}$$

$$b = \frac{\operatorname{cov} x\, y}{s_{x^2}} \qquad a = \bar{y} - b\bar{x}$$

Regression equation Y on X $y = a + bx$

Regression equation X on Y $x = a + by$

note here to calculate $b = \frac{\operatorname{cov} xy}{s_{y^2}}$

$a = \bar{x} - b\bar{y}$

work

Calculate regression equation for the following data

X – 2, 3, 2, 5, 8

Y – 3, 2 , 4, 1, 10

Answer

Ho – there is no relationship between dependent and independent variable

HA - there exists a relationship between the dependent and independent variable

Sl no	X	Y	X^2	Y^2	XY
1	2	3	4	9	6
2	3	2	9	4	6
3	2	4	4	16	8
4	5	1	25	1	5
5	8	10	64	100	80
Total	Σx=20	$\Sigma y = 20$	$\sum x^2 = 106$	$\sum y^2 = 130$	$\sum xy = 105$

Calculation

Mean $\bar{x} = \frac{\sum x}{n}$ = 20/5 =4

$\bar{y} = \frac{\sum y}{n}$ =20/5 =4

Standard deviation

$$sx^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

=106-20^2/5
=106- 400/5
=106 – 80
= 26

$sx^2 = 26$

$$sy^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

= 130- 20^2/5
= 130 – 400/5
= 130 – 80
= 50

$sy^2 = 50$

$$\operatorname{cov} x\, y = \Sigma xy - \frac{\Sigma x \cdot \sum y}{n}$$

$= 105 - 20^2/5$

$= 105 - 400/5$

$= 105 - 80$

$\operatorname{cov} x\, y = 25$

$$b = \frac{\operatorname{cov} x\, y}{s_{x^2}}$$

$$b = \frac{25}{26}$$ b =0.961

$$a = \bar{y} - b\bar{x}$$

$a = 4 - 0.961 \text{x } 4$

$= 4 - 3.84$

$a = 0.156$

Regression equation Y on X $y = a + bx$

Y = 0.156 + 0.961X

Let’s do a new one

Calculate regression equation for the following data

X – 2 , 7 , 8 , 3 , 5

Y – 10 , 12 , 3 , 10 , 10

DESCRIPTIVE VS INFERENTIAL STATISTICS

DESCRIPTIVE STATISTICS

- Descriptive Statistics describes the characteristics of a data set. It is a simple technique to describe, show and summarize data in a meaningful way. You simply choose a group you're interested in, record data about the group, and then use summary statistics and graphs to describe the group properties. There is no uncertainty involved because you're just describing the people or items that you actually measure. You're not aiming to infer properties about a large data set.

INFERENTIAL STATISTICS

- Inferential statistics involves drawing conclusions about populations by examining samples. It allows us to make inferences about the entire set, including specific examples within it, based on information obtained from a subset of examples. These inferences rely on the principles of evidence and utilize sample statistics as a basis for drawing broader conclusions.

DESCRIPTIVE VS INFERENTIAL

DESCRIPTIVE STATISTICS

- Summarize essential features of the data.
- It concerned with describing the population under study
- Organize, analyze & present data in a meaningful way
- Find results in graphs, charts & tables
- Deals with central tendency & spread of frequency distribution
- Conclusions cannot be made beyond the given data.

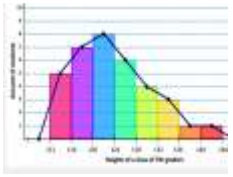
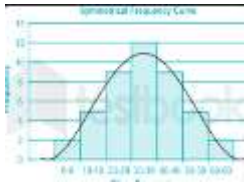
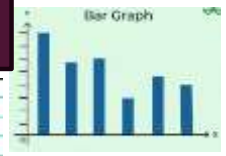
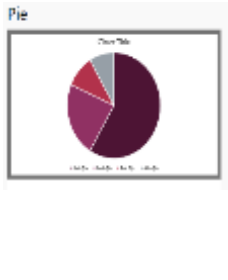
INFERENTIAL STATISTICS

- Estimates predictions, forecasts, generalize.
- Focused on drawing conclusions about the population on the basis of sample analysis and observation.
- Compares, test and predicts data
- Final result in probability
- More details such as hypothesis tests and confidence intervals are studied.
- The educated predictions & guesses can be made on the basis of the parameters of the given population.

DESCRIPTIVE STATISTICS

Organize

- tables
- Graphs
  - histogram
  - Frequency polygon
  - Frequency curve
  - Bar diagram
  - Pie Chart

Summarize

- Central tendency
  - Mean
  - Median
  - Mode
- Variation
  - Mean deviation
  - Standard deviation
  - varian ce

ADVANTAGES OF DESCRIPTIVE STATISTICS

- **It collects and summarizes large amount of data and information in a manageable and organized manner.**
- **It is fairly straight forward process that can easily translate results into a distribution of frequency, percentage and overall averages.**
- **It forms the basis of rigorous data analysis**
- **It is easier to work with, interpret, and discuss than raw data.**
- **Helps in examining the tendencies, variability and normality of a data set.**
- **It can be rendered both graphically and numerically.**
- **It forms the basis for more advanced statistical methods.**
- **Deals with immediate data and single variables rather than trying to establish conclusions.**

