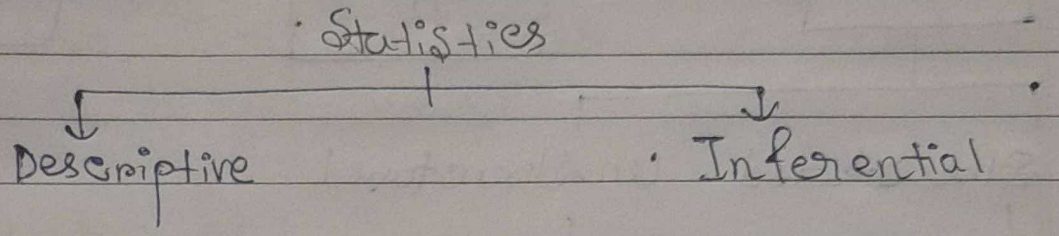
Statistics

1. Defination: Statistics is a branch of applied maths that involves the collection, description, analysis, drawing and inference of conclusions from ~~quantitive~~ quantitative data.

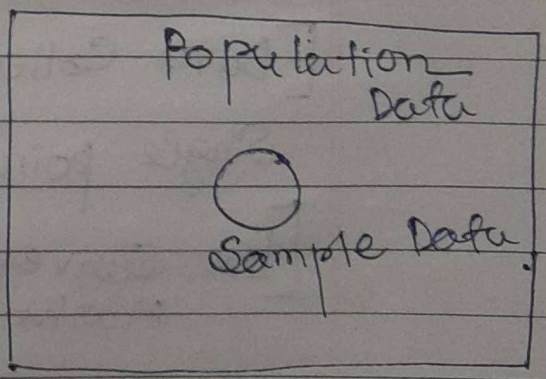


(i) Descriptive Statistics: All about summarizing and describing the data you have.

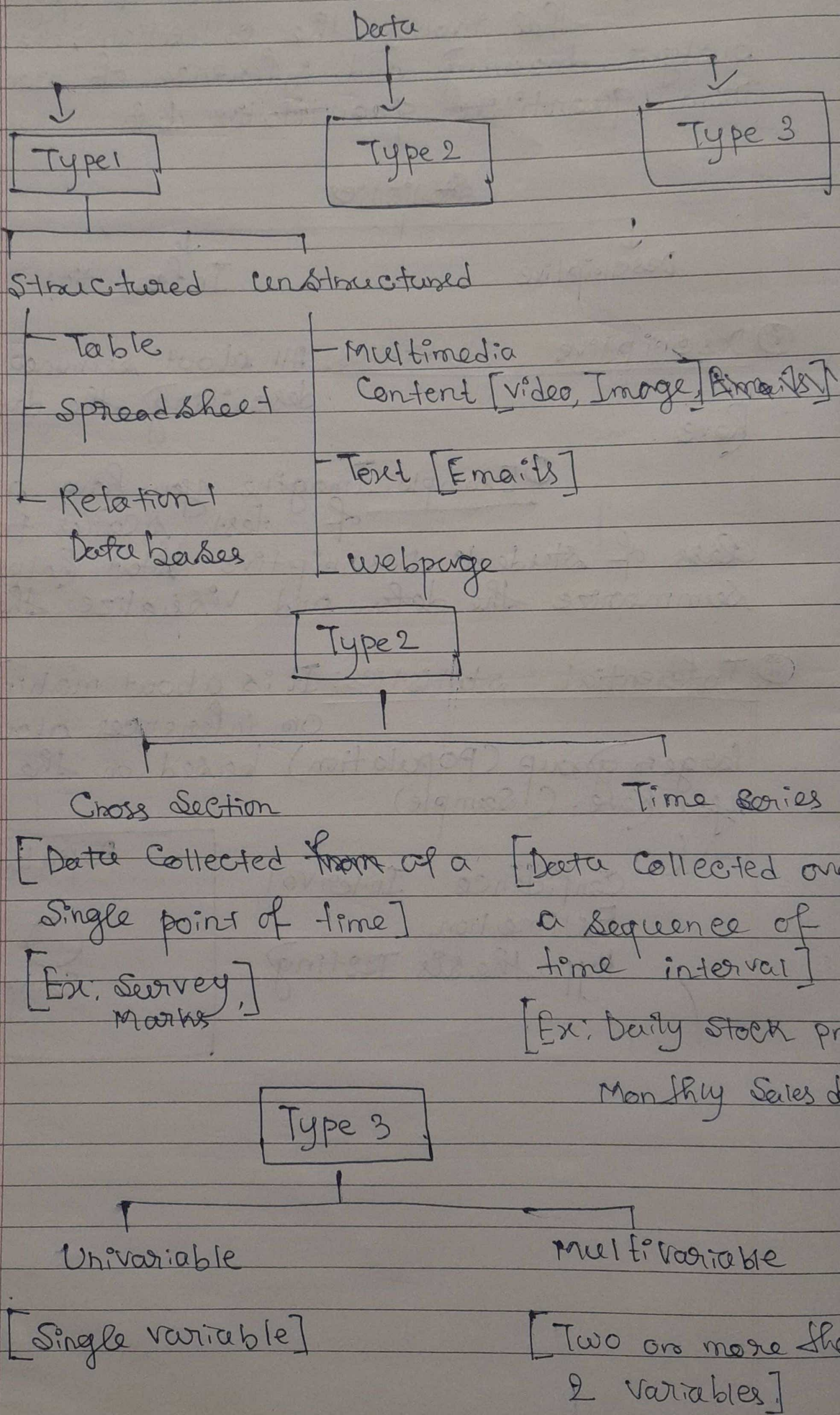
Example: Imagine you have a bunch of test scores from a class of students. Descriptive Stat helps you to summarize the data and visualize the data

