

## ① Nominal data and ordinal data.

⇒ Nominal data are categories that don't have any inherent order or ranking.

Examples includes colours, gender categories, types of animals, or cities. They represent distinct and separate groups, but there's no way to quantify the differences between them in meaningful manner.

⇒ Ordinal data are categories that have a specific order or ranking, but the intervals between the categories are not necessarily equal or meaningful. Examples include ratings (eg. 1 star, 2 stars, 3 stars) or educational levels (eg. high school, bachelor's degree, master's degree). While you can determine which category is higher or lower, you can't say by how much one category is greater than another.



## 2. Scales of measurement.

⇒ There are 4 commonly recognized scales of measurement.

i) Nominal scale.

ii) Ordinal scale

iii) Interval scale

iv) Ratio scale.



### 3) STANDARD ERROR OF MEAN (SEM)

It is a measure, that quantifies the variability (or) uncertainty of sample means around the true population mean.

It is calculated by dividing the standard deviation of the sample by the ~~sq~~ square root of the sample size.

$$SEM = \frac{\sigma}{\sqrt{n}}$$

$\sigma$  = Standard deviation  
 $n$  = no. of samples

### STANDARD DEVIATION

It is a measure that quantifies the amount of variability or dispersion in a set of data points.

A higher standard deviation indicates greater variability, while a lower standard deviation suggests that the data points are closer to the mean.

It is calculated by square root of variance.

Formula

The "Population" standard deviation  $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$

The "Sample" standard deviation  $s = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N-1}}$

where,  $N$  = no. of ~~data points~~ total sample

$x_i$  = ~~value of each data~~ Each of the values of data

$\mu$  = mean



#### 4) DESCRIPTIVE STATISTICS:

Descriptive statistics are used to summarize and describe the main features of data set.

This includes measures like mean, median, mode, standard deviation, range, which help provide a clear overview.

#### INFERENTIAL STATISTICS:

It is the branch of statistics that involves using sample data to make inferences or predictions about a large population.

This technique is used to extrapolate information from a smaller sample to make predictions and draw conclusions about a larger population.

Engl



## 5) Types of data distribution.

→ The Statistical data can be divided into two broad Categories:

1. Quantitative
2. Qualitative.

### 1. Quantitative data (numerical data)

- a) discrete data (whole number data)
- b) Continuous data (data in fraction)

### 2. Qualitative data (categorical data)

- a) Nominal data
- b) Ordinal data
- c) Categorical data
- d) Binary data
- e) ternary data

## 6) Quartile and interquartile range.

→ Quartile divide a dataset into four equal parts, and the interquartile range (IQR) measures the spread of the middle 50% of the data.

First Quartile ( $Q_1$ ) = 25% of data

Second Quartile ( $Q_2$ ) = 50% of data

Third Quartile ( $Q_3$ ) = 75% of data

$IQR = Q_3 - Q_1$ , It provides a measure of how spread out the middle 50% of the data is.



~~Q~~ Why mean is most preferred central tendency?

Ans) mean is widely preferred as the best measure of central tendency because it is the measure that includes all the values in the data set for its calculation and any change in any of the scores will affect the value of the mean.

• It can be used with both discrete and continuous data, although its use is most often with continuous data.