DISADVANTAGES OF DESCRIPTIVE STATISTICS

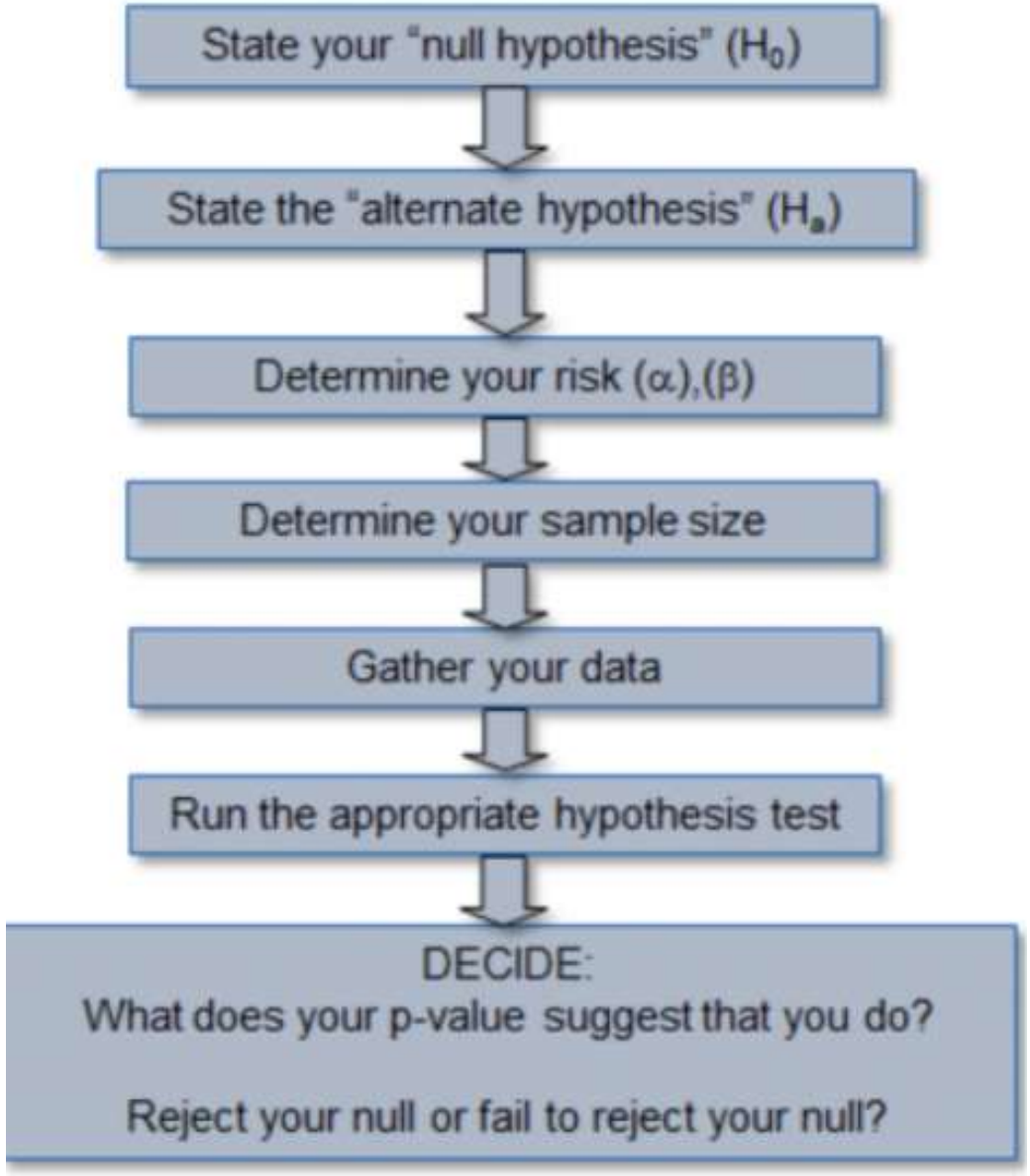
- **It can be misused, misinterpreted and is incomplete.**
- **It offers little information about causes and effects**
- **It can be dangerous if not analyzed completely**
- **Can be of limited use when samples and populations are small**
- **There is a risk of distorting the original data or losing important details.**
- **It does not account for randomness or provide statistical calculations that can lead to hypothesis or theories of population studied.**

HYPOTHESIS TESTING

Hypothesis testing in statistics is a way for you to test the results of an experiment to see if you have meaningful results. The outcome of your hypothesis test will tell you whether your results are caused by pure chance or not. If your results are due to pure chance, then you will have a difficult time replicating or repeating them. But, if your results were due to something significant happening, you can use that knowledge to repeat the results.

STEPS IN HYPOTHESIS TESTING

The hypothesis testing process and analysis involves using sample data to determine whether or not you can be statistically confident that you can reject the Ho. If the Ho is rejected, the statistical conclusion is that the alternative or alternate hypothesis Ha is true.



PROCESS

1. Restate the question as a Research Hypothesis & Null Hypothesis about the population. (provide suitable examples)
2. Determine the characteristics of the comparison distribution,(mean or median must be equal)
3. Determine the cutoff sample score on the comparison distribution at which the Null Hypothesis should be rejected.(C.V are based on degree of freedom & level of significance)
4. Determine sample score on the comparison distribution(statistical test)
5. Decide whether to reject or accept the Null Hypothesis.

DECISION ERRORS
(Situations in which the right procedure leads to the wrong decisions)

TYPE I ERROR

- **Rejection of true Ho(Null Hypothesis)**
- **Probability of committing type I error is called level of significance.**
- **Often denoted by α (alpha)**
- **Value of α is always set before the experiment or study is undertaken.**
- **It is the probability of overreacting.**
- **α= 0.05 means there is 5% risk of making type I error.**

TYPE II ERROR

- **Acceptance of Ho when it is false**
- **The probability of not committing a type II error is called the power of the test**
- **Often denoted by β (beta)**
- **β is not usually stated at the begining of the hypothesis testing procedure.**
- **It is the probability of under reacting.**

		THE TRUTH	
		The null hypothesis (H_0) is true (H_a is false)	**The null hypothesis (H_0) is not true** (H_a is true)
THE DECISION THE ANALYST MAKES	**Reject H_0** (support H_a)	**TYPE I (α) error/ Alpha Risk/ p – value** **Overreacting** **(1 - α) = the Confidence level of the test**	**Correct Decision (1 - β)** **Power of the test**
	Fail to Reject H_0 (do not support Ha)	**Correct Decision**	**TYPE II (β) error/ Beta Risk** **Underreacting**

KRUSKAL WALLIS ANOVA

- Rank based non parametric test.
- Used when assumptions of one way ANOVA doesn't meet .
- It tells if there is any significant difference between the groups. However it won't tell you which group is different .
- It computed with medians not mean.
- Test data should be transferred to rank order format.
- **Just like MANN WHITNEY U TEST Kruskal Wallis ANOVA also rank the whole data together.**
- It compares the critical value to chi square table value.

Ho – Null Hypothesis – there is no significant difference between the groups.

HA – Alternative Hypothesis – there exist significant difference between the groups .

(Rank the data as a whole.)

TABULATION

Si no	X_A	R_A	X_B	R_B	X_C	R_C
1						
2						
3						
.						
.						
n						
		$\sum R_A=$		$\sum R_B=$		$\sum R_C=$
	$(\sum R_A)^2=$		$(\sum R_B)^2=$		$(\sum R_C)^2=$	

CALCULATION

$$H = \frac{12}{N(N+1)} \Sigma \frac{R^2}{n} - 3(N+1)$$

$$N = n_A + n_B + n_C$$

$$\sum \frac{R^2}{n} = \frac{(\Sigma R_A)^2}{n} + \frac{(\sum R_B)^2}{n} + \frac{(\sum R_C)^2}{n}$$

INTERPRETATION

H calculated ≤ χ^2 critical

Accept Ho

H calculated > χ^2 critical

Reject Ho & accept HA

WORK

Reaction time of Group A , Group B , Group C are given below. Compute Kruskal Wallis ANOVA for the groups.

