1) Define the following with examples from The dataset

* Types of data

1) ~~Nto~~ Numerical data :- Numerical data expresses information in the form of a measurable quantity, it answers the question "how many?" or "how much?".

2) categorical data :- categorical data expresses information that describes qualities or characteristics. it answers the quection "what kind?" or "which group".

* Types of statistics

1) Descriptive statistics :- descriptive statistics is all about summarizing and discribing the features of a dataset.

key components :

1) measures of central Tendency: These discribe the center of the data. mean, median, mode.

2) measures of variability :- These discribe the spread or scatter of the data. range, standard deviation.

3) Frequency Distributions: Displaying data in Tables or graphs To show the occurrence of different values or categories.

3) inferential statistics :- inferential statistics uses data from a sample to draw conclusions , predictions , or generalizations about a larger population

key Techniques :-

* Hypothesis Testing :- using sample data to test the validity of a claim about a populactio parameters .
* confidence intervals :- calculating a range of values within which the true populaction parameters is likely to fall along with a specified probadbility .
* Regression Analysis :- modeling the ~~relat~~ relationship between variables to predict the value of one variable based on one or more other variables.

* what is ~~di~~ descriptive statistics ?

descriptive statistics is a Branch of statistics that deals with summarizing , organizing and presenting the Basic features and characteristics of a set of data

2) Explain the difference between mean, median, mode?

mean :- The mean is the ~~value~~ ~~that~~ arithmetic average of all values.

Formula

$$\text{mean} = \frac{\text{sum of all values}}{\text{Total number of values}}$$

median :- The median is the middle values of all sorted data.

Formula

$$\text{median} = \frac{\text{count of value} + 1}{2}$$

mode :- The value is the most frequently in the dataset

~~formula:~~

~~mode~~ =

Explain the difference between range, variance and standard deviation?

range:- The Range is the simplest measure of spread it is the difference between the largest and smalles value in the dataset.

Formula..

range = maximum value - minimum value

variance:- variance is the average of the squared differences from the mean it quantifies how far the data point are spread out from the Average value.

Formula:-

$$\text{variance} = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

standard deviation:- The standard deviation is the square root of the variance it is the most widely used measure of variability because it brings the units back to the original scal of the data it represents the typical distance of the data points from the mean.

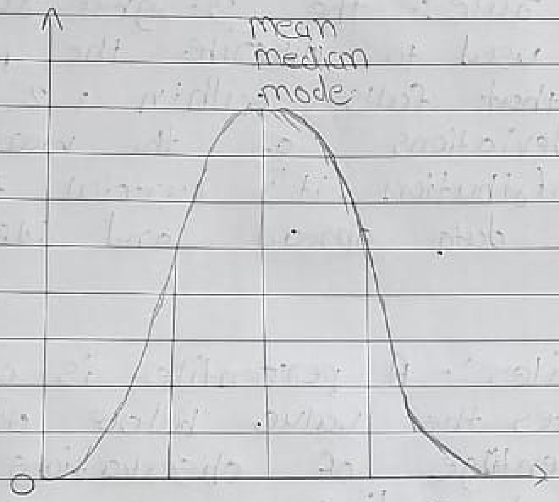
Formula:-

$$s = \sqrt{\text{variance}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

3) Explain the following term with neat and ~~clain~~ clean ~~dig~~ diagram along with its Formula:

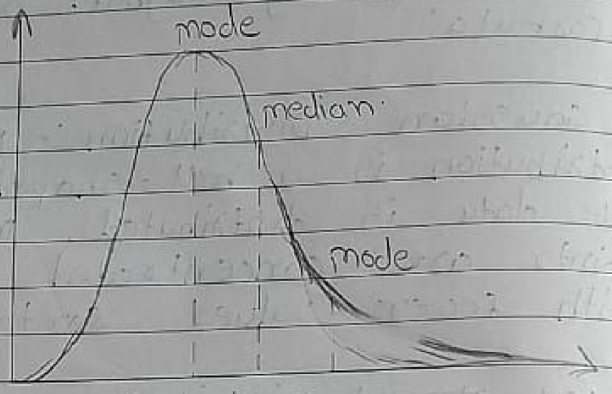
* Gaussian Distribution :- A Gaussion distribution is a bell - shaped curve that show how data is distributed, most of the data points are concentrated around the mean with fewer values farther away.

* ~~log~~ - ~~Normal~~ ~~Distribution~~.



* log - Normal distribution :- A log - Normal distribution is a continuous probAbility distribution of a random variable whose Natural logarithm is normally distributed if a random variable x is log - normally distributed then y = in (x) follows a normal distribution this distribution is always positively skewed and can only model variables that are greater than zero such ~~as~~

as stock price, income or cell size.



* 3-sigma rule :- the 3-sigma rule is a shorthand used to discribe the percentage of data that falls within 1, 2, and 3 standard deviations of the mean for a Normal distribution it's curcial for understanding data spread and identifying outliers.