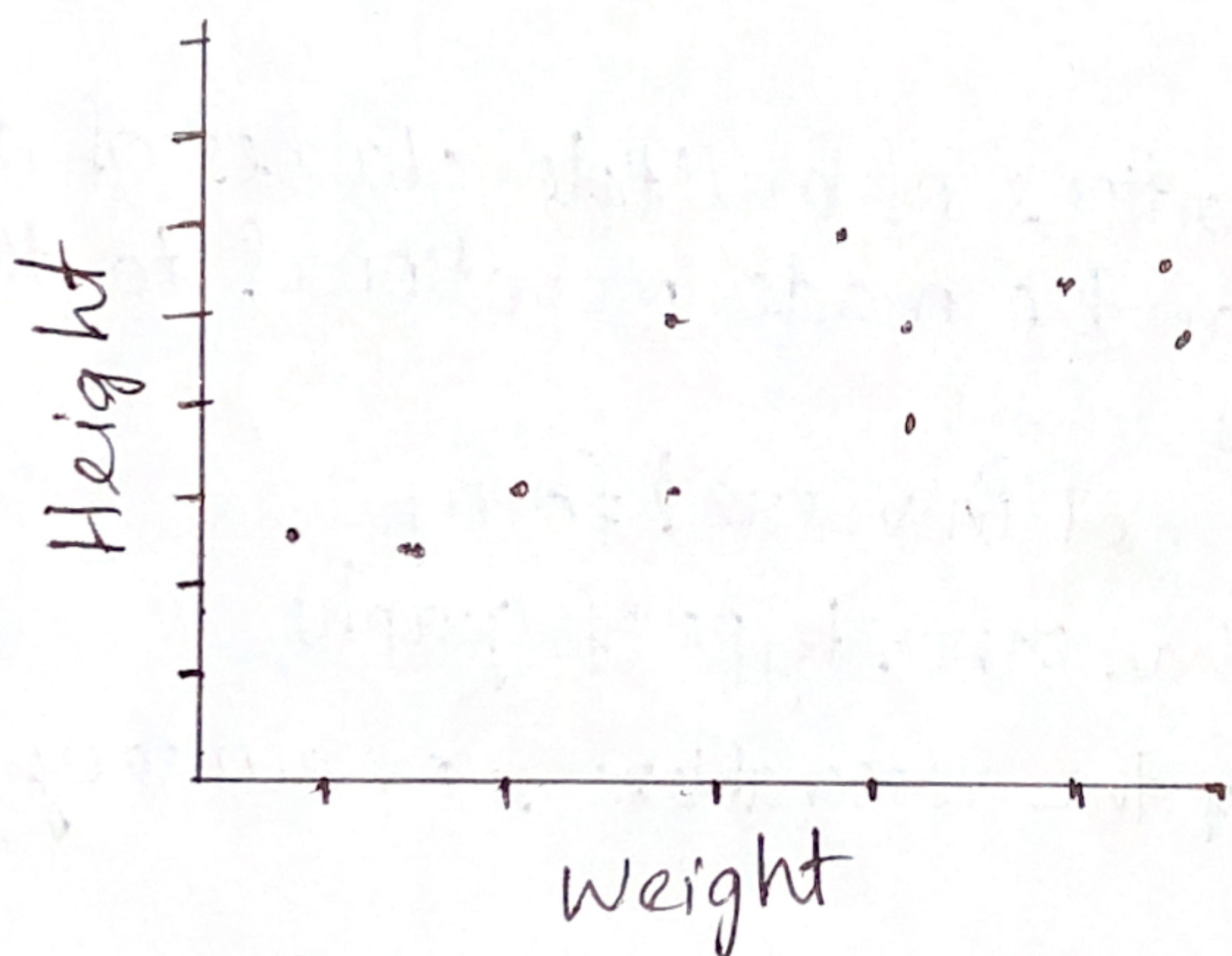
Qans

## Scatter plots

Scatter plots also called correlation diagrams or scatter diagrams at this time. This type of plot or diagram is commonly used when presenting results of experiments.

A scatter diagram is a graphic representation of the relationship between two variables.

Scatter diagrams help teams identify and understand cause effect relationship





9ans

## PIE CHART

A pie chart is a two dimensional circle, at its most basic divided into a few slices. as a whole the chart represents the sum of all its data. Individual slices represents a percentage of the whole

### Advantages

- Display relative proportions of multiple classes of data
- size of the circle can be made proportional to the total quantity it represents
- summarize a large set in visual form
- Be visually simpler than other type of graphs
- Permit a visual check of the reasonableness or accuracy of calculations
- A pie chart presents data as a simple and easy to understand picture

### Disadvantage

- Do not easily reveal exact value
- Many pie charts may be needed to show changes over time
- Fail to reveal key assumptions, cause, effects or patterns
- Be easily manipulated to yield false impressions
- A pie chart becomes less effective if it uses too many pieces of data. For example a chart with more than 10 becomes less so, especially if it contains many similarly sized slices



## 10. Semi logarithmic plots

- \* It is a graphical representation of a time series.
- \* It is defined by an arithmetic scale of time  $t$  plotted on the abscissa axis and a logarithmic scale plotted on the ordinate axis, meaning that the logarithm of observed value  $Y_t$  will be transcribed in ordinate on an arithmetic scale.
- \* In general, printed semi logarithmic sheets of paper are used.
- \* In a semi logarithmic graph, one axis has logarithmic scale and other axis has linear scale.

## 11. Bivariate distribution

- \* A bivariate distribution is a statistical method used to analyse the relationship between two random variables.
- \* It can be represented as table, graph or function.
- \* A bivariate distribution is often displayed as a table.
- \* The outcomes for variable 1 are listed in top row, and outcomes for variable 2 are listed in first column. The probabilities for each set of outcomes are listed in individual cells.
- \* The last row and column contains the marginal probability distribution.
- \* The bivariate distribution is important for determining risks and probabilities in many situations.
- \* It is frequently used in fields such as insurance, health care and sales.
- \* An example of a bivariate normal 'distribution' would be rolling two fair dice.



## 12. POWER OF STUDY:

Ans:

- Power is the probability that the statistical test results in rejection of  $H_0$  (null hypothesis) when specified alternative is true.
- The stronger the power, the better the chance that the null hypothesis will be rejected, when in fact  $H_0$  is false.
- The larger the power, the more sensitive is the test.
- Power is defined as  $1 - \beta$ .
- The power of a test should be calculated before an experiment is conducted. and power is used to help compute the sample size.
- There is no traditional values of  $\beta$  to use, as in case for  $\alpha$ , where 5% is usually used.
- Larger the sample size, larger the power.
- Larger sample size will be needed in order to have strong power to detect small differences.
- A simple way to compute the approximate power of a test is to use the formula for sample size.

$$N = \left( \frac{\sigma}{\Delta} \right)^2 (Z_\alpha + Z_\beta)^2.$$

$$\therefore \text{Power of study} = Z_\beta = \left( \frac{\Delta}{\sigma} \right) (\sqrt{N} - Z_\alpha).$$



13. UNPAIRED TWO SAMPLE 't-test' :

Ans: Unpaired t-test is used when we want to compare the 2 means, and used when 2 independent random samples come from normal populations, having unknown (or) same variable.

Testing the null hypothesis that is, 2 population means are same i.e.  $\mu_1 = \mu_2$  against an appropriate 1 sided / 2 side alternative hypothesis.

Assumptions:

- The samples are Random and Independent of each other.
- The distribution of dependent, variable is normal.
- The variance are Equal in both the groups.

Formula: Test statistic is given by,

$$t = \frac{\text{Mean}_1 - \text{Mean}_2}{\text{SE}(\text{Mean}_1 - \text{Mean}_2)}$$

$$\text{SE}[\text{Mean}_1 - \text{Mean}_2] = S \sqrt{\left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}$$

$$\therefore S = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

Where,  $S_1^2$  and  $S_2^2$  are respectively called  $SD^2$  of 1<sup>st</sup> and 2<sup>nd</sup> gp.