A – 15, 21, 26, 33

B – 36, 23, 37, 12

C – 20, 25, 30, 38

Critical value for χ^2 square @ 0.05 level of significance = 5.99

@ 0.01 level of significance = 9.21

Ho – Null Hypothesis – there is no significant difference between the groups.

HA – Alternative Hypothesis – there exist significant difference between the groups .

Si no	X_A	R_A	X_B	R_B	X_C	R_C
1	15	2	36	10	20	3
2	21	4	23	5	25	6
3	26	7	37	11	30	8
4	33	9	12	1	38	12
		$\sum R_A$=22		$\sum R_B$=27		$\sum R_C$=29
	$(\sum R_A)^2$=484		$(\sum R_B)^2$=729		$(\sum R_C)^2$=841	

$$H = \frac{12}{N(N+1)} \Sigma \frac{R^2}{n} - 3(N+1)$$

$$= \frac{12}{12(12+1)} * \frac{484+729+841}{4} - 3(12+1)$$

$$= \frac{12}{12*13} * \frac{2054}{4} - 3*13$$

$$= 39.5 - 39$$

$$\underline{H = 0.5}$$

H calculated **<** **χ^2 critical**

0.5 < 5.99 &9.21(0.05 & 0.01 level)

Hence we Accept Ho. There is no significant difference between the groups.

Describe Kruskal Wallis ANOVA . An experimenter is interested in examining the effectiveness of three methods of teaching. A group of 15 subjects were randomly divided into 3 groups . The scores are given below. Examine whether the three methods of teaching differ in terms of effectiveness or not.

Subject	Method 1	Method 2	Method 3
A	1	2	4
B	3	0	2
C	2	1	3
D	3	2	4
E	2	1	3

Critical value for χ^2 square @ 0.05 level of significance = 5.99

@ 0.01 level of significance = 9.21

Si no	X_A	R_A	X_B	R_B	X_C	R_C
1	1	3	2	7	4	14.5
2	3	11.5	0	1	2	7
3	2	7	1	3	3	11.5
4	3	11.5	2	7	4	14.5
5	2	7	1	3	3	11.5
		$\sum R_A$=40		$\sum R_B$=21		$\sum R_C$=59
	$(\sum R_A)^2$=1600		$(\sum R_B)^2$=441		$(\sum R_C)^2$=3481	

$$H = \frac{12}{N(N+1)} \Sigma \frac{R^2}{n} - 3(N+1)$$

$$= \frac{12}{15(15+1)} * \frac{1600+441+3481}{5} - 3(15+1)$$

$$= \frac{12}{15*16} * \frac{5522}{5} - 3*16$$

$$= 55.22 - 48$$

$$\underline{H = 7.22}$$

H calculated	>	χ^2 critical
7.22	>	5.99 (0.05 level)

Hence we reject Ho & Accept HA . The teaching methods differ in terms of effectiveness.

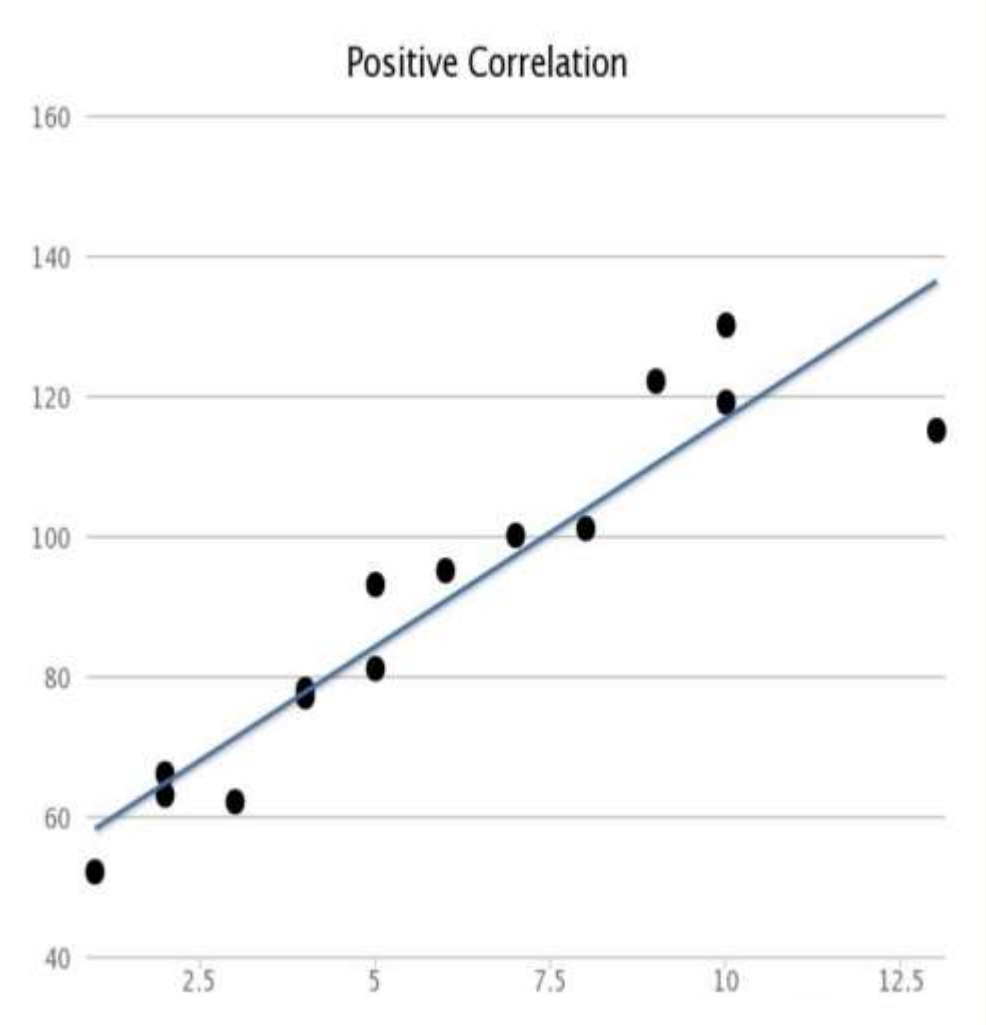
H calculated	<	χ^2 critical
7.22	<	9.21 (0.01 level)

Hence we Accept Ho . The teaching methods does not differ in terms of effectiveness.

