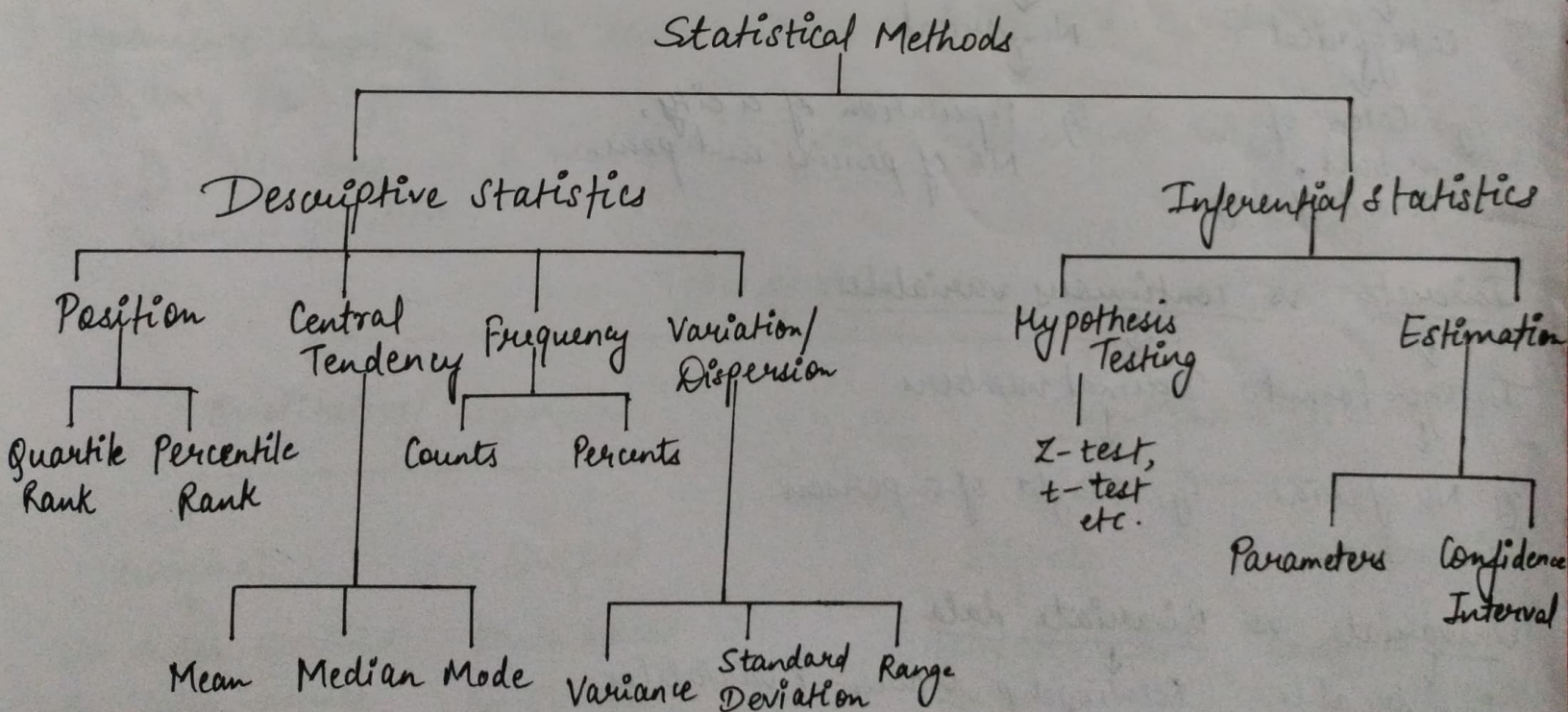
STATISTICS

Statistics is the branch of Mathematics where we collect, organize, analyse and represent the data for better decision making. We apply statistics to different problems.

Statistics refers to a scientific approach used to:

1) Collect data
2) Interpret and Analyze data
3) Assess the reliability of conclusions based on sample data.