PAIRED - TWO SAMPLES 't-test' :

These are used, when paired data of observation from one 'sample' only, when each individual gives a pair of observations.

Same individuals are studied more than once in different - circumstances, Measurements made on, sample people before and after interventions.



### Assumptions:

The outcome variable should be continuous the difference between pre-post measurements should be Normally distribute.

### Formula:

$$t = \frac{\bar{d}}{\frac{SD}{\sqrt{n}}}$$

$d \rightarrow$  difference, bet<sup>n</sup>  $x_1$  and  $x_2$ .

$\bar{d} \rightarrow$  Average of  $d$ .

S.D  $\rightarrow$  Std. Deviation for the difference

$n \rightarrow$  sample size.



14, Why ANOVA is preferred over student-t-test :-

- 1) ANOVA is an observable technique used to compare the means of more than two population groups.
- 2) ANOVA equates three or more such groups.
- 3) ANOVA is one-sided test due to no negative variance.
- 4) ANOVA is used for huge population counts.



## 15) One way Anova

A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution. The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are equal.

### \* Limitations of the one way Anova

A one way Anova will tell you that at least two groups were different from each other. But it won't tell you which groups were different. If your test returns a significant F-Statistic, you may need to run an ad hoc test (like the least significant difference test) to tell you exactly which groups had a difference in means.



16. Wilcoxon rank sum test :-

This is the non-parametric test to compare the value in 2 independent groups. This test is used to determine whether two independent samples have drawn from the same population.

- > Assigns the rank one to the lowest value, 2 to next 3, so on in ascending order by considering two groups of values together.
- > Find the sum of the ranks of values in the first group.
- > Find the sum of the ranks of values in the second group.
- > Calculate  $U_1$  and  $U_2$  as -

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

where,  $U$  = Wilcoxon rank sum  
 $W$  = Rank sum of group  
 $n$  = Sample size of group

If the calculated value of  $U$  is less than the tabulated value of  $U$  at  $n_1, n_2$  degree of freedom, we reject the null hypothesis and conclude that two samples are not from the same population.



## ■ Wilcoxon rank signed test :-

In this method we have to find the difference between each pair of values and assign rank to differences from the smallest to the largest without regard sign. The positive or negative signs of each differences are then put to corresponding ranks and the test statistics  $T$  is computed, which happens to be the smaller of the two sums viz., the sum of the negative ranks and sum of the positive ranks.

Then the standard deviation ( $\sigma$ ) is given by

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

where,  $n = (\text{No. of matched pairs} - \text{No. of dropped out pairs})$

Now use test statistics  $Z$  as

$$Z = \frac{T - V_T}{\sigma_T}$$

$$\text{where, } V_T = \frac{n(n+1)}{4}$$



## Linear Regression & Correlation

Regression analysis is used in Stats to find trends in data. For example, you might guess that there's a connection between how much you eat and how you weigh. Regression analysis can help you quantify that. Regression analysis will provide you with an equation for a graph so that you can make predictions about the data. For example, if you've been putting on weight over the last few years, it can predict how much you'll weigh in ten years time if you continue to put on weight at the same rate.