(ii) Inferential statistics: It is about making predictions or inferences about a larger group (population) based on the data you have. (Sample)

↳ Confidence Interval
Estimation
Hypothesis Testing



2. Types of Data:



3. Types of variables:

(i) Nominal: Genders, Colours
(ii) Ordinal: Rich-poor, Education Levels, Customer Rating
(iii) Numerical: Income, Age, Price
(iv) Categorical: Types of cars, Product Categories
(v) Interval: Temp, IQ Series
(vi) Ration: Height, weight

4. Population and Sample:

(i) Population: It is the entire group of individuals

Ex: All people in India, All users of Netflix

(ii) Sample: It is a subset of population.

5. Why Sample is needed?

Ans: (i) To reduce the cost of data collection.
(ii) When a full census of data can not be taken.

6. Sampling Techniques:

(i) Random Sampling: Every members of the population has an equal chance of being selected.

(ii) Stratified Sampling: The population is divided into subgroups (Strata) and random samples are taken from each.

(iii) Systematic Sampling: Every nth member of the population is selected after a random starting point.

(iv) Cluster Sampling: The population is divided into clusters, some clusters are randomly selected, and all members of chosen clusters are sampled.

7. Factors to choose the Sampling Techniques:

Ans: (i) Nature of population
(ii) Research Objective.
(iii) Available Resource.

8. Steps for statistical data Analysis.

Ans: (i) Define the problem on research question.
(ii) Data collection.
(iii) Data Cleaning.
(iv) Exploratory Data Analysis [EDA]
(v) Data Transformation.
(vi) Hypothesis Formulation.
(vii) Statistical Testing
(viii) Interpretation of Results.
(ix) Draw Conclusions.
(x) Document the Analysis Process / Report Making.

9. Measures of Central Tendancy:

(i) Mean: The average of data points. Use when data is evenly distributed without extreme outliers. $\bar{x}$

(ii) Median: The middle value, when the data points are ordered. Use, when data has outliners or is skewed, as it is not affected by extreme values.

In case of even numbers, median is the average of the two middle value.

$x = \{3, 1, 7, 5, 9\}$

$= \{1, 3, (5), 7, 9\} \rightarrow$ odd

Median

$x = \{1, 3, 5, 9, 11, 13\}$ ← Even

$\frac{5+9}{2} = \frac{14}{2} = 7$

[Median]

(iii) Mode: The most frequently occuring value(s) in dataset.

→ Use, when you need to identify the most common value, especially in categorical data.