CORRELATION

CORRELATION

Correlation **describes the relationship between variables. It can be described as either strong or weak, and as either positive or negative.**

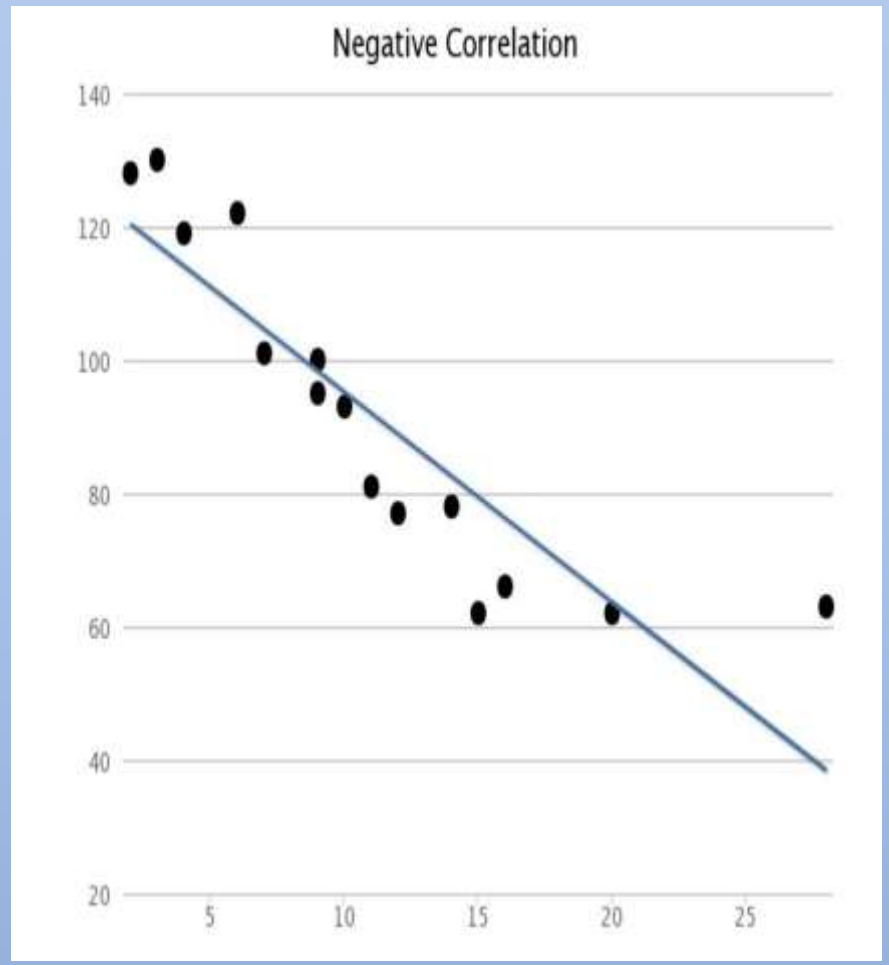


POSITIVE CORRELATION

THERE IS A *POSITIVE LINEAR CORRELATION* WHEN THE VARIABLE ON THE X-AXIS INCREASES AS THE VARIABLE ON THE Y-AXIS INCREASES. THIS IS SHOWN BY AN UPWARDS SLOPING STRAIGHT **REGRESSION LINE**.

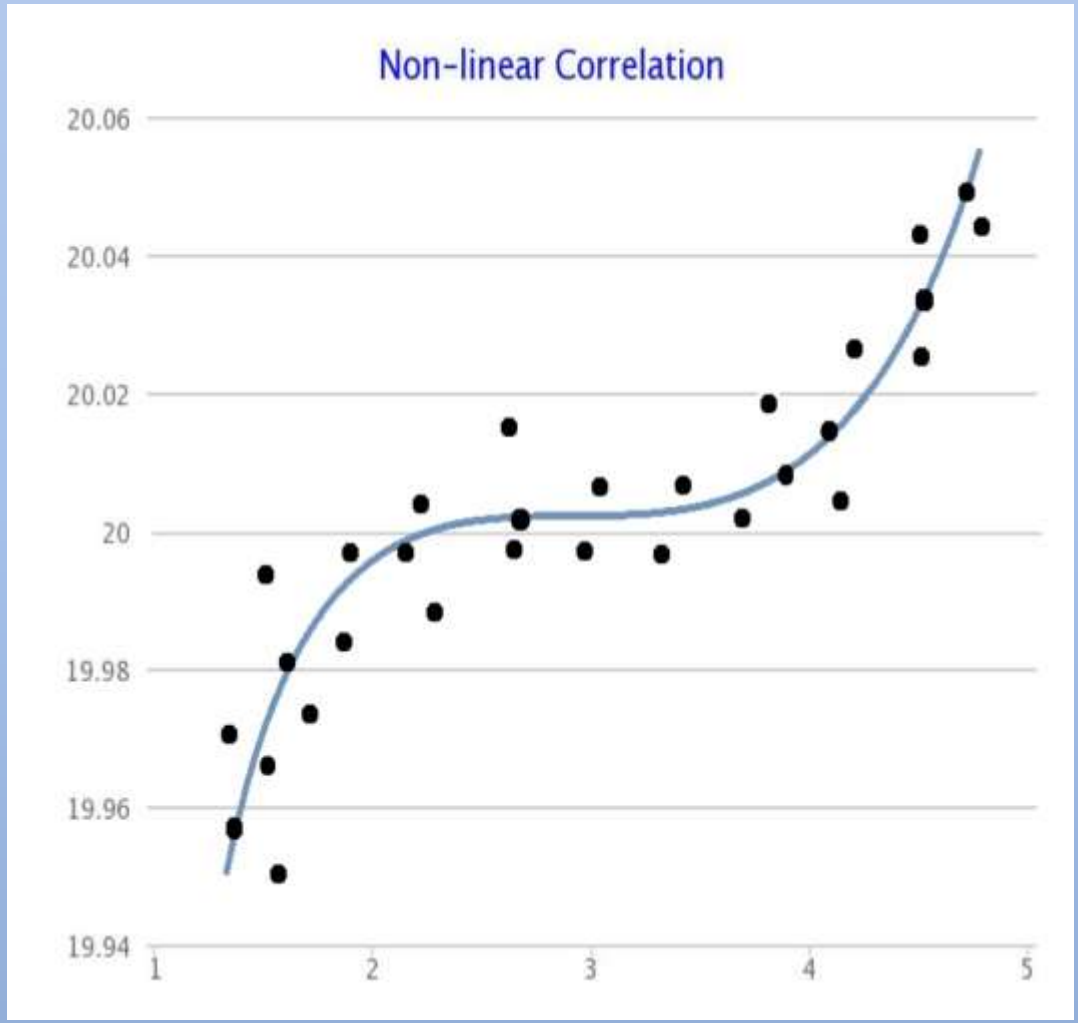
NEGATIVE CORRELATION

There is a *negative linear correlation* when one variable increases as the other variable decreases. This is shown by a downwards sloping straight regression line.



