KANAV BANSAL

MODE: The value that appears most frequently in a data set.

A set of data may have one mode, more than one mode, or no mode at all.

NOTE:

* The mode can be the same value as the mean and/or median, but this is usually not the case.

Mode is Unimodal and Bimodal.

MEASURE OF SPREAD: Also called as measure of dispersion / UNIVARIATE ANALYSIS

It is used to describe the variability in a sample or population. It is usually used in conjunction with a measure of central ~~tender~~ tendency, such as the mean or median, to provide

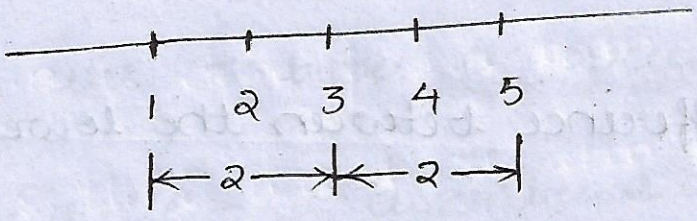
an overall description of a set of data.

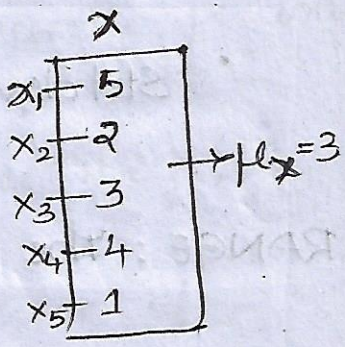
VARIANCE: The measurement of the spread between numbers in data set.

(or)

How far the data points are from the mean.

Example:





$\rightarrow \mu_x - x_1 = 3 - 5 = -2$

$\rightarrow \mu_x - x_5 = 3 - 1 = 2.$

As the variance needs to be positive, we use:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}}\ |\mu_x - x_i| = \frac{1}{n}\sum_{i=1}^{n}(\mu_x - x_i)^2$$

↑ Average

variance is average square distance b/w mean & data points

* As the modulus value increase the magnitude of the value, we use standard deviation.

So, the measure of spread is standard deviation. (12)

STANDARD DEVIATION: The statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

$$\text{Std dev} = \sqrt{\text{var}}$$

RANGE: The difference between the lowest and highest values.

Example:

5	2	3	4	1

Range = 5 − 1 = 4

PERCENTILES: Also called as centile.

A type of quantile which divides the given probability distribution, or sample into 100 equal-sized intervals.

↳ This allows the data to be analyzed in the

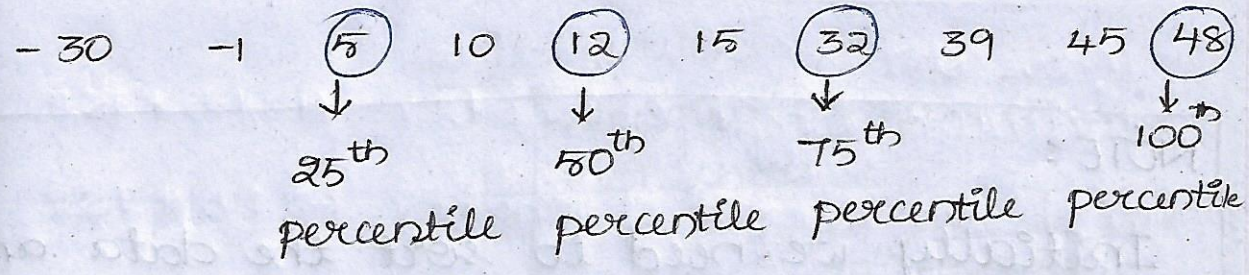
terms of percentages.

Example:

I SCORED 48/100 IN MATHEMATICS

↓

100 percentile - gives the rank.



There are 3 types of percentiles. They are

1. Quartile - It describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observations. (25%, 50%, 75%)

2. Quintile - It describes a division of observations into five defined intervals based on the values of the data as 20%, 40%, 60%, 80%.

3. Decile - It describes a division of observations into ten defined intervals based on the values of the data as 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%.

NOTE:
Initially we need to sort the data and then the percentile is taken.