$x = \{(2), 4, 3, 4, (2), 6, 7, (2)\}$

Mode = 2

(*) Summary:

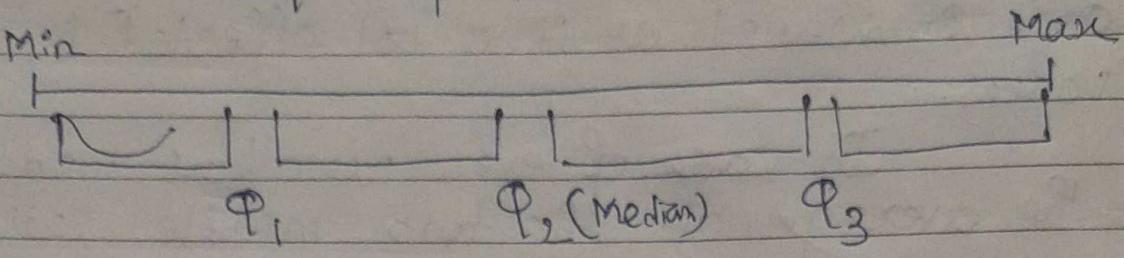
(i) Distribution of data (Depend)
(ii) Numerical Variable without outliers (mean)
(iii) Numerical Variable with outliers. (Median)
(iv) Categorical variable. (Mode)

b. Measures of Dispersion:

It describes the spread / variability of data points within a dataset.

(i) Range: Max - Min.
- (i) Sensitive to extreme values.
- (ii) It does not considers the distribution.
- (iii) Suitable for small dataset

(ii) Quartile: Value that divide a dataset in 4 equal parts.



Q_1 —— 25% —— Q_2 —— 50% —— Q_3 —— 75% ——

$Q_1 = (n+1)/4$ $Q_2 = (n+1)/2$ $Q_3 = \frac{3(n+1)}{4}$

Example: {4, 6, 7, 8, 10, 23, 34} $n = 7$

$Q_1 = \frac{(7+1)}{4} = \frac{8}{4} = 2$ $Q_1 = 6$

$Q_2 = \frac{8}{2} = 4$ $Q_2 = 8$

$Q_3 = \frac{3(n+1)}{4} = \frac{3(7+1)}{4} = \frac{24}{4} = 6$

$Q_3 = 23$

(iii) Percentile:

Formula: $P = \frac{P}{100} \times (n+1)$

A value, below which a given percentage of data points fall.

Use to understand the relative standing of a data point within a dataset.

(iv) Interquartile Range (IQR): The difference between first (Q_1) and third (Q_3) quartiles.

Formula: $Q_3 - Q_1 = 50\%$ of the data.

To see the middle portion of the data.
Less sensitive to extreme values.
Use to measure variability and to identify outliers in skewed distributions.

(v) Variance:

Formula: $(v) = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

N = Total number of data points
x_i = Individual data points
μ = mean

dataset = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

μ (mean) $= \frac{(5+6)}{2} = 5.5$

$(x_i - \mu)^2 = (1-5.5)^2$
$= (2-5.5)^2$
$= (3-5.5)^2$
$\vdots$
$= (10-5.5)^2$

Sum / N

The average of the squared differences from the ~~mar~~ mean.
Use to measure the overall variability within a dataset.

(vi) Standard deviation (Std): The square root of Variance, indicating the average distance from the mean. Use when you need to measure of spread that is in the same units as the data and to understand data dispersion.

Formula: $SD = \sqrt{Variance}$

(ii) Frequency: No of times, a value of data occurs
Use in Categorical variables.

(ii) Relative Frequency: Percentage or portion of the data value present in a complete dataset.

Formula: $\frac{Frequency}{Total\ no\ of\ Observations} \times [100\ \text{if Percentage}]$

(iii) Cumulative Frequency: It is the running total of frequencies upto a certain point in a dataset.