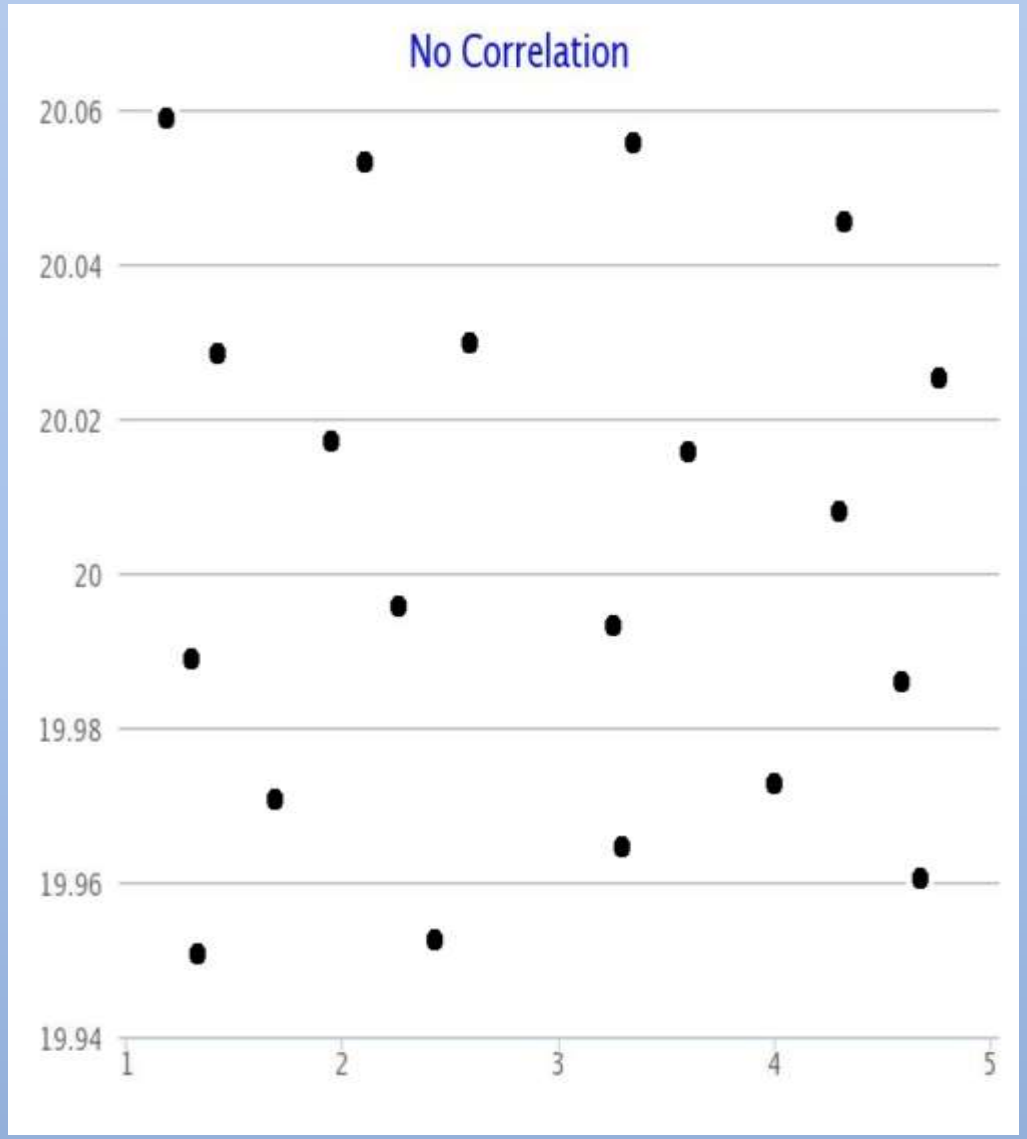
CURVILINEAR CORRELATION

There is a *non-linear correlation* when there is a relationship between variables but the relationship is not linear (straight).



NO CORRELATION

There is *no correlation* when there is no pattern that can be detected between the variables.

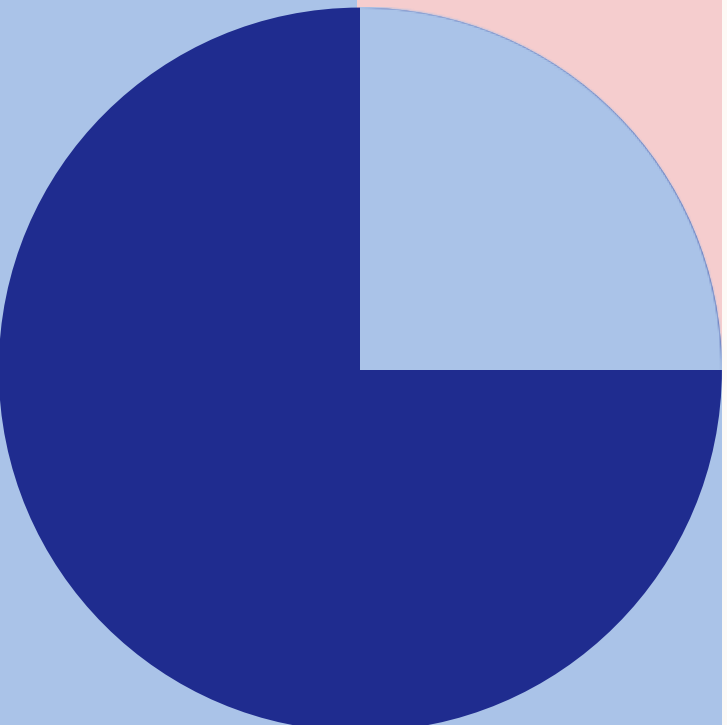


ZERO ORDER CORRELATION

a zero-order correlation simply refers to the correlation between two variables (i.e., the independent and dependent variable) without controlling for the influence of any other variables.

PARTIAL CORRELATION

a partial correlation is the correlation between an independent variable and a dependent variable after controlling for the influence of other variables on both the independent and dependent variable. For instance, a researcher studying occupational stress may be interested in the correlation between the length of time a person has worked with a company and their level of stress while controlling for one or more potentially confounding variables, such as age and pay rate. In a partial correlation, the influence of the control variables on both the independent and dependent variables are taken into account. In our example, this would mean that the partial correlation between time with the company and stress would take into account the impact of age and pay rate on both time with the company AND stress.