For Example: Let us consider that a researcher wants to know which medicine is more effective - A or B, based on the symptoms, then, using the statistical methods, the researcher might conclude with 95% confidence that one medicine was superior to the other.



Descriptive statistics: It is a summary that describes or summarizes the collection of information/data. It summarizes the sample data rather than learning about the population that sample data is representing.

Inferential statistics: It is the process of data analysis where we make the conclusions about population data using sample data.

Population: It is the entire group that we want to draw the conclusions about.

Sample (n): It is a small part of population.

Variables in Statistics:

Qualitative vs. Quantitative Variables:

- Qualitative → Categorical → Eg: Color of a ball, Breed of dog.
- Quantitative → Numerical → Eg: Population of a city, No. of pencils and pens.

Discrete vs continuous variables:

- Discrete → Integer format → Eg: No. of coins in a box.
- Continuous → Decimal numbers → Eg: Weight of a person.

Univariate vs. Bivariate data:

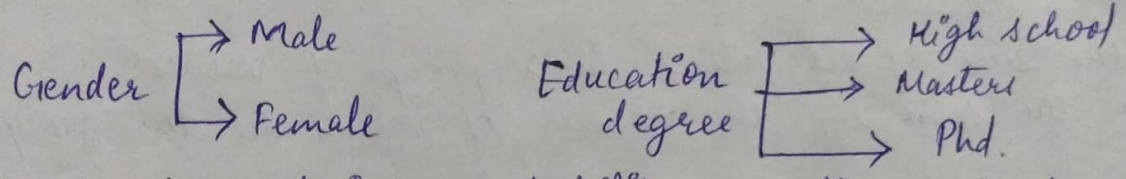
- Univariate → One Variable → Eg: Average weight of all students
- Bivariate → Relationship between 2 variables → Eg: Relationship between height and weight of students.

Sampling Techniques:

1] **Simple Random Sampling**: Every member of the population (N) has an equal chance of being selected for your sample (n).

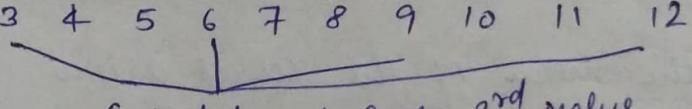
2] **Stratified Sampling**: Population (N) is split into non-overlapping groups.

Eg:



3] **Systematic Sampling**: It is a probability sampling method where researchers select members from population at n^{th} interval.