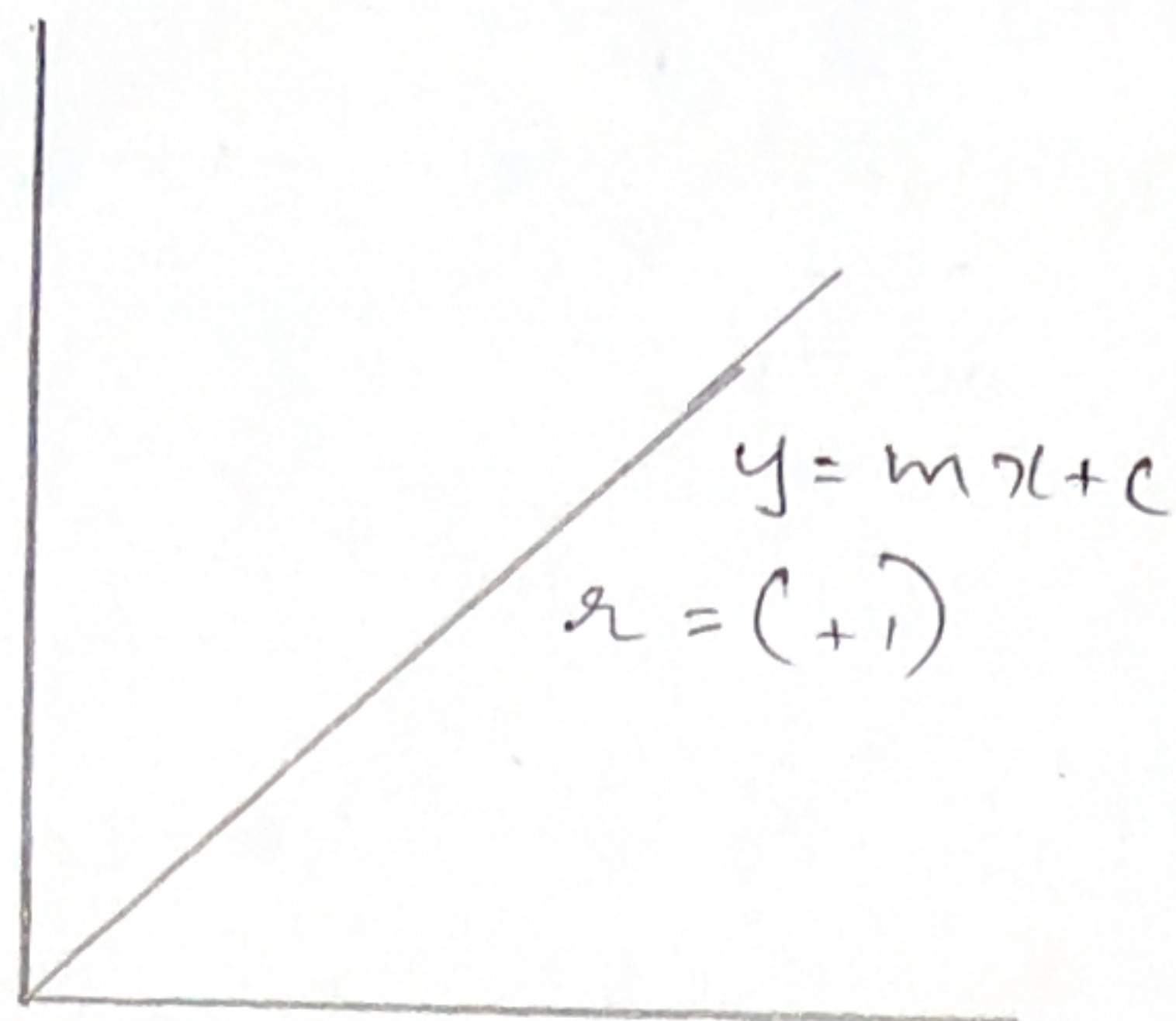


It will also give you a slew of statistics (including a p-value and a correlation co-efficient) to tell you how accurate your model is. Most elementary Stats courses cover very basic techniques, like making Scatter plots and performing linear regression. However, you may come across more advanced techniques like multiple regression.

