

INTRODUCTION OF STATISTICS

Points to be covered in this topic

1. INTRODUCTION
2. MEASURES OF CENTRAL TENDENCY
3. MEASURES OF DISPERSION
4. CORRELATION

INTRODUCTION

- Statistics is defined as, "the **discipline** that concerns with the **collection, organization, analysis, summarization, interpretation** and **presentation of data**".
- **Descriptive statistics**: Descriptive statistics is a summary statistic that **quantitatively describes** or summarizes features of **collected information**. It is a process of **using and analyzing a set of information** that has been collected only.
- **Inferential statistics**: Inferential statistics is the **process of data analysis** to deduce properties of an underlying probability distribution. It is used to **deduce properties of a population** by **testing hypotheses** and deriving estimates.
- Biostatistics the branch of statistics that **deals with data relating to living organisms**.
- Biostatistics applied to the **collection, analysis, and interpretation of biological data** and especially data relating to **human biology, health, and medicine**.

Steps in Biostatistics:

1. Generation of hypothesis.

2. Collection of experimental data.
3. Classification of the collected data.
4. Categorization and analysis of collected data.
5. Interpretation of data.

Mean

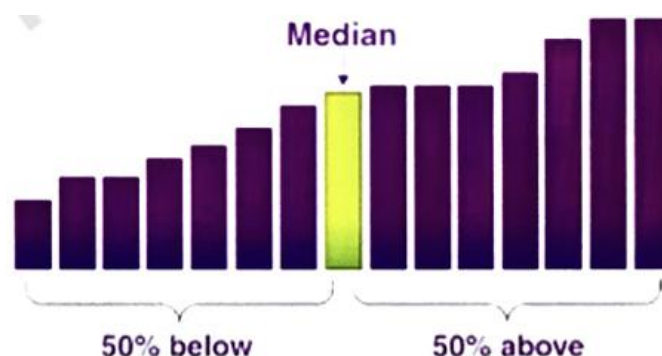
- Mean is the **average of the given numbers** and is calculated by **dividing the sum of given numbers** by the **total number of numbers**.
- Mean is nothing but the average of the given set of values.
- It denotes the **equal distribution of values** for a given data set.

$$\text{Mean } (\bar{x}) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n}$$

- To calculate the arithmetic mean of a set of data we must first add up (sum) all of the **data values (x)** and then divide the result by the number **of values (n)**. Since $\sum$ is the symbol used to indicate that values are to be summed.

Median

- Median, in statistics, is **the middle value of the given list of data** when **arranged in an order**.
- The arrangement of data or observations can be made either in **ascending order or descending order**.



Odd Number of Observations	Even Number of Observations
If the total number of observations given is odd, then the formula to calculate the median is:	If the total number of observation is even, then the median formula is
$\text{Median} = \left(\frac{n+1}{2} \right)^{\text{th term}}$	$\text{Median} = \frac{\left(\frac{n}{2} \right)^{\text{th term}} + \left(\frac{n}{2} + 1 \right)^{\text{th term}}}{2}$

Mode

- In statistics, the **mode is the value that is repeatedly** occurring in a given set.
- A mode is defined as the **value that has a higher frequency** in a given set of values.
- It is the value that **appears the most number of times**.
- Example: In the given set of data: 2, 4, 5, 5, 6, 7, the mode of the data set is 5 since it has appeared in the set twice.

Mean Median Mode Comparison

Mean	Median	Mode
Mean is the average value that is equal to the ratio of sum of values in a data set and total number of values.	Median is the central value of given set of values when arranged in an order.	Mode is the most repetitive value of a given set of values.
For example, if we have set of values = 2,2,3,4,5, then; Mean = $(2+2+3+4+5)/5 = 3.2$	Median = 3	Mode = 2

MEASURE OF DISPERSION

- The measure of dispersion presents the **scatterings in the data**.
- Dispersion can also be called as **variability, scatter or spread**.
- It helps to interpret the **variation of the data from one another** that is to **know how much homogenous or heterogeneous** the data is, and gives a clear picture about the **distribution of the data**.

Absolute measure of dispersion	Relative measure of dispersion
This type of measure of dispersion contains the same unit as the original data set. This type states the variations as average of deviations of observations like standard or means deviations. Examples: Range, SD, quartile deviation.	Relative measures of dispersion, also known as coefficient of dispersion (CD), are obtained as ratios or percentages of the average. Examples: Coefficients of range, variation, SD, quartile deviation and mean deviation.

Range

- Range is the **difference between two extreme observations** of the **data set such as lowest and highest values**.
- It is the **most common and comprehensive measure of dispersion**.

$$\text{Range} = X_{\max} - X_{\min}$$

Standard deviation

- SD is a measure of the **amount of variation** or **how spread out numbers** in a set of SD is also represented by the lower case **Greek letter sigma (σ)** for the population SD Latin letter S for the sample SD.

- The **low value of SD indicates** that the values tend close to the mean, also called the '**expected value**', of the data set.
- The high value of indicates that, the values are spread out over a wider range.
- The **SDS of a random variability statistical population**, data set and probability distribution is the square root of its variance.

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

$$SD = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum (x)^2}{n}\right)}$$

Characteristics of SD

1. It includes **algebraic signs** and is less affected by sampling fluctuations.
2. **Small SD** has **high probability** of **getting a value close to the mean** and if it is large, **the value is further away from the mean**.
3. SD is independent of origin, but not of the scale.
4. SD can be **mathematically manipulated**.
5. It is **most widely and very satisfactorily** used measure of dispersion.

Application of SD

1. It **describe the variation** (deviation) of a large distribution from mean that mean it is used as a unit of variation.
2. Indicates whether the **variation of difference** of an individual from the mean is **by chance** i.e. **natural or real due to some special reasons**.

3. It helps to **find out error**, which determines whether the differences between means of samples is by chance or real.
4. It also helps in **finding the suitable size of sample** for valid conclusion.
5. It is used to **measure confidence in statistical** conclusions.
6. SD is used to **compare real world data against a model to test the model**.

Pharmaceutical example

Question: The relative humidity in syrup production department of a pharmaceutical unit is given in table, calculate the SD in percent relative humidity.

Days	MON	TUE	WED	THU	FRI	SAT	Σ of days = 6
X	60	62	65	69	75	65	$\Sigma X = 396$
x ²	3600	3844	4225	4761	5625	4225	$\Sigma x^2 = 26280$

Solution: The mean percent relative humidity and SD are calculated as:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{396}{6} = 66$$

$$SD = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$$

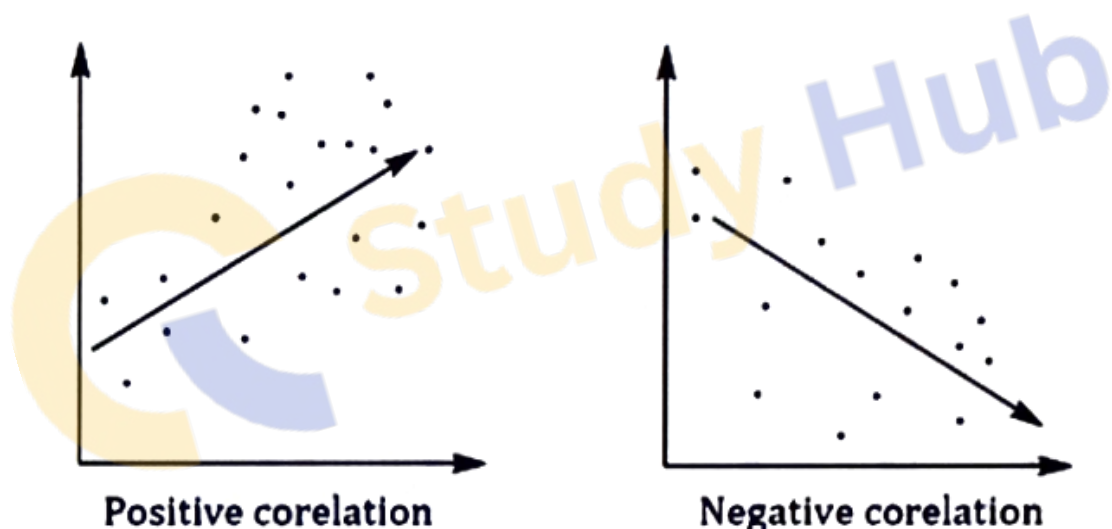
$$SD = \sqrt{\frac{26280}{6} - \left(\frac{396}{6}\right)^2}$$

$$SD = \sqrt{4380 - 4356}$$

$$SD = \pm 4.89$$

CORRELATION

- Correlation is a **statistical measure of relation that indicates the extent** to which **two or more variables fluctuate together**.
- A **positive correlation is the extent to which** those variable **either increase or decrease in parallel**.
- A **negative correlation is the extent to which** one variable increases as the other decreases or vice versa.
- If the **change in values of one variable makes a constant ratio with the change in value of other variable**, then such type of relation known as linear correlation.
- The correlation is said to be a **non-linear** if the **value in one variable does not make a constant ratio with change in the value of other variable**.



Karl Pearson's Coefficient of Correlation

- Karl Pearson's Coefficient of Correlation is **used to measure the degree of linear relationship** between **two variables**.
- It is also called moment **correlation coefficient**.
- It is denoted by 'r' and defined as:

$$r = \frac{\sum XY}{N\sigma_x\sigma_y}$$

Where $X = X - \bar{X}$ and $Y = Y - \bar{Y}$

N = No of pair of values of variations

σ = Standard deviation

- Another form of correlation coefficients as:

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

Merits

1. It is important method to **give a precise and quantitative result** with a **meaningful interpretation**.
2. It also **gives a direction** (i.e. positive or negative) as well as the degree of the correlation between the variable.

Demerits

1. This method is a **time consuming**.
2. The **limitation of value of correlation** is $(-1 < r < +1)$.

Multiple correlation

- In multiple correlation, we study the relationship **between three or more variable**. Suppose the dependent variable is z and ' x and y ' both are independent variables.

$$R_{zxy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$

Pharmaceutical example

Question: Find the value of the correlation coefficient for the blood sugar level data given in table for patient with different age group.

Subject	Age (x)	Blood sugar level (y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Solution:

Subject	Age (x)	Blood sugar level (y)	xy	x ²	Y ²
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241

Subject	Age (x)	Blood sugar level (y)	xy	x ²	Y ²
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
n = 6		Σx = 247	Σy = 486	Σxy = 20485	Σx ² = 11409
					Σy ² = 40022

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{6(20485) - (247 \times 486)}{\sqrt{[6(11409) - (247)^2] \times [6(40022) - (486)^2]}}$$

$$r = 0.5298$$

As we know the range of the correlation coefficient is from -1 to 1, the result $r = 0.5298$ or 52.89 %, indicates that variables has a moderate positive correlation.