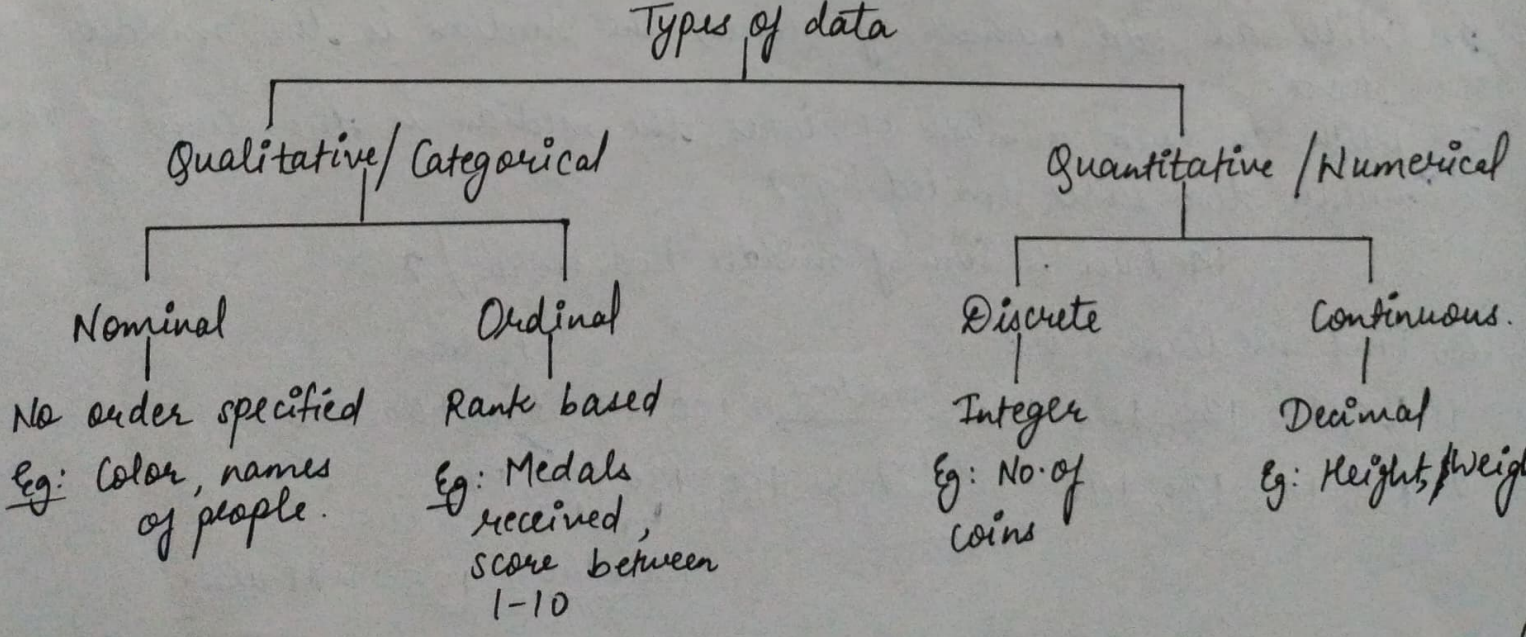
Eg:



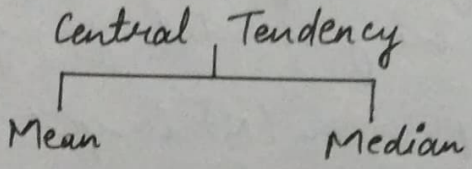
Get data of every 3rd value.

4] **Convenience sampling**: Only those who are interested in the survey will only participate.

Eg: Standing at a mall or a grocery store and asking people to answer questions.



Measures of Central Tendency: Central Tendency is the statistical measure that identifies a single value as a representative of an entire distribution.



To compute the mean:

1. Find the sum of all values in a group of values.
2. Divide the sum by the number of values in the group.

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

μ - Population mean $\quad$ $\bar{X}$ - Sample mean.

Q. Compute the mean of the scores below

100, 170, 130, 140

$$\mu = \frac{540}{4} = \underline{135}$$

To compute the median:

1. List scores from smallest to largest.
2. With an odd number of scores, the median is the middle score.
3. With an even number of scores, the median is the sum of the middle two scores divided by 2.

Median = Sum of middle two scores / 2

Q. Find median score

1. 100, 170, 130, 140, 160 $\xrightarrow{\text{sorting}}$ 100, 130, 140 (Median), 160, 170
2. 100, 170, 130, 140 $\xrightarrow{\text{Sorting}}$ 100, 130, 140, 170

$$\frac{130+140}{2} = \underline{135} \rightarrow \text{Median}.$$

As measures of central Tendency, the mean and the median each have advantages and disadvantages.

- The median may be a better indicator of the most typical value if a set of scores has an outlier. An outlier is an extreme value that differs greatly from other values.
- However, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency.

Eg: If we consider annual income of 4 people to be 7,00,000, 8,00,000, 9,00,000 & 1,00,00,000, then the 4th value will be an outlier. If we choose a measure to estimate the annual income of a ~~typical~~ person in general, the mean will greatly over-estimate the income (because of the outlier); while the median will not.

Mode: It is the most frequent score in the dataset. Normally, the mode is used for categorical data where we wish to know which is the most common category.
A problem with mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the dataset.