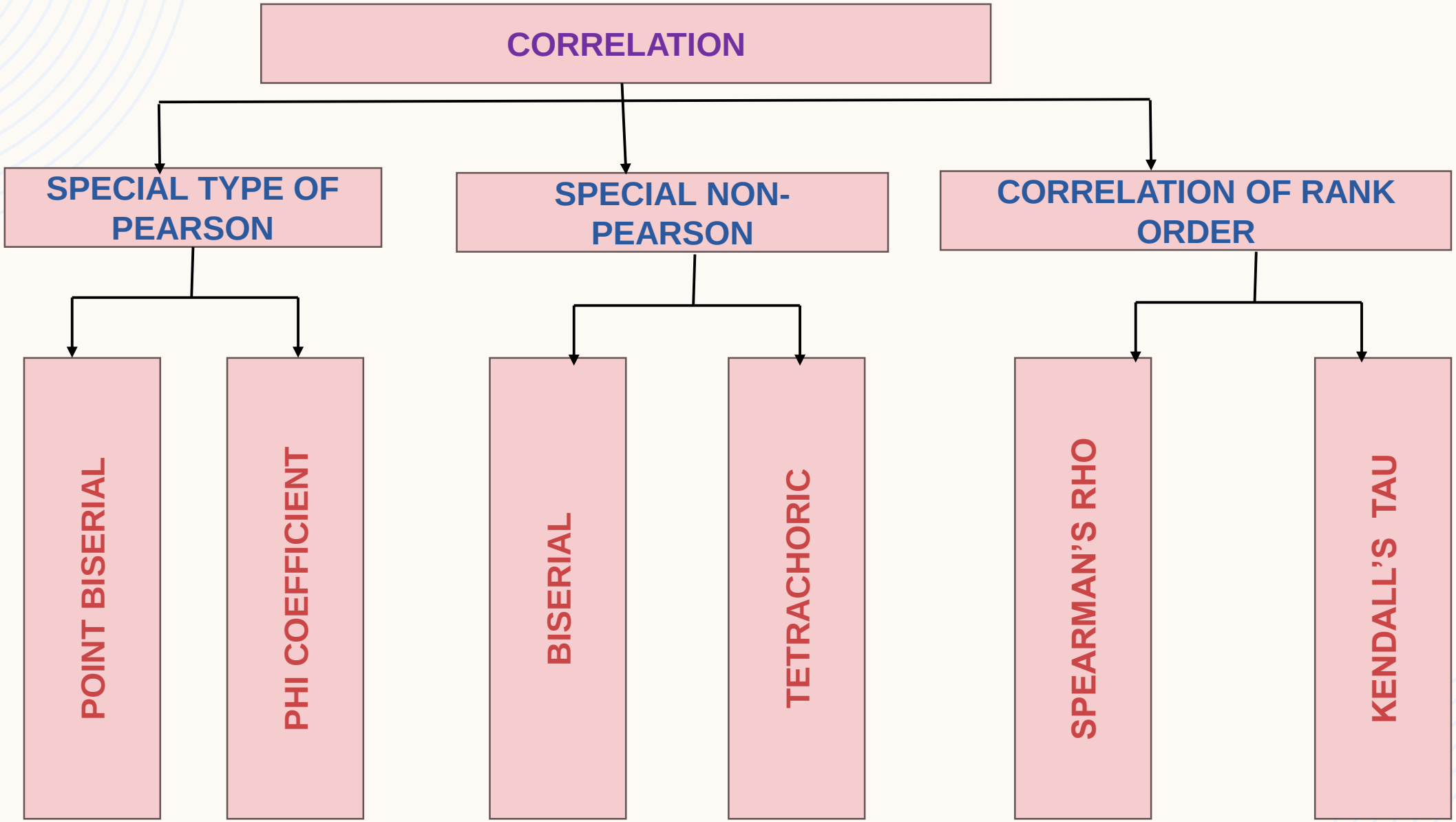


PART CORRELATION

Like the partial correlation, the part correlation is the correlation between two variables (independent and dependent) after controlling for one or more other variables. However, for the part correlation, only the influence of the control variables on the independent variable is taken into account. In other words, the part correlation does not control for the influence of the confounding variables on the dependent variable. In terms of our earlier example, this means that the part correlation between time with the company and stress would only take into account the impact of age and pay rate on time with the company.

MULTIPLE CORRELATION

The correlation is said to be Multiple when three variables are studied simultaneously. Such as, if we want to study the relationship between the yield of wheat per acre and the amount of fertilizers and rainfall used, then it is a problem of multiple correlations


CORRELATION
SPECIAL TYPE OF PEARSON
SPECIAL NON-PEARSON
CORRELATION OF RANK ORDER
POINT BISERIAL
PHI COEFFICIENT
BISERIAL
TETRACHORIC
SPEARMAN'S RHO
KENDALL'S TAU

POINT BISERIAL CORRELATION

The point biserial correlation coefficient (r_{pb}) is a correlation coefficient used when one variable (e.g. *Y*) is dichotomous; *Y* can either be "naturally" dichotomous, like whether a coin lands heads or tails, or an artificially dichotomized variable. In most situations it is not advisable to dichotomize variables artificially. When a new variable is artificially dichotomized the new dichotomous variable may be conceptualized as having an underlying continuity. If this is the case, a biserial correlation would be the more appropriate calculation.

The point-biserial correlation is mathematically equivalent to the Pearson (product moment) correlation coefficient; that is, if we have one continuously measured variable *X* and a dichotomous variable *Y*, $r_{XY} = r_{pb}$. This can be shown by assigning two distinct numerical values to the dichotomous variable.

PHI COEFFICIENT

A Pearson correlation coefficient estimated for two binary variables will return the phi coefficient.

A **Phi Coefficient** (sometimes called a *mean square contingency coefficient*) is a measure of the association between two binary variables.

For a given 2×2 table for two random variables x and y:

	y = 0	y = 1
x = 0	A	B
x = 1	C	D

The Phi Coefficient can be calculated as:

$$\Phi = (AD-BC) / \sqrt{(A+B)(C+D)(A+C)(B+D)}$$

Example: Calculating a Phi Coefficient

Suppose we want to know whether or not gender is associated with political party preference. We take a simple random sample of 25 voters and survey them on their political party preference. The following table shows the results of the survey:

	Dem	Rep
Male	4	9
Female	8	4

We can calculate the Phi Coefficient between the two variables as:

Φ = (4*4-9*8) / √(4+9)(8+4)(4+8)(9+4) = (16-72) / √24336 = **-0.3589**

BISERIAL CORRELATION

Biserial Correlation: In biserial correlation, the relationship is measured between a continuous variable and an artificially dichotomous variable.

Correlation: It is a measure of association between two or more variables and this relationship is determined not only in terms of direction, whether negative or positive.