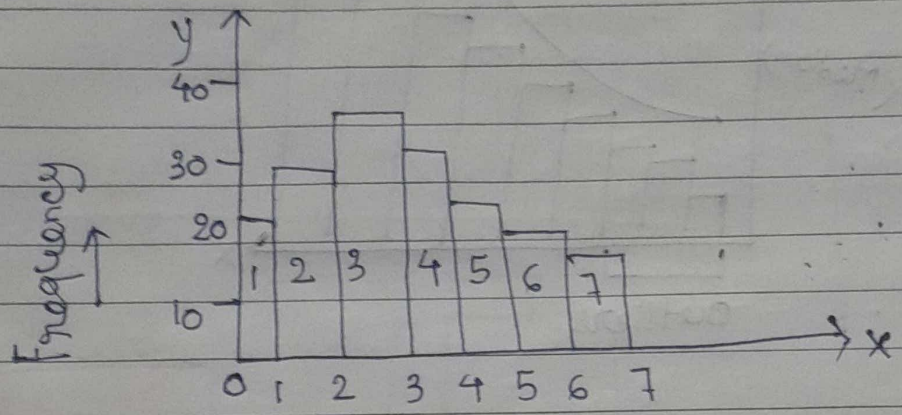
Ex: If the frequency of test scores are 2, 3, 5, then the cumulative frequency are → 2, 5=[2+3], 10=[2+3+5]

* Dataset = {1, 2, 1, 2, 3, 4, 2, 3, 4, 1, 1, 2, 3, 4, 4, 2, 1, 3, 4, 1, 2, 2, 3, 4}

values	Frequency	cumulative Frequency
1	6	6/24
2	7	7/24
3	5	5/24
4	6	6/24
	24	

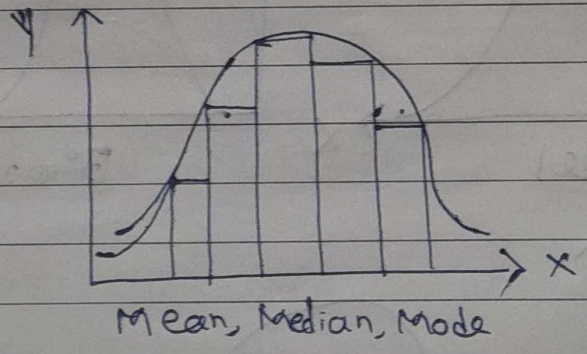
12. Graphical Representation:

(i) Histogram: A bar graph, telling the frequency distribution of a dataset.



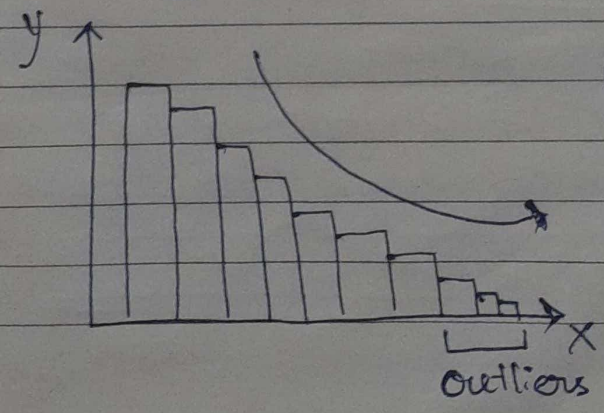
(ii) Types of Skewed Histograms:

(i) Symmetric (Normal Distribution): A histogram, where left and right sides are approximately mirror images.