Measures of dispersion: Statisticians use summary measures to describe the amount of variability or spread in a set of data. The most common measures of variability are the range, the interquartile Range (IQR), variance and standard deviation.

1. Range: It is the difference between the largest and smallest values in a set of values.

Eg: 2 4 9 5 7 3 ⟶ Range = 9 - 2 = 7

It is easy to calculate but one drawback is, it ignores the middle values.

Eg: 1 2 3 4 5 6 7 8 9 → Range = 9-1 = 8

1 1 1 1 1 1 1 1 9 → Range = 9-1 = 8

Range is same in both cases, but there is a huge difference between the distribution of the data points.

2. Inter Quartile Range (IQR): It is a measure of variability, based on dividing a dataset into quartiles.

What is a quartile?

⇒ Before understanding quartiles, we need to understand percentiles.

Percentile: A percentile is a value below which a certain percentage of observations lie.

Eg: 99 percentile – It means a person has performed better than 99 % of the people.

$$\text{percentile} = \frac{\text{No. of values below a given no}}{\text{Total no. of values}}$$

Eg: $x = \{1, 2, 3, 4, 5, 5, 7, 20, 11, 10\}$

$y = 10$

$$\text{percentile} = \frac{7}{10}$$

Quartile: 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 and Q_3 respectively.

Relationship between quartiles and percentiles.

$Q_1 - P_{25}$, $Q_2 \rightarrow P_{50}$, $Q_3 \rightarrow P_{75}$. Q_2 is the median value in a dataset.

The InterQuartile Range (IQR) is a measure of variability, based on dividing a dataset into quartiles. The advantage of IQR over Range is that it takes all the values of dataset into consideration.

3. Variance : It is the average squared deviation from the population mean.

$$\sigma^2 = \Sigma (X_i - \mu)^2 / N$$

where, σ^2 = population variance
μ = population mean
X_i = ith element from population
N = no. of elements in population.

For sample variance, $s^2 = \Sigma (x_i - \bar{x})^2 / (n-1)$

Why sample variance is divided by n-1 ?

⟹

Population (N)	Sample (n)
$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$	$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu_x)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

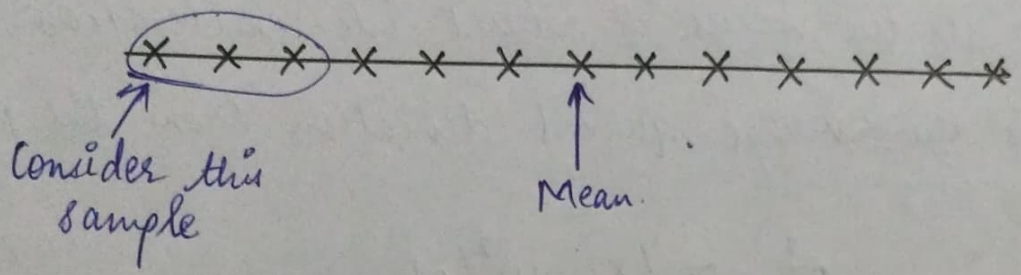
When we take out any sample, we need to infer some information from it so that we can draw some conclusions about the population data. This is possible only when we have

Population mean ≈ Sample mean

and

Population variance ≈ Sample variance

Consider the below example,



Now, when we calculate, we get ~~Rep~~ huge difference between the population mean and sample mean and the same with population variance and sample variance.

