Describing Data

BY… SAKEEB SHEIKH

Describing Data

- 5 Characterstics :-
  - Center
  - Variation
  - Distribution
  - Outliers
  - Changes Over Time.

Measurement of Central Tendency

- Three Ways to find the middle of the dataset
  - *Mean*
  - *Median*
  - *Mode*

Mean

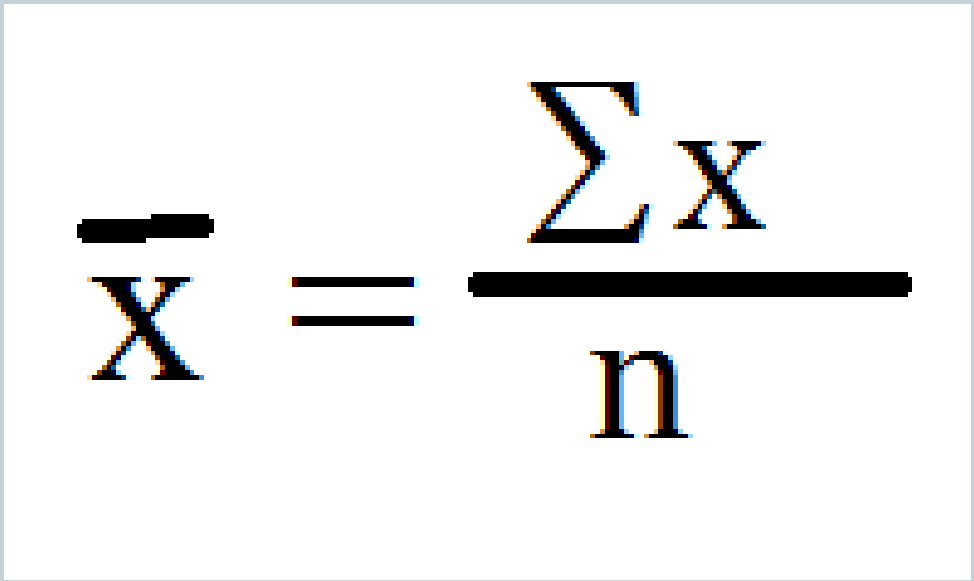
- Sum of all the Data Points divided by the number of data points.

- Formula :

$$\text{Mean} = \frac{\sum x}{n}$$

n → Number of Data Points

- $\sum$ → Sum
- X → Data Value
- n → no of items in the sample
- N → no of items in the Population
- $\bar{x}$ → Sample Mean
- μ → Population Mean


$\bar{x} = \frac{\sum x}{n}$

Example

[5.40, 1.10, 0.42, 0.73, 0.48, 1.10]

$$\bar{x} = \frac{5.40 + 1.10 + 0.42 + 0.73 + 0.48 + 1.10}{6}$$

$$\bar{x} = 1.54$$

Median

- The MIDDLE value of a Data Set.
- Must be in Order
- Find Middle Value
  - If ODD no of values the median will be the middle no.
  - If EVEN no of values the median will be the MEAN of the two middle no's.

Example

- Example-1

[1, 3, 4, 5, 6, 7] → Median = 4.5

- Example-2

[8, 3, 5, 11, 13, 4, 6] → Median = 6

- Example-3

[3, 4, 5, 6, 8, 11, 13, 412] → Median = 7

Comparing Mean and Median

- Calculating Mean

[5.40, 1.10, 0.42, 0.73, 0.48, 1.10]

$$\bar{x} = 1.54$$

- Calculating Median

[5.40, 1.10, 0.42, 0.73, 0.48, 1.10]

$$M = 0.915$$

Problem With Mean

- For example you have collected the data of 100 people's monthly earning but in the dataset there are 2 people who are earning 20000 per month while the other 98 peoples have earning ranging between 500 to 10000.
- Now in the above Scenario mean will get highly affected and will give some unusually results.

- So its better to obtain median value in this case.
- Suitable when there is presence of OUTLIERS .

MODE

- The Most Repeated Value the Data Set.

Example-1

[5.40, 1.10, 0.42, 0.73, 0.48, 1.10]

Mode = 1.10

Example-2

[2, 2, 2, 5, 5, 5, 7, 8]

Mode = 2 and 5

Mode Example

Example-3

[1, 2, 3, 5, 7, 8, 9, 10]

Mode = Φ

Summarizing the terms

- $\sum$ → Sum
- X → Data Value
- n → no of items in the sample
- N → no of items in the Population
- $\bar{x}$ → Sample Mean
- μ → Population Mean

What is Population And Sample

- Population Defines All the values from a given variable
  - Ex :- Set of all the peoples in a country.
- Sample is the subset of the population.
  - The Way How you Collect Data.
  - Ex :- selecting some people from the population.

What is the need of Sampling and How to Do Sampling ?

Need of Sampling

- For example, to measure the diameter of each nail that is manufactured in a mill is impractical.
- But we need some technique to select those nails from all the nails manufactured in mill.
- How to Do It ?

Using Random Sampling

Random Sampling

- Use a random sample of data taken from a population to describe and make inferences about the population.
- Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible.

Simple Random Sample

- Fixed Sized Random Samples
- Four Sampling Techniques:-
  - Convenient Sampling.
  - Systematic Sampling.
  - Stratified Sampling.
  - Cluster Sampling.

Outliers

- Data Points that are way outside the normal Data points.

Frequency Distribution

Helps in understanding Trends within the data

Gears	Cars
3	15
5	12
8	3

Frequency Distribution

- A List of values with corresponding frequencies.

- **CLASS WIDTH** → Difference between the two lower class limits.
- **LOWER CLASS LIMIT** → Smallest value belonging to a class.
- **UPPER CLASS LIMIT** → Largest value in a class.

- **CLASS MIDPOINT** is the
  (Upper Class Limit + Lower Class Limit)/2

- **CLASS BOUNDARIES**
  used to separate Classes Without Gaps.
  (Lower Class Limit second class +
  Upper Class Limit of first class)/2

Steps to find Frequency Distribution

- Step-1 : Determining No. of Classes → 8
- Step-2 : Calculating CLASS WIDTH

$$\frac{\text{Max Value - Min Value}}{\text{No. of Classes}}$$

- Step-3 : Start With the Smallest Value.
- Step-4 : Create Classes with Class Width.

  Create Lower Class Limits.

Example

- Consider we have two variables **AGE** and **Count Of Peoples**.
- And we need to organize this data to understand the distribution of data between different age groups.
- Now by seeing the range of age variable which ranges from 18 to 44, we decide the no. of classes.
- And then we calculate the Class width

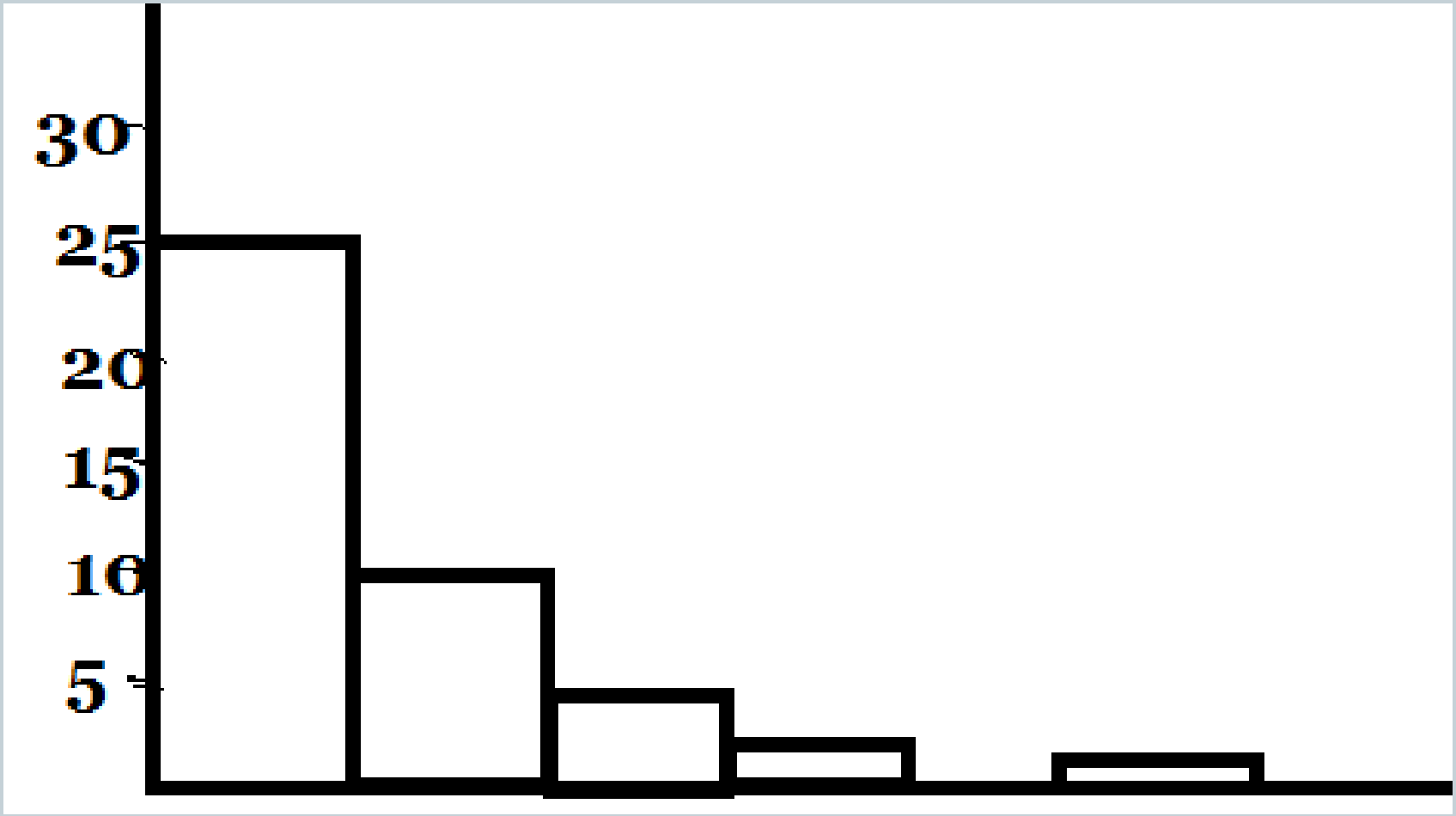
Therfeore, **(44-18)/8=3.25 = roundup(3.25) = 4**

Frequency Distribution

AGE	frequency f	Relative frequency (%)	Cummulative frequency
18-21	25	58.1 %	25
22-25	10	23.3 %	35
26-29	4	9.3 %	39
30-33	2	4.7 %	41
34-37	1	2.3 %	42
38-41	0	0 %	42
42-45	1	2.3 %	43
46-49	0	0 %	43
	n=43		

Normal Frequency Distribution

Representing Frequency Distribution using Histogram



Histogram

- Histogram → Like a Touching Bar Chart
- Horizontal Axes → Class Midpoint or Boundaries
- Vertical Axes → Frequences

Finding Mean of Frequency Distribution

Ages	Frequency f	Mid Point of Class X	f.x
21 – 30	28	25.5	714
31 – 40	30	35.5	1065
41 – 50	12	45.5	546
51 – 60	2	55.5	111
61 – 70	2	65.5	131
71 – 80	2	75.5	151
	n = 76		**$\sum$ f. x = 2718**

Finding Mean of Frequency Distribution

$$\overline{x} = \frac{\sum f \cdot x}{n}$$

$$\overline{x} = \frac{2718}{76} = 35.76$$

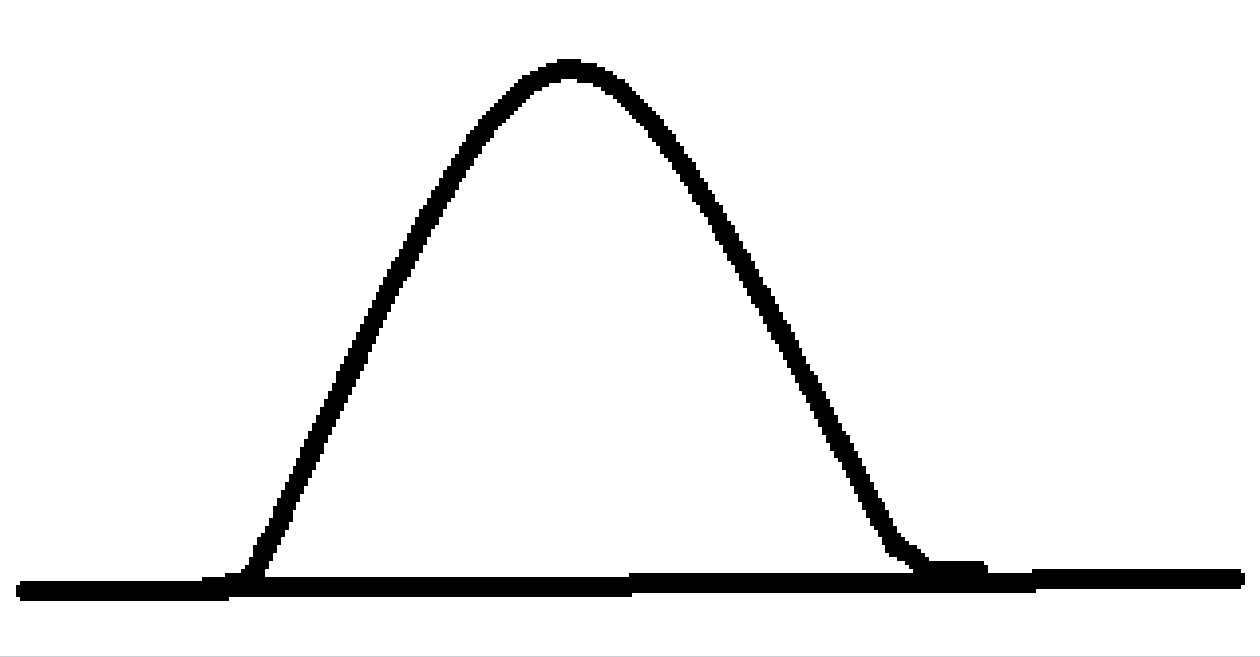
Mean of Weighted Frequency Distribution

Heads	Weights (w)	Points Scored(X)	W . X
HW	15%	70	10.5
Test-1	20%	90	18.0
Test-2	20%	68	13.6
Test-3	20%	85	17.0
Final	25%	95	23.75
	$\sum w = 100$		$\sum x.w = 82.85$

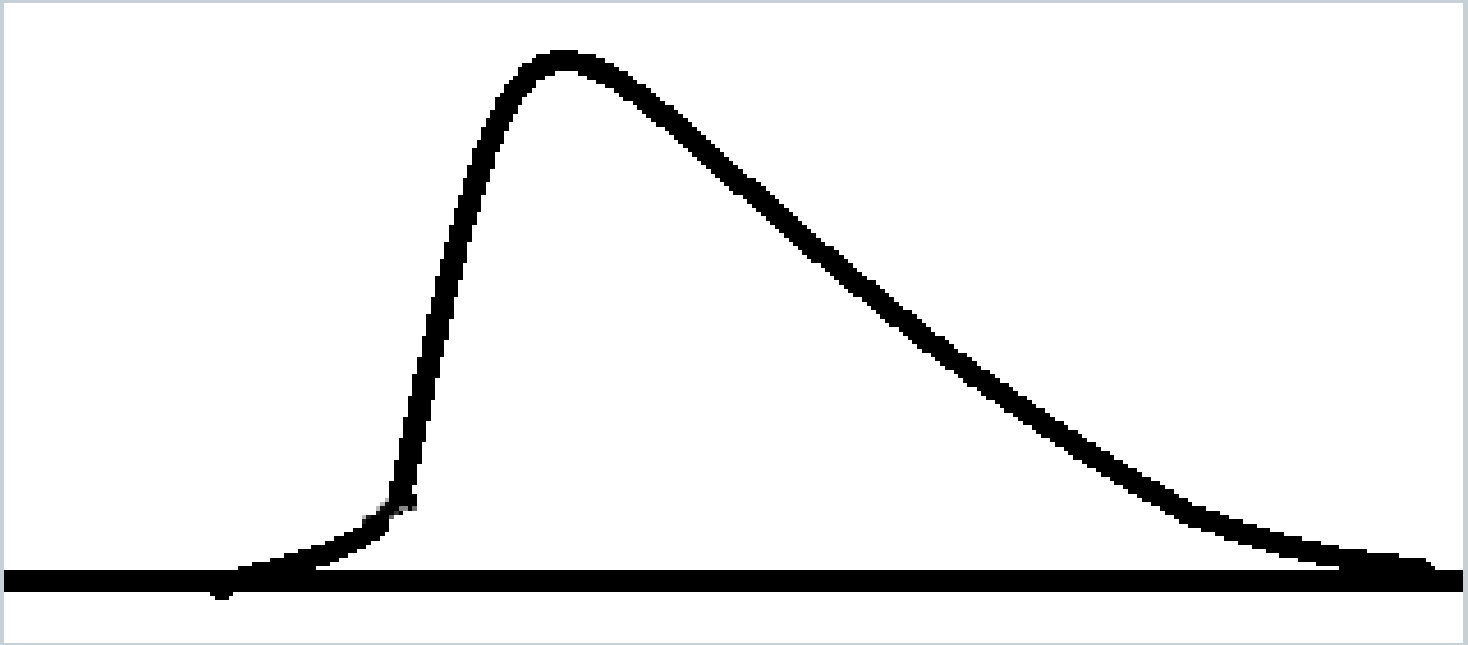
$$\overline{X} = \frac{\sum x.w}{} = \frac{82.85}{100} = 0.7654$$

Skewness

- Lack of Symmetry (Effect Of Outliers)
- Normal Distribution



- Skewed Right



- Outlier On right side to the data
- Lack of Symmtry

Measure of Variation

- Variance :- How the Data is Spread.
- Ways To Measure Variation

First Way :

1. **Range** :- Max_Value – Min_Value
   - Do no consider all the values.

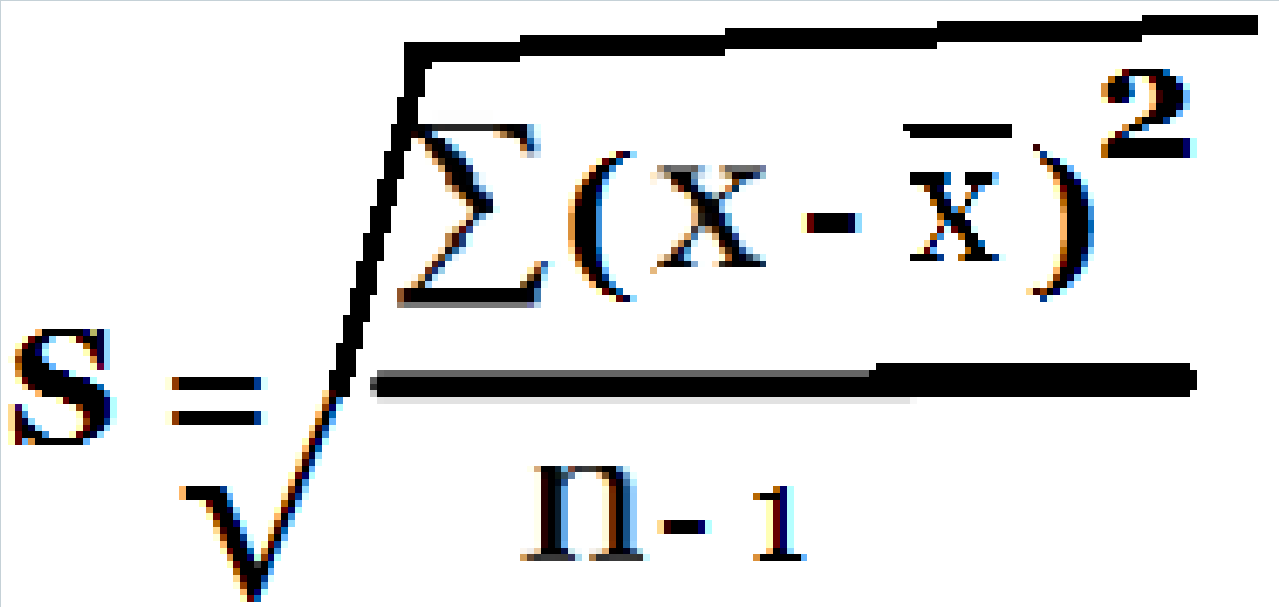
Value-1	Value-2	Value-3	Variance
6	6	6	$\bar{X} = 0$
3	4	7	$\bar{X} = 4$

Second way to find Variance

- Standard Deviation :
  - **Measures the Average Distance your Data Values are from the mean.**

1. **Never Negative.**
2. **Never Zero unless all the data points are same.**

Standard Deviation



Calculating Standard Deviation

X	$X - \bar{X}$	$(X - \bar{X})^2$
1		
3		
14		

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

Central Limit Theoram

- We use Sample to represent Population.
- We Assume that we are sampling from a population which has some mean and some standard deviation.
- **Central limit theorem States that by increasing the sample size we get nearer to normal distribution.**

What is 'Hypothesis Testing'

- An act to **tests an assumption** regarding a population parameter.
- methodology depends on the **nature of the data** used and the **reason for the analysis**.
- used to **infer the result** of a hypothesis performed on sample data from a larger population.

4 Steps to Perform Hypothesis Testing

- state the two hypotheses so that only one can be right.
- formulate an analysis plan, which outlines how the data will be evaluated.
- carry out the plan and physically analyze the sample data.
- analyze the results and either accept or reject the null.

Methods for Hypothesis Testing

- T-Test (for samples size less than 30)
- Z-Test (for large samples size but max two variables)
- ANOVA Test (Multi variable)
- Chi-Square Test

Z-Test for Hypothesis testing

- A **z-test** is used for testing the mean of a population versus a standard, or comparing the means of two populations, with large ($n \geq 30$) sample size .
- Its based on the Z-statistic, which follows the standard normal distribution under the null hypothesis.
- A one-sample location test, two-sample location test, paired difference test and maximum likelihood estimate , where you may perform z-tests.

- You can also use Z-tests to determine whether predictor variables in logistic regression have a significant effect on the response.
- The null hypothesis states that the predictor is not significant.
- For n<30, you may perform T-test instead !

Assumption for Z-test

1. Interval or ratio scale of measurement (approximately interval)
2. Random sampling from a defined population
3. Characteristic is normally distributed in the population

Hypothesis Testing: T-Tests

- **Hypothesis testing** uses statistics to choose between hypotheses regarding whether data is statistically significant or occurred by chance alone.
- **T-Test** examine whether two means are statistically significantly different from each other or whether the difference between them simply occurred by chance.

Assumption for T-test

1. Interval or ratio scale of measurement (approximately interval)
2. Random sampling from a defined population
3. Characteristic is normally distributed in the population

Steps For **T-Test** Hypothesis Testing

1. **Determine a null and alternate hypothesis.**

Null Hypothesis : *Height of men & women are the same*

Alternate Hypothesis : *Height of men & women are the different*

2. **Collect sample data**
3. **Determine a confidence interval and degrees of freedom**

$$df = n_x + n_y - 2$$

4. **Calculate the t-statistic**

- t-statistic can be calculated using the below formula.

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

M = mean
n = number of scores per group

$$S^2 = \frac{\sum (x - M)^2}{n - 1}$$

x = individual scores
M = mean
n= number of scores in group

- where, Mx and My are the mean values of the two samples of male and female.
  Nx and Ny are the sample space of the two samples
  S is the standard deviation

- Calculate the critical t-value from the t distribution
- A p-value is the probability that the results from your sample data occurred by chance.

One-Sample T-Test

- A **One-Sample T-Test** compares a sample mean to a known population mean to determine whether the difference between the two means is statistically significant or occurred by chance alone..

Independent Sample T-Test

- An **Independent Samples t-test** compares the means for two groups.

Pair Sample T-Test

- A **Paired sample t-test** compares means from the same group at different times (say, one year apart).

ANOVA Test

- ANOVA – Analysis Of Variance

FOR MORE THAN TWO SAMPLES

Basic Assumption

- Dependent Variable should be measured at least interval.
- Independence of Data.
- Normality
- Homogeneity of variance.