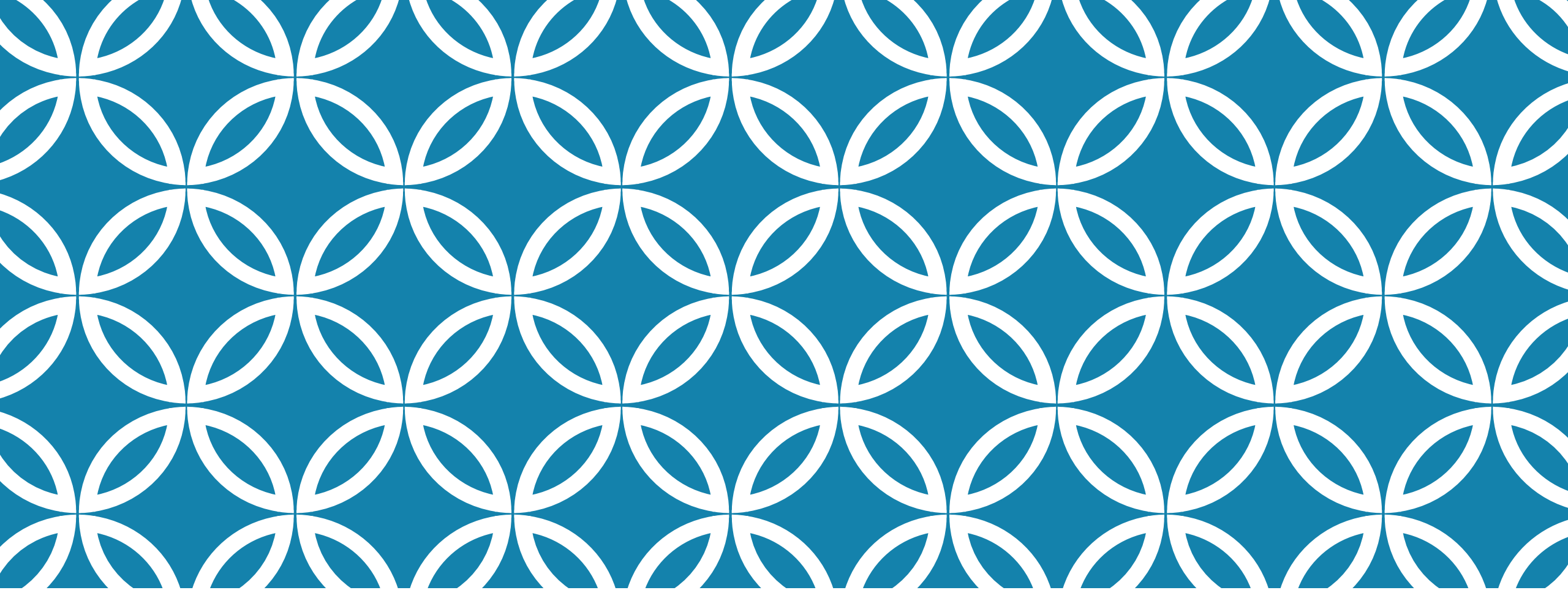
TETRACHORIC CORRELATION

Tetrachoric correlation is a measure of the correlation between two binary variables – that is, variables that can only take on two values like “yes” and “no” or “good” and “bad.”

This type of correlation is often used in surveys and personality tests in which the questions being asked only have two possible response values.

THANK YOU

USMITHA K S

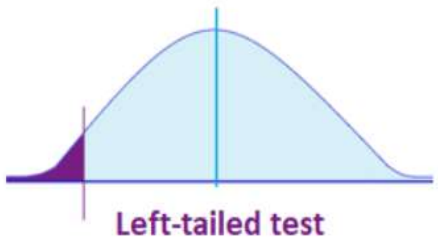


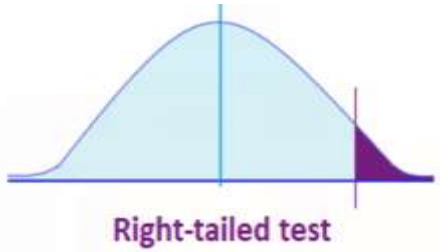
ONE TAILED AND TWO TAILED

Hypothesis testing

ONE TAILED VS TWO TAILED HYPOTHESIS TESTING

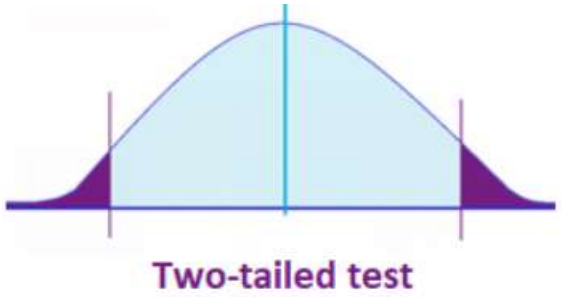
One tailed





A statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both.

Two tailed



A method of calculating statistical significance in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

ONE TAILED VS TWO TAILED HYPOTHESIS TESTING

One tailed

- Region of rejection is on one side of the sampling distribution
- Directional test
- More powerful
- Tells you the effect of a change in a direction

Two tailed

- Region of rejection is on both side of the sampling distribution
- Non directional
- Less powerful
- Tells the effect of change in both direction

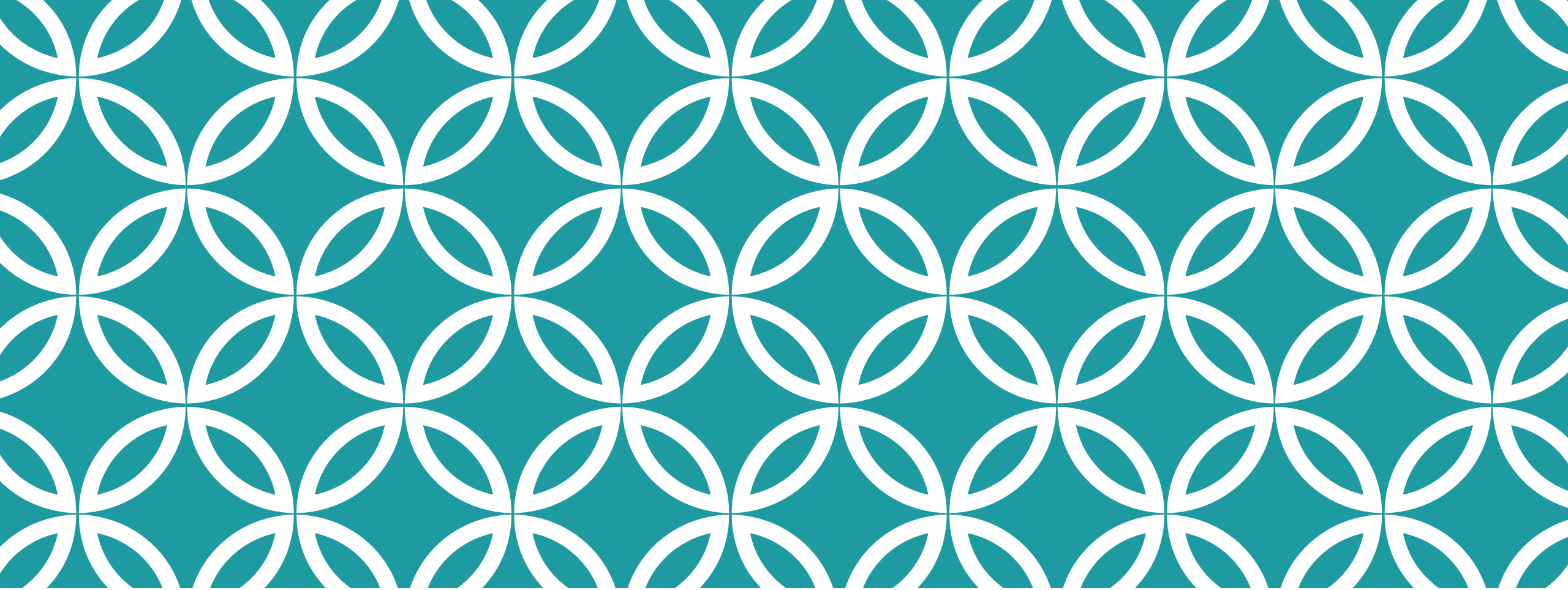
ONE TAILED VS TWO TAILED HYPOTHESIS TESTING