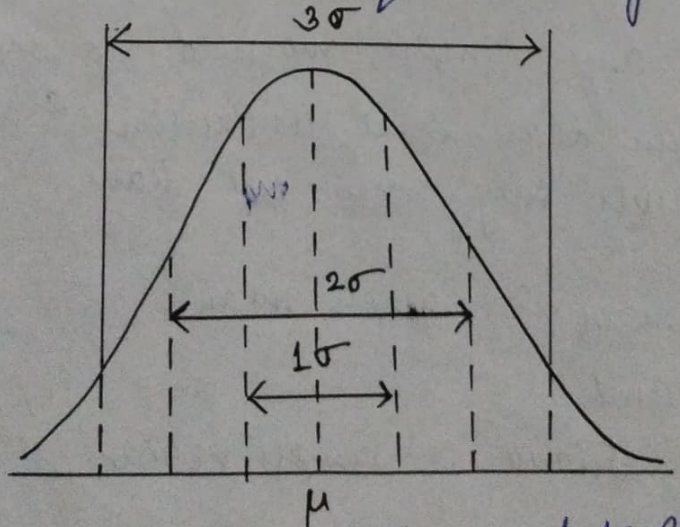
So, after a lot of research, ~~that~~ it was found out that instead of having 'n' in the denominator of sample variance, if we have n-1, n-2, n-3, ... then

Sample ~~mean~~ variance $\approx$ Population variance

And hence, it was found that "n-1" gave the closest answer ~~when~~ as compared with n-2, n-3..

Hence, sample variance is divided by n-1.

4. <u>Standard Deviation</u>: It is the square root of the variance.

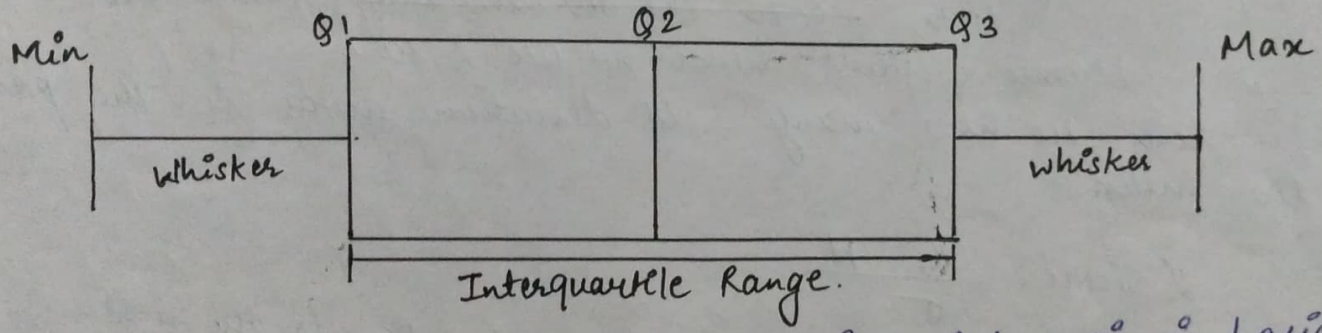


This is also used to identify whether a data point is an outlier or not.

Five Number Summary (In order to remove outliers):

1. Minimum
2. First Quartile (25 percentile - Q1)
3. Median (50 percentile - Q2)
4. Third Quartile (75 percentile - Q3)
5. Maximum

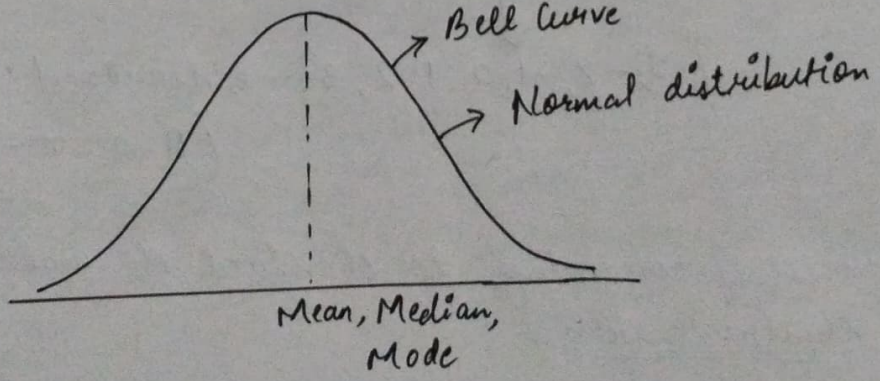
$$IQR = Q3 - Q1$$
$$\text{Lower Fence} = Q1 - 1.5*(IQR)$$
$$\text{Higher Fence} = Q3 + 1.5*(IQR)$$



The plot shown above is called the Box plot. This is basically used to get the outliers. Whatever points lie beyond this box plot are treated as the outliers.