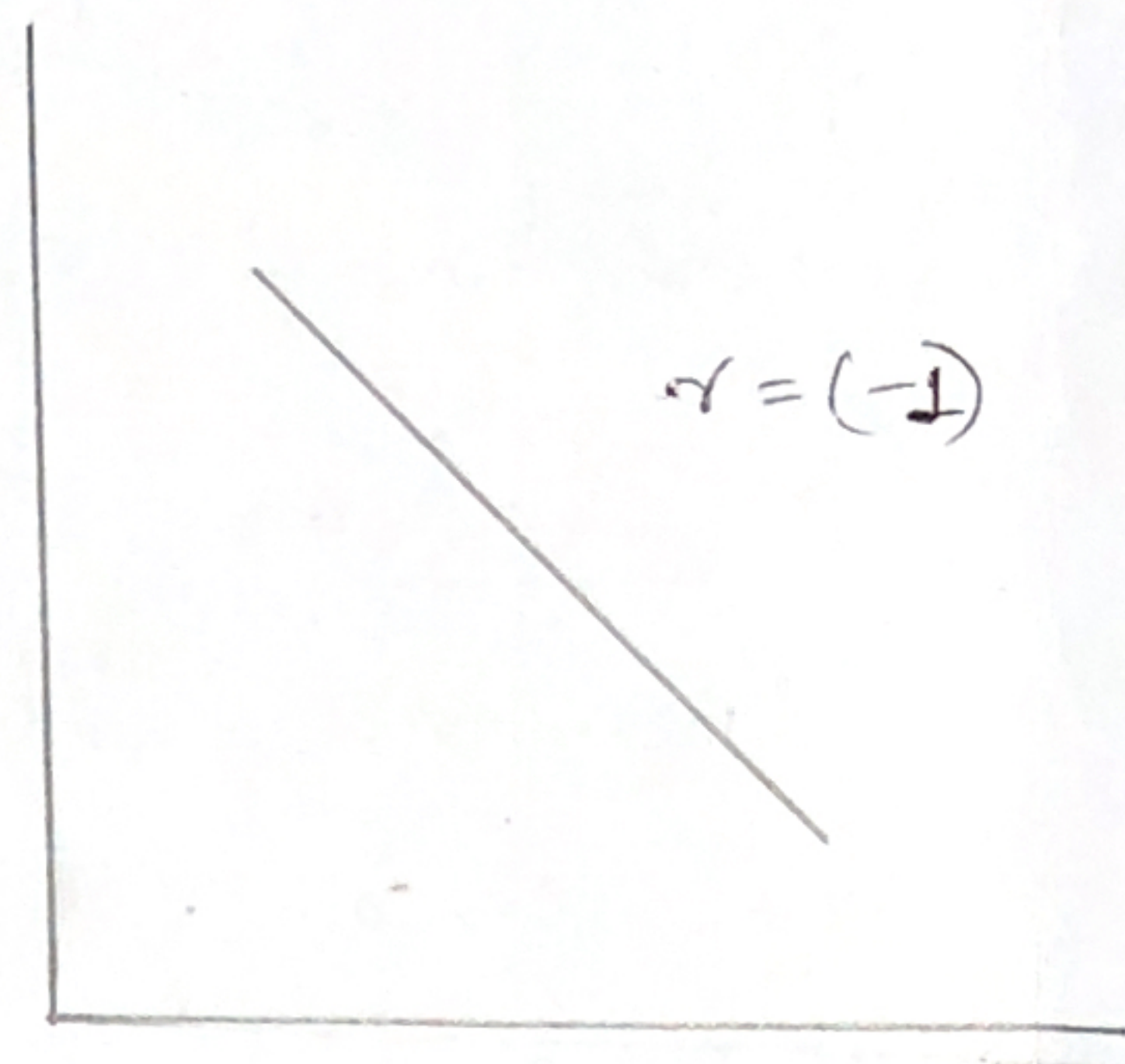


# 18 R values of Regression

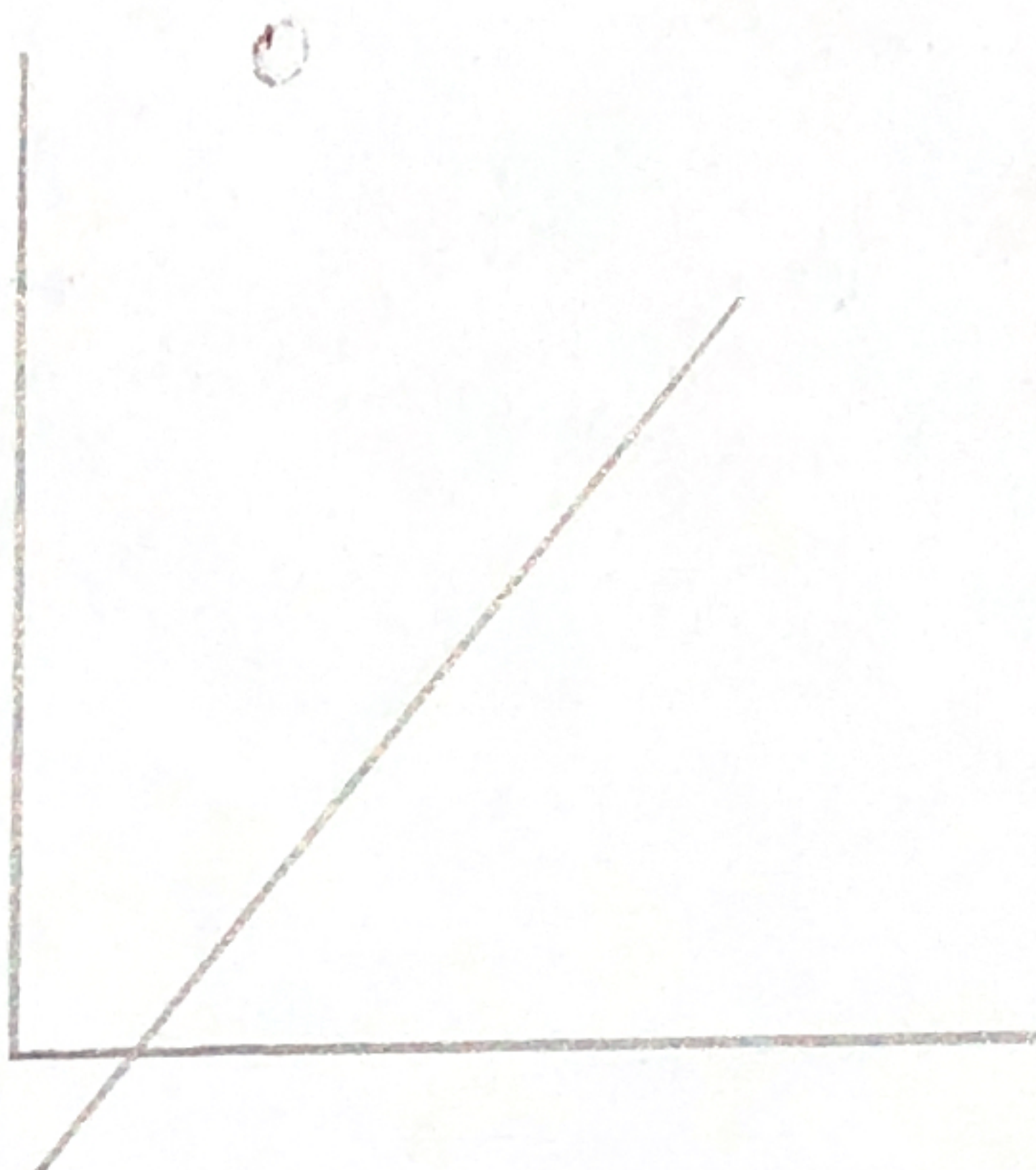
- Positive Co-relation



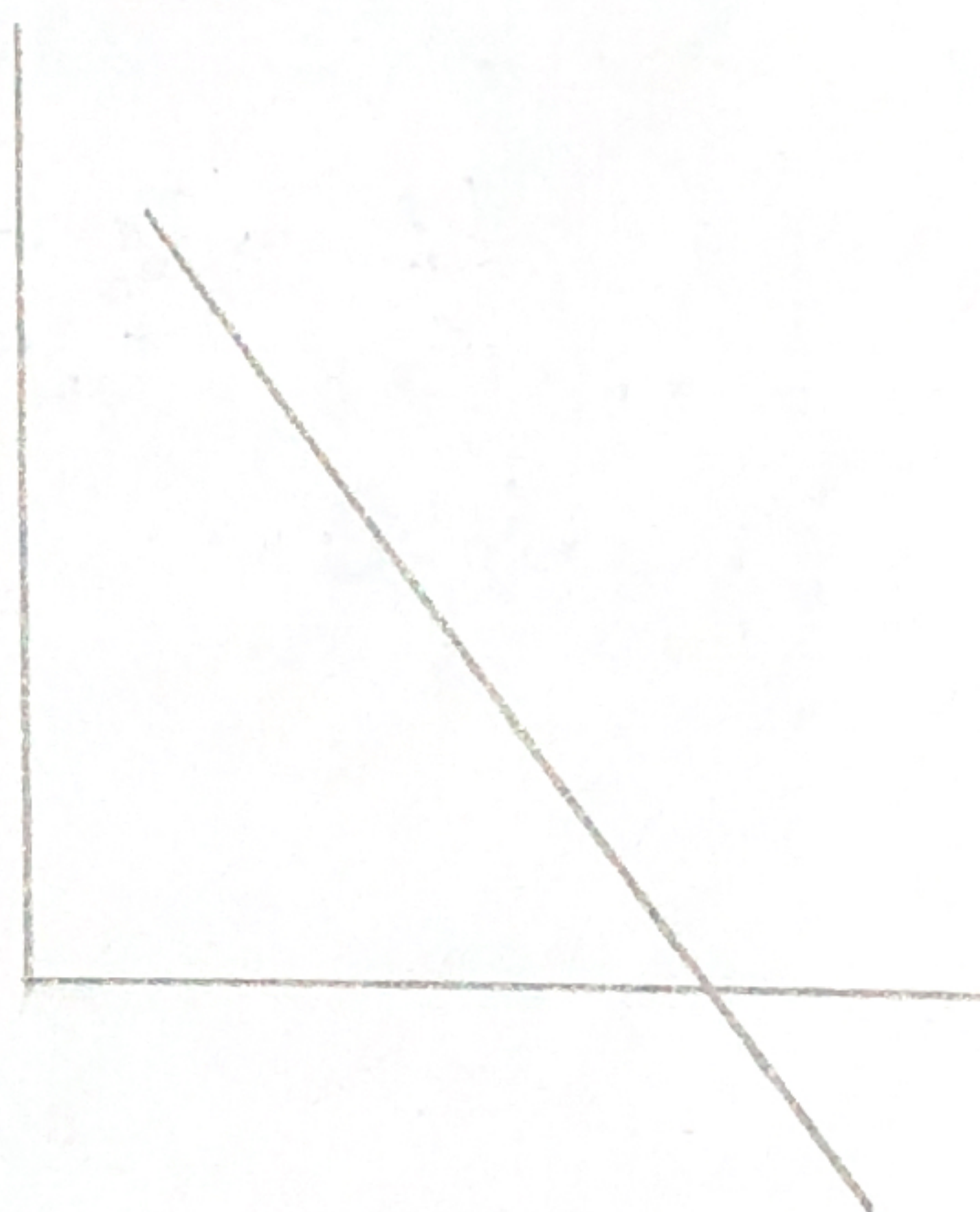