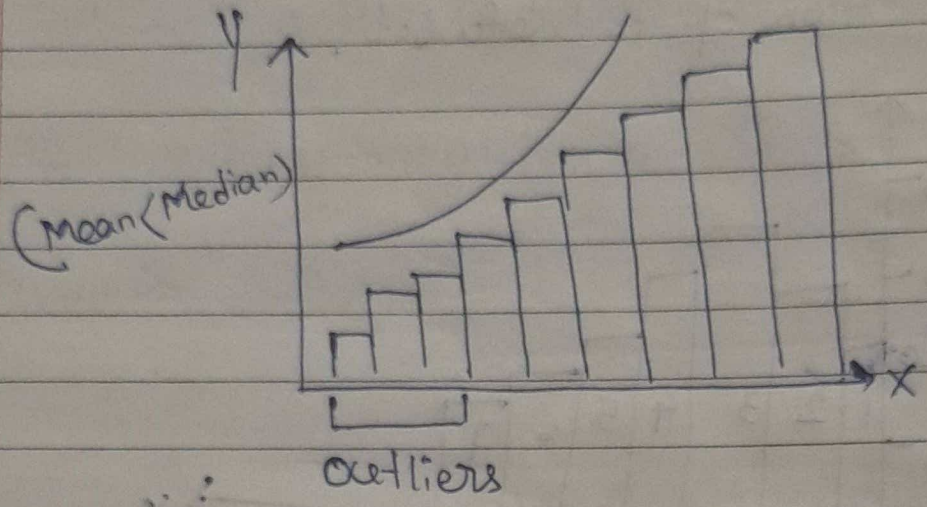


Center of the distribution

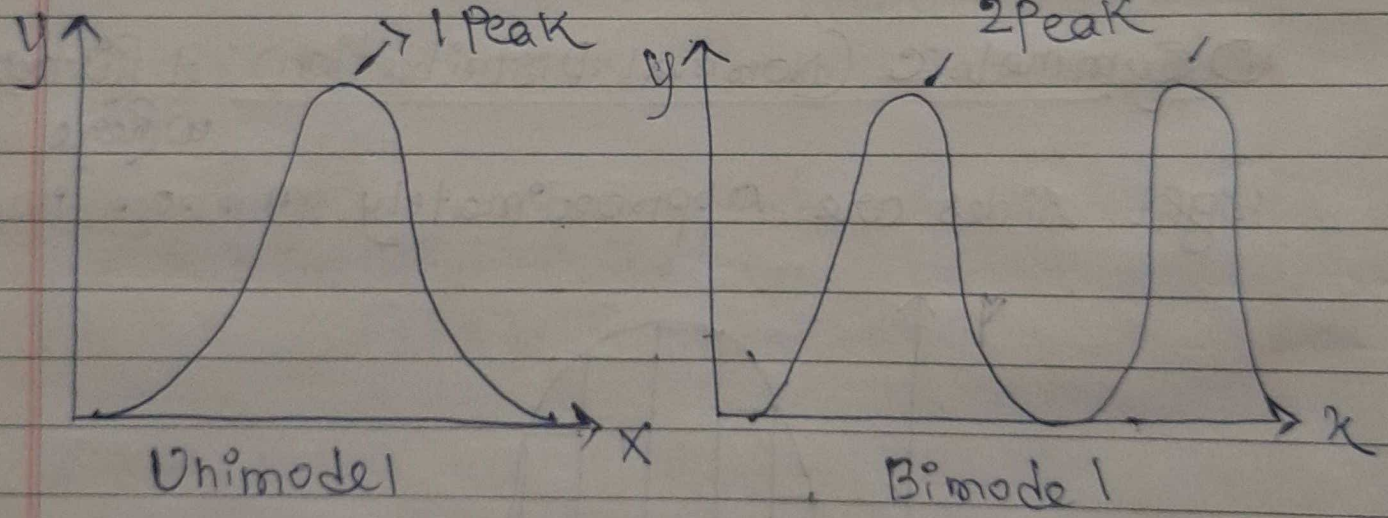
(ii) Right Skewed (+ve) Histogram: A histogram with the longer tail on the right side.

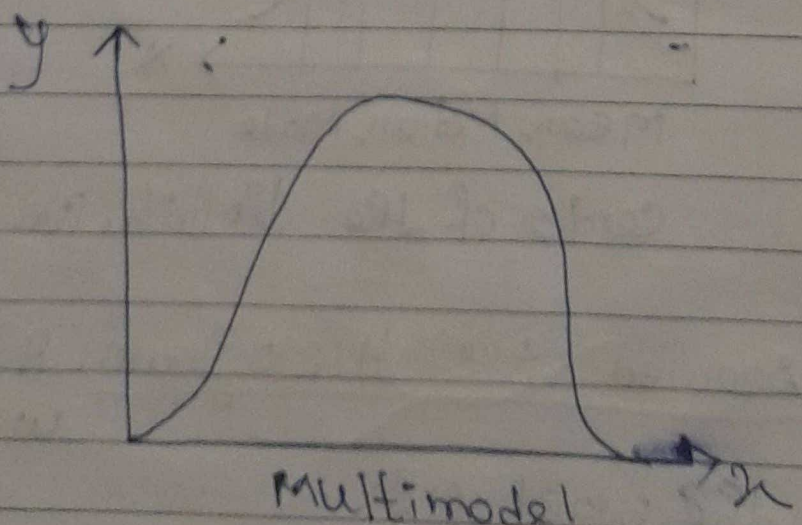


(ii) Left Skewed (-ve) Histogram: A Histogram with the longer tail on the left side.