Distributions:

① Gaussian / Normal Distribution:



Empirical formula - (68-95-99.7% rule) :

68% of data lies in 1st std. deviation.
95% of data lies in 2nd std. deviation.
99.7% of data lies in 3rd std. deviation.

This rule is applied on Gaussian/Normal distribution.

Z-score :

Let us consider, $\mu = 4$ & $\sigma = 1$

When we say 4.5, it means +0.5 std. deviation away from mean.

~~4.75, +0.75 std. deviation away from mean~~ calculation becomes difficult, hence we use Z-score.

Z-score tells us how many std. deviations away is the point from the mean.

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

$$= \frac{4.75 - 4}{1} = 0.75 \text{ std. deviation to the right.}$$

Dataset - {1, 2, 3, 4, 5, 6, 7}

⇓

Z-score

⇓

{-3, -2, -1, 0, 1, 2, 3} → Standard Normal distribution.

($\mu = 0$, $\sigma = 1$)

The process of converting to standard Normal distribution is called Standardization.

Normalization: $\{\mu = 0, \sigma = 1\}$

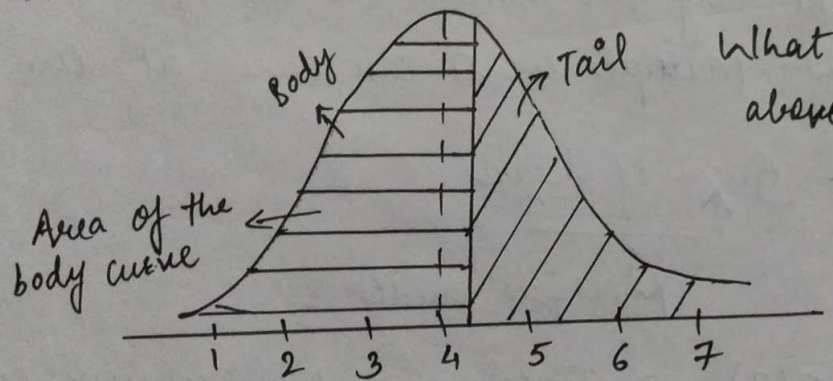
If we want to shift the values between a particular range then we use normalization. For example, between 0 to 1.

Min Max Scaler: (0 to 1)

Where to use Normalization?

⇒ In deep learning, we have images which have pixels which range between (0-255) where we do normalization and bring it in between 0 to 1.

Q. What percent of data lies above 4.25?



$$\Rightarrow \quad Z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Z-score: For area in a left tail, we need to look at the left tail z-table. Or 1 − right Area

$\therefore$ Area of tail = 1 − 0.5987 = 0.4013

i.e. 40% of data lies above 4.25

Probability: It is a measure of the likelihood of an event.

Eg: Flipping a coin - {H, T}

Probability of head or tail = $\frac{1}{2}$

Addition Rule (Probability, "or")

Mutual Exclusive Event: Two events are mutually exclusive if they cannot occur at the same time.

Eg: Rolling a dice {1, 2, 3, 4, 5, 6}

We cannot get any two numbers at the same time.

Non Mutual Exclusive Event: Multiple events can occur at the same time.

Eg: Deck of cards - Queen of hearts

↑ ↑

↑

Multiple events.

$P(A \text{ or } B) = P(A) + P(B)$ ⟶ In case of mutual exclusive event.

$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$ → In case of non-mutual exclusive event.