- Negative Correlation



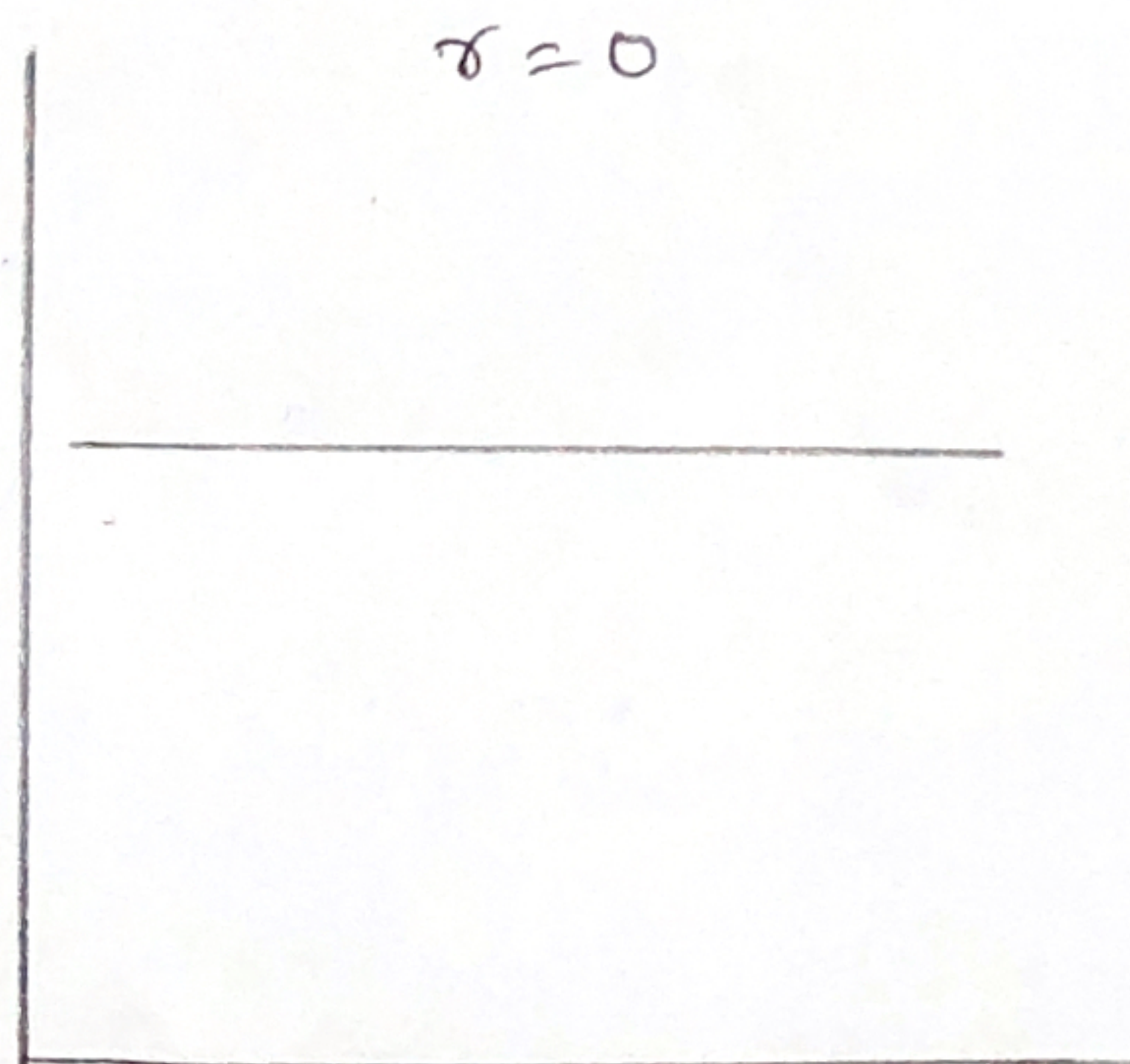
- Moderately Positive



- Moderately negative



- No co-relation





(19)