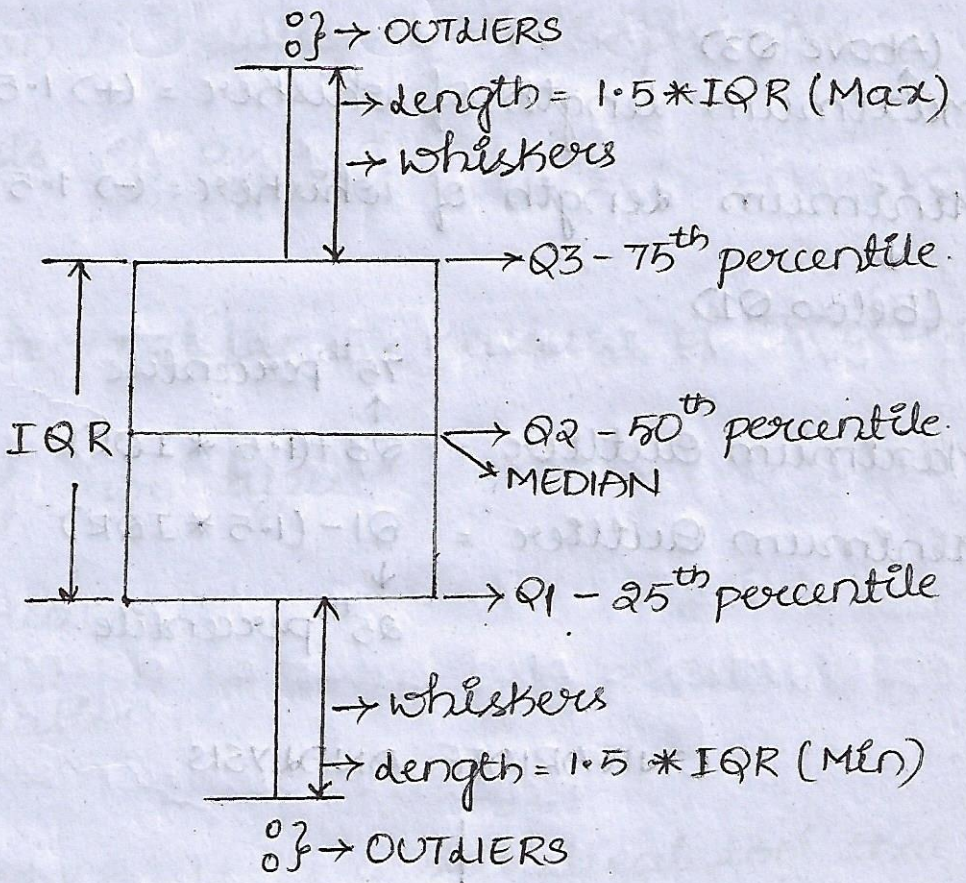
IQR : INTERQUARTILE RANGE

It is a measure of variability, based on dividing a data set into quartiles.

↳ The quartiles divide a rank-ordered data set into four equal parts.

↳ The values that divide each part are called the first, second and third quartiles; and they are denoted by Q_1, Q_2, Q_3 respectively.

BOXPLOT : Also known as box and whisker plot. It is a type of chart often used in EDA to visually show the distribution of numerical data and skewness through displaying the data quartiles (percentiles) and averages.



In statistical theory;

length of whiskers = 1.5 * IQR

OUTLIERS : It is defined as a data point that is located outside the whiskers of the boxplot.

↳ An outlier is an observation that is numerically distant from the rest of the data.

> Outliers → Extremely small values or large values.

(Above Q3)
↑
Maximum length of whisker = (+) 1.5 * IQR

Minimum length of whisker = (−) 1.5 * IQR
↓
(Below Q1)

75th percentile
↑
Maximum Outlier = Q3 + (1.5 * IQR)

Minimum Outlier = Q1 − (1.5 * IQR)
↓
25th percentile.

UNIVARIATE ANALYSIS

- MEASURE OF SPREAD
  - VARIANCE – Impacts with outliers
  - STD. DEVIATION – "
  - RANGE – Impacts with outliers
  - IQR – Doesn't impacts with outliers
- MEASURE OF CENTRAL TENDENCY
  - MEAN – Impacts with outliers
  - MEDIAN – Doesn't Impacts with outlier
  - MODE – "

BIVARIATE ANALYSIS : It is one of the simplest form of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them.

↳ This can be helpful in testing simple hypotheses of association.

Bivariate Analysis → Measure of Relationships

It is of two types.

1. Covariance.
2. Correlation.

1. COVARIANCE : It is a statistical tool that is used to determine the relationship between the movement of two variables.

↳ When two stocks tend to move together, they are seen as having a positive covariance.

i.e., if $X \& Y$ is taken then $X \propto Y$

↳ When the two stocks tend to move inversely then the covariance is negative i.e., if $X \& Y$ are taken then $x \propto \frac{1}{y}$

SIGN MATTERS for covariance. ➡

↓

$COV(x,y) = \pm$ values.

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

If $COV(x,y)$ is +ve. $(x\uparrow, y\uparrow) \rightarrow x \propto y$

If $COV(x,y)$ is -ve $(x\uparrow, y\downarrow) \rightarrow x \propto \frac{1}{y}$

2. CORRELATION : Also called as dependence or (ρ) Pearson correlation coefficient.

In statistics, correlation is any statistical relationship, whether casual or not, between two random variables or bivariate data.

↳ When the two sets of data are strongly linked together we say that they have a

- HIGH CORRELATION.

↳ When the correlation is positive, when the value increases together.

↳ The correlation is negative, when one value decreases as the other increases.

The word correlation is made of :

co-(together) and relation.

correlation $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X * \sigma_Y}$ and lies between $-1 \leq \rho_{xy} \leq 1$

If $\rho = -1$ $(X\uparrow, Y\downarrow)$ - All points lies on same line.

If $\rho = 0$ - All points lies in a random manner

If $\rho = 1$ $(X\uparrow, Y\uparrow)$ - All points lies on same line.

$-1 < \rho < 0$	$0 < \rho < 1$
Negative Relationship and variance is a lot between the points.	Positive Relationship and variance is a lot between the points

NOTE:

correlation can only detect the linear relationships.

DRAWBACK : Can't detect the patterns apart from the linear relationships.