Multiplication Rule:

Independent Events:

Eg: {1, 2, 3, 4, 5, 6}

First time I can get 1, next time 3 and so on.

Each and every event is independent of each other.

Dependent Events:

Eg: Bag of 3 Red marbles and 2 green marbles

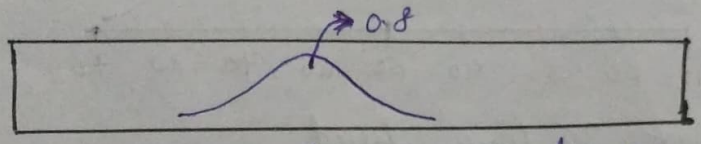
So picking out any marble will have impact on ~~on~~ probability of other marble.

Permutation and Combination:

Permutation: $^{n}P_{r} = \frac{n!}{(n-r)!}$

Combination: $^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$

p Value:



Suppose, this is a space bar of a laptop. The area marked is the area where a person touches the most. Suppose out of 100 times, the person touches 80 times in the place shown above.

p-value is the probability that a particular statistical measure, such as the mean or standard deviation, of an assumed probability distribution will be greater than or equal to observed results.

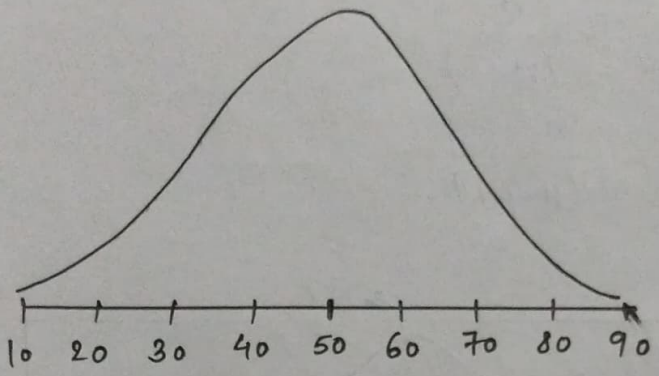
Hypothesis Testing:

Consider an example of tossing a coin 100 times. If we get head for 50 times then we say that the coin is fair.

Null Hypothesis: Coin is fair

Alternate Hypothesis: Coin is unfair.

Let, the standard deviation be 10.