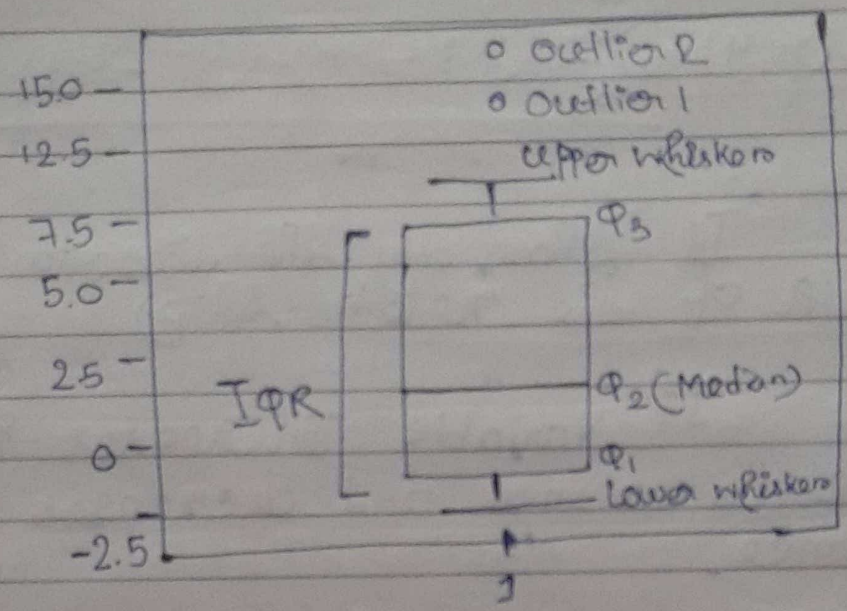
(iii) Types of Histogram based on number of modes:





(ii) Box plots: A graphical representation showing the distribution of the data and identify outliers. Also known as whisker plot.

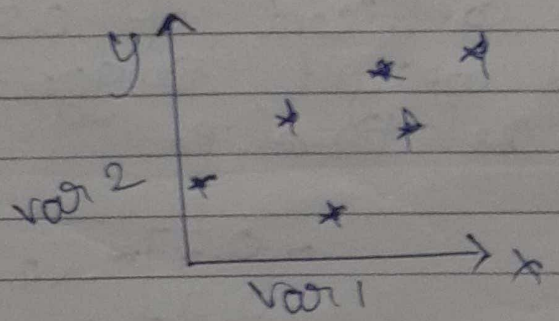
Box - IQR = $Q_3 - Q_1$



upper whisker = $Q_3 + (1.5 \times IQR)$
Lower whisker = $Q_1 - (1.5 \times IQR)$

5 Summary statistics

(iii) Scatter plots: Relationship between two continuous Variable.



Helps identifying patterns, trends, correlation, useful for identifying outliers & understanding strength and direction of relation.

13. (i) Covariance: A measure of how two variables change together.

Ex: positive covariance between study time and test scores indicates they tend to increase together.

$$Cov(x,y) = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$$

It shows only the direction (not the strength of the relationship)

If variable increases it is positive covariance and decreases negative covariance.

(ii) Correlation: A standarized measure of the strength and direction of the relationship between two variables. It shows linear relation. $(-1 < P < 1)$

-ve 0 +ve

No correlation

Ex: Correlation coefficent of 0.8 between height and weight suggests is a strong positive relationship.

$-0.5 < P < 0.5$ — No correlation or weak correlation

-0.7] -0.9] -ve correlation strong correlation

Pearson's correlation coefficent

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

$$Cov(x,y) = \sum_{i=1}^{N} \frac{(x_i - \mu_x)(y_i - \mu_y)}{N}$$

(iii) Causation: A relationship, where one variable directly affects another. It is direct connection in which one variable influences the other.

Ex: Increased hours of sunlight (causation) leading to higher plant growth.

Limitations:

(i) Outliers

(ii) It does not measure non linear relationship.

14. Removing outliers:

(i) Z score

(ii) $IQR = Q_3 - Q_1$

$Q_3 - 1.5 \times IQR <$ outlier

$Q_1 - 1.5 \times IQR >$ outlier

$$Z\text{ score} = \frac{X - \mu}{\sigma}$$