One tailed

- Looks for an increase or decrease in the parameter
- Two outcomes – 'B' is better than 'A' or it isn't
- Can lead to biased result

Two tailed

- looks for any change in the parameter (both increase & decrease)
- Three outcomes – 'B' is better than 'A'
  - 'B' is same as 'A'
  - 'B' is worse than 'A'
- Lead to accurate result



HYPOTHESIS AND TYPES

HYPOTHESIS

It is a claim/statement/ assumption about one or more population parameters. It is referred to as the statistical decision making process.

TYPES OF HYPOTHESIS

Null Hypothesis :- Ho – means insignificant or 'No' relationship b/w variables.

Alternate Hypothesis :- HA- rejection of null hypothesis or 'there is relationship b/w the variables'

Note : “ Alternative hypothesis will always be the counter part or complement of Null hypothesis”.

When one of the hypothesis become false the other must be true .

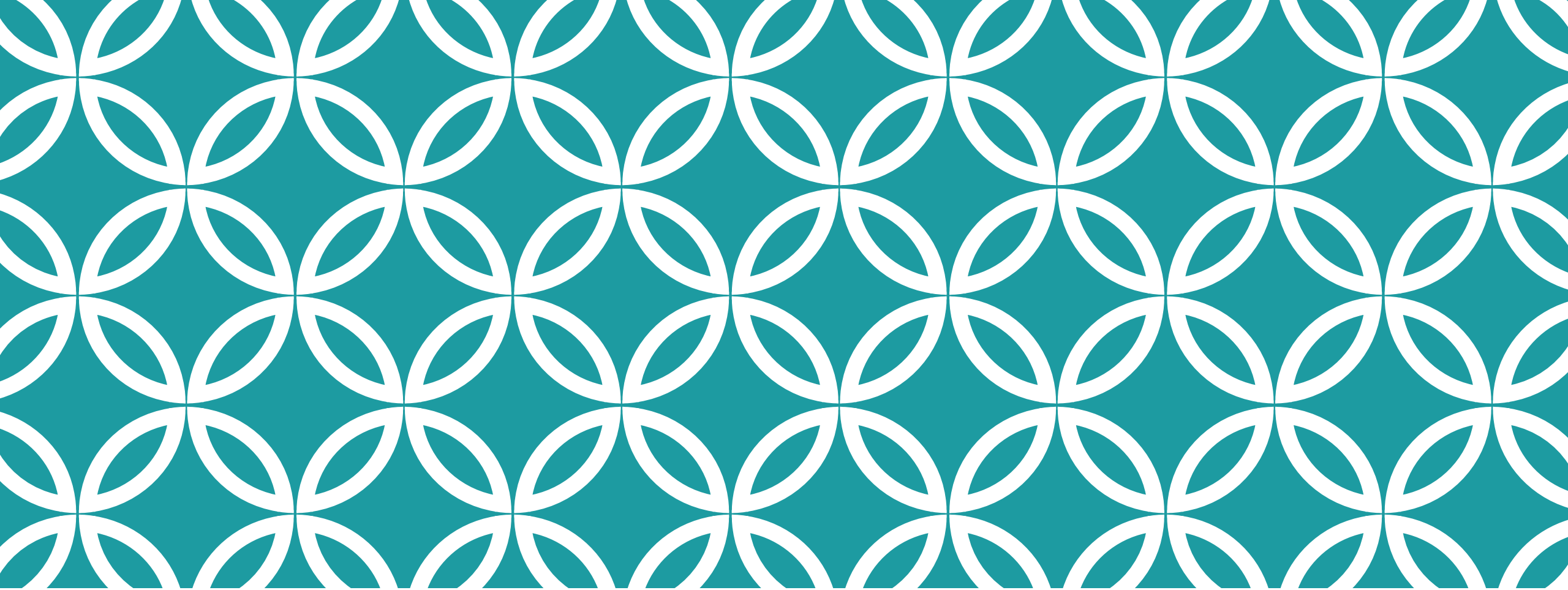
NULL VS ALTERNATIVE HYPOTHESIS

Null Hypothesis

- A statement which usually claims a zero difference
- Statement about the value of a population parameter
- Directly tested statistically
- Hypothesis researcher wish to reject
- Always stated as equality/no relation/no significant difference etc.

Alternative hypothesis

- Statement usually postulates a non-zero difference or relationship
- Statement about the value of a population parameter that must be true if Ho is false
- Not directly tested , accept or rejected based on Ho
- The hypothesis which researcher support . Known as research hypothesis
- Stated as inequality/relationship/existence of significant difference etc.



MERITS AND DEMERITS OF TWO WAY ANOVA

MERITS AND DEMERITS OF TWO WAY ANOVA

MERITS

- More efficient than its one way ANOVA counter part as there is reduced error variation
- Can test the effect of two factors at the same time
- Test for independence of the factors is possible, provided there are more than one observation in each cell
- Usually have a smaller total sample size
- Removes some of the random variability
- We can look at interactions between factors
- Can look at effect on second variable after controlling the first variable

DEMERITS

- If the assumptions are not fulfilled, it may provide us spurious result
- Difficult and time consuming
- As the number of factors are increased in a study, the complexity of analysis is increased and interpretation of the result become difficult
- Requires high level arithmetical and calculative ability
- Requires high level of imagination and logical ability to interpret the result
- For test of independence, the number of observations in each cell has to be equal.