* percentiles :- A percentile is a measure that indicates the value below which a given percentage of observations in a group of observations

* Quartiles :- quartiles are special percentiles that divide an ordered dataset into four equal parts Each quartiles contains 25% of the data.

* five number of summary :- the five number of summary provides a quick overview of the spread and center of a dataset it includes.

Term	Discription
minimum	The smallest ~~num~~ value.
Q1	First Quartile (25th parcentile)
Q2 (median)	second Quartile (50th parcentile)
Q3	third Quartile. (75th percentile)
maximum	The largest value

* Skewness :- skewness is a statistical measure ~~on~~ that discribes the asymmetry of a data distribution around its mean it helps identify if the data is symmetrically distributed or tilted towards one side.

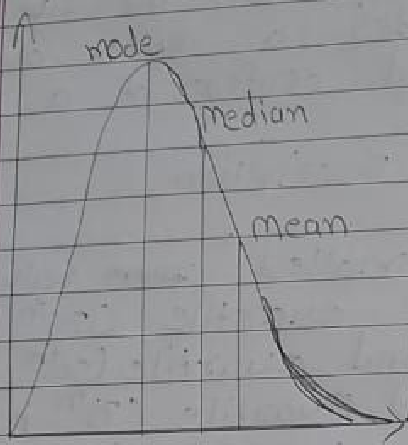
Types of skewness :-

1) symmetric :- data is evenly distributed is a bell shape curve.

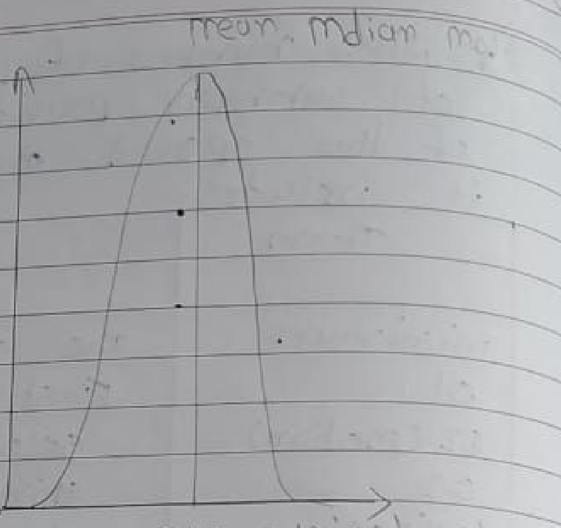
2) positive skewness :- data ~~sis~~ is tail on the Right side is longer. mean > median

3) Negative skewness :- data is tail on the left side is longer mean < median.

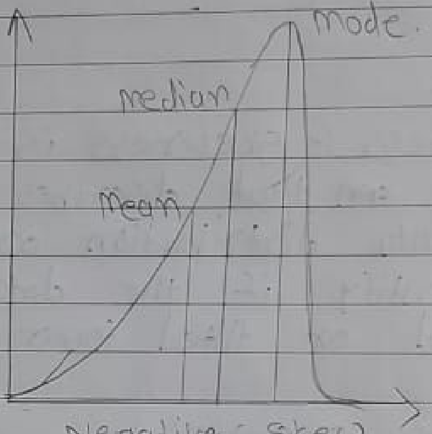
2025/12/04 12:37



positive skew



symetrical skew



Negative skew

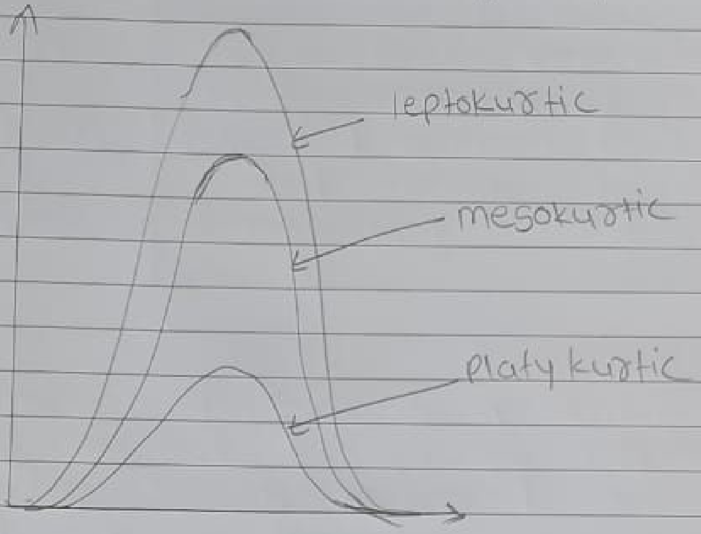
* kurtosis :- kurtosis measures the tailedness of a distribution - how havy or light the tail are compared to Normal distribution it tells us how it is to get ouliers.

Types of kurtosis :-

i) Mesokurtic :- Normal distribuction (Baseline) The shape are moderate peak and tails.

2) leptokurtic :- more peaked, heavier tails the shape are high peak, fat tails.

3) platykurtic :- flatter peak, lighter tails the shape are ~~Bo~~ Brod peak thin tails


leptokurtic
mesokurtic
platykurtic