## Regression ~~correlation~~: Analysis

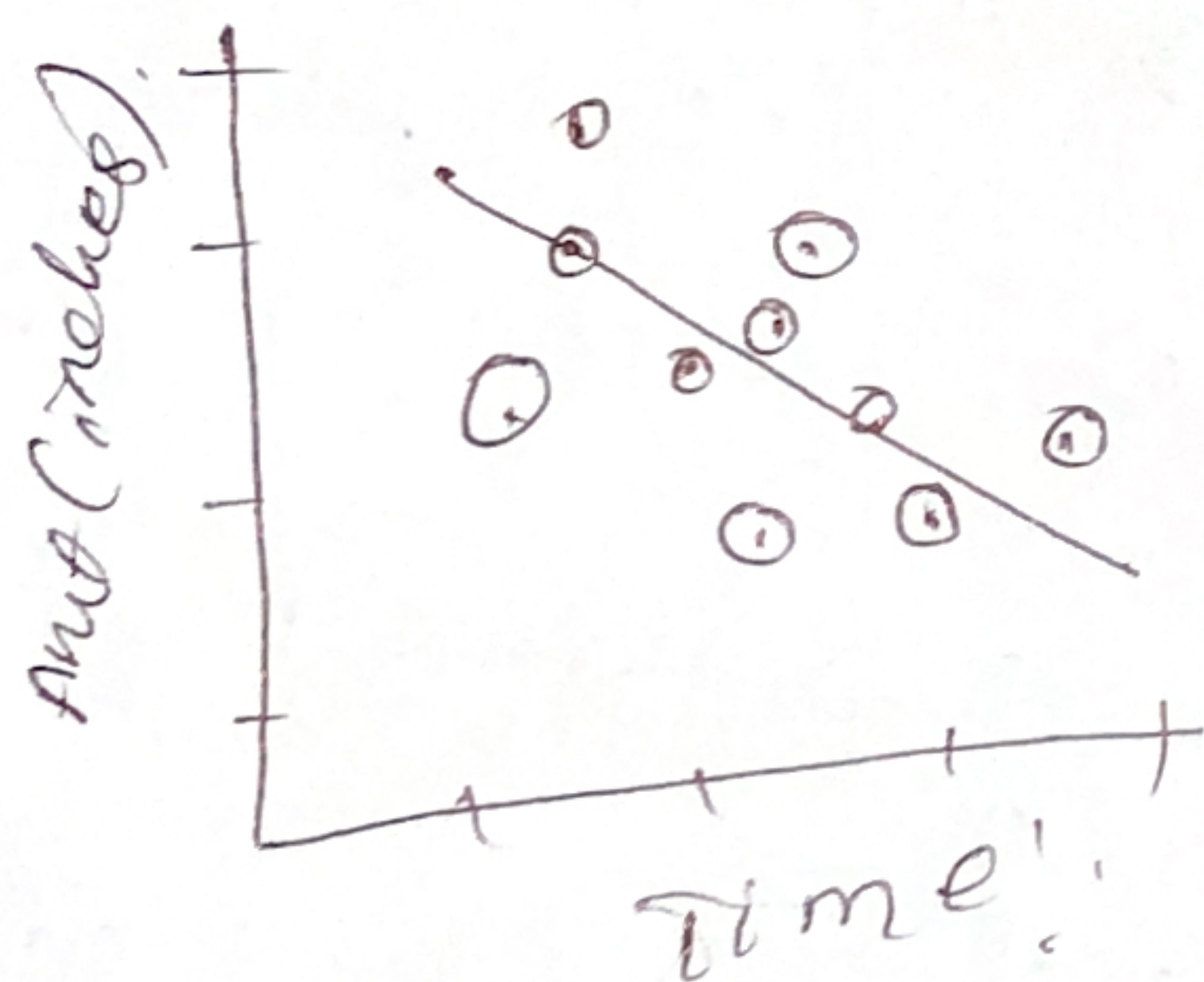
→ It is most commonly used techniques for investigating the relationship bet<sup>n</sup> two quantitative variables are correlation and linear regression.

Correlation quantifies the strength of the linear relationship between a pair of variables whereas the relationship in the form of equation.

Eg. Regression is used to find trends in data:

for e.g. connection bet<sup>n</sup> how much you eat and how much you weigh.

It provide eq<sup>n</sup> for a graph to make predictions about data.





## 20) CORRELATION COEFFICIENT :-

- Denoted by ' $r$ ', is a measure of the strength of the straight-line or linear relationship between two variables. The Correlation coefficient takes on values ranging between  $+1$  &  $-1$

- The quantity  $r$ , called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables

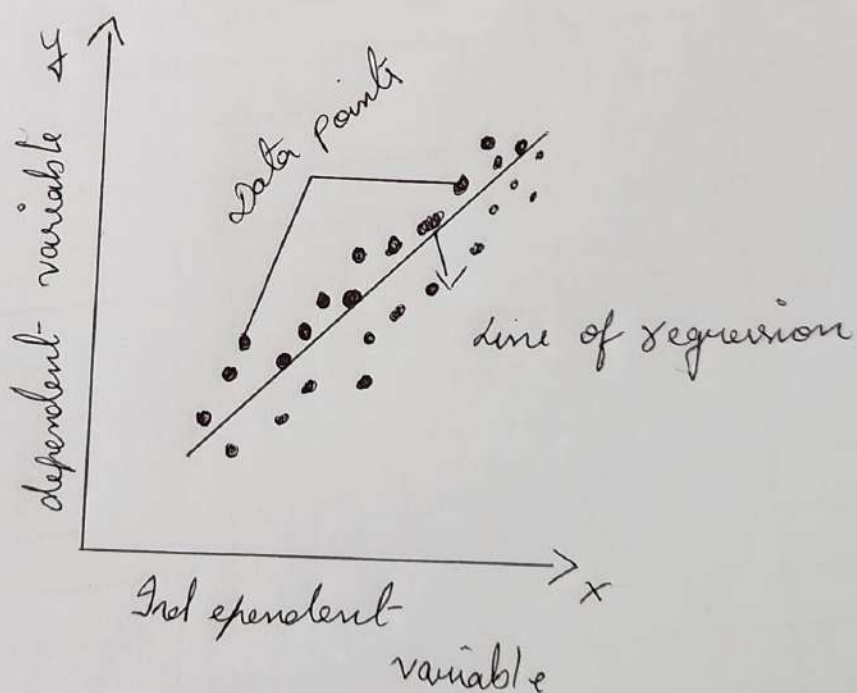
### TYPES OF CORRELATION COEFFICIENT :-

- Perfect Positive Correlation
- Perfect negative Correlation
- Moderately Positive Correlation
- Moderate negative Correlation
- Absolute no correlation



## → LINEAR REGRESSION: -

- Regression analysis is used in stats to find trends in data.
- Linear regression used to predict the value of a variable based on the value of another variable.
- The variable you want to predict is called the dependent variable.
- The variable you are using to predict the other variable's value is called the independent variable.





2021 What is linear regression and correlation?

Ans Linear Regression

- Regression analysis is used in statistics to



find trends in data - For example, you might guess that there is a connection between how much you eat and how much you weigh.

- Regression always provide you with a equation for a graph so that you can make predictions about your data.

### Correlation

- Two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

- The correlation is a statistical tool which studies the relationship between 2 variables.

- Correlation always involves various methods and techniques used for studying and measuring the extend of relationship b/w 2 variables.



## 23) Epidemiological markers. ~~sig~~

A) These are biological markers are used to characterize microorganisms or discriminate b/w genomes based on the genetic variation among microbial isolates.

### Types of markers

#### 1. Phenotypic markers -

It is the composite of the organism's observable characteristics or traits, including its morphology or physical form & structure, its developmental processes, its biochemical & physiological properties, its behaviour & the products of behaviour.

eg - Biotyping, Antimicrobial testing, serotyping.

#### 2. Genetic markers &

These are variants in the DNA sequence that can be typed directly.

### uses

1. Classify isolates of microorganism

2. Catalogue genetic variation

3. Define relatedness of but of it b/w microbial species or genera

4. It is applied in infection control too.

## 24) Direct endpoint & Surrogate endpoint.

A) A <sup>direct</sup> ~~clinical~~ endpoint is an outcome that represents direct clinical benefit, such as survival, decreased pain, or the absence of disease.

eg - clinical endpoint in cancer research is overall survival.

Surrogate endpoints do not represent direct clinical benefit, but

instead predict clinical benefit

eg - Tumour shrinkage can be used as a surrogate endpoint for longer survival in clinical trials for drugs intended

to treat some cancer



(25) 1. Statistical software used for decision making.

SAS - is a statistical software suite developed by SAS Institute for the data management, advanced analytics, multivariate analysis, business intelligence & emotional investigation & predictive analytics.

- In Business Intelligence - refers to strategies & technologies used by any enterprise for data analysis of business information.
- It provides insight regarding predictive, current and historical views of business working.
- The analysis of data helps the senior board with scope for decision making.
- SAS is developed by SAS Institute & release - 1976.

(26) Statistical software used for questionnaires & feedback survey

Epi info is statistical software for epidemiology developed by centres for disease control & prevention (CDC) in Atlanta, Georgia.

Epi info is currently available for Microsoft, Windows, Android & iOS along with a web & cloud version.

Epi info has been in development over 20 yrs.

- Uses:-
- ability to rapidly develop a questionnaire
  - Customize the data entry process
  - Quickly enter data into the questionnaire.
  - Analyze the data
  - Epidemiological uses.



②6) statistical software used for decision making:

→ SAS is a statistical software suite developed by SAS institute which is used for decision making. and used for

- data management
- advanced analytics
- multivariate analysis
- business intelligence
- criminal investigation
- predictive analytics

→ SAS provides a graphical point and click user interface for non-technical user and more through SAS language.

→ It has 3 Data steps ,

- ① Retrieve the data
- ② manipulate data
- ③ Analyze the data.

②7) statistical software used for Questioner and feedback survey:

→ Epi info is the statistical software used for Questioner and feedback survey.

→ It is developed by centers for Disease control and prevention (CDC) in Atlanta Georgia (US).

→ It is currently available for microsoft, windows, Android and ios along with a web and cloud version.

→ This program allows for

- \* electronic survey creation
- \* Data entry
- \* Data analysis
- \* Stratification
- \* Survival analysis etc.