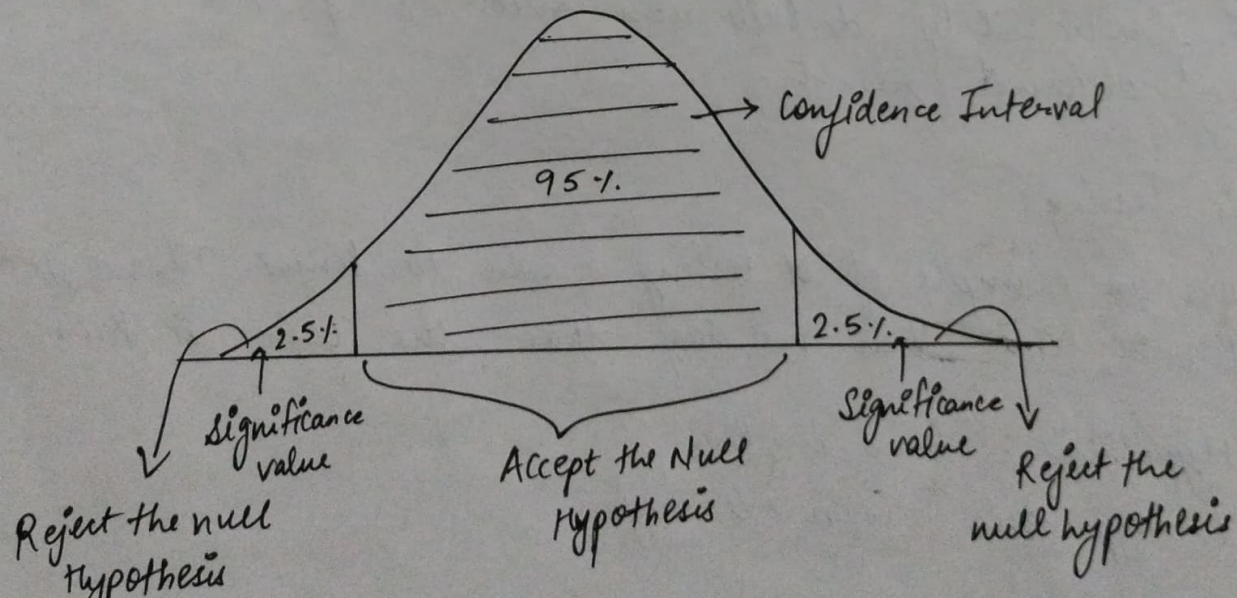
1] Suppose we got 30 times head

In order to prove coin is fair, we need to have the value nearer to the mean.

Now, how much away can the value be from mean is given by Significance value.

Let $\alpha = 0.05 \rightarrow$ Defined by domain expert.

$1 - 0.05 = 0.95 \rightarrow$ Confidence Interval.



Significance value $\neq$ P-value

Type 1 and Type 2 error:

Null Hypothesis (H_0)

Alternate Hypothesis (H_1)

Outcome 1: We reject the Null Hypothesis, when in reality it is false → Yes, good decision.

Outcome 2: We reject the Null Hypothesis, when in reality it is true → No, Bad decision.

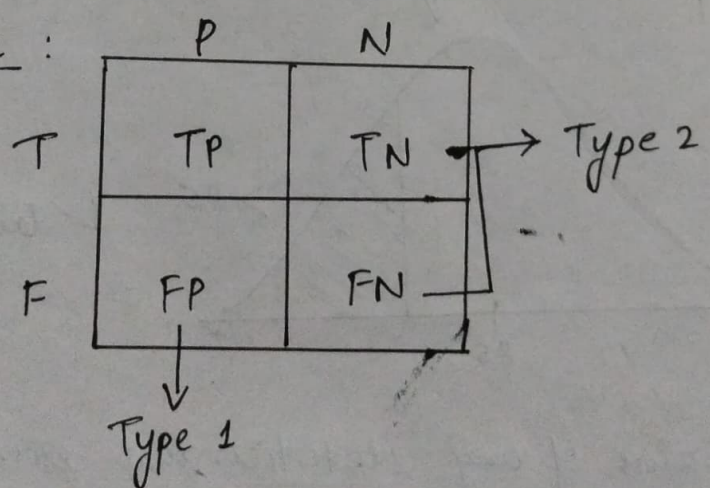
Type 1 error

Outcome 3: We retain the Null Hypothesis, when in reality it is False → NO, Bad Decision

Type 2 error

Outcome 4: We retain the Null Hypothesis, when in reality it is True → Yes, Good decision.

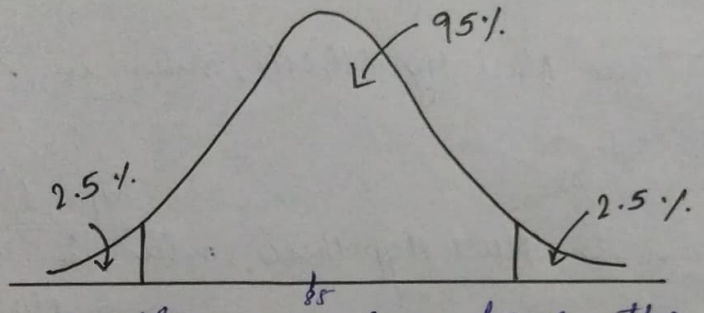
Confusion Matrix :



1 Tail and 2 Tail test :

Q. Colleges in Karnataka have 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%, with a standard deviation of 4%. Does this college have a different placement rate than other colleges? $\alpha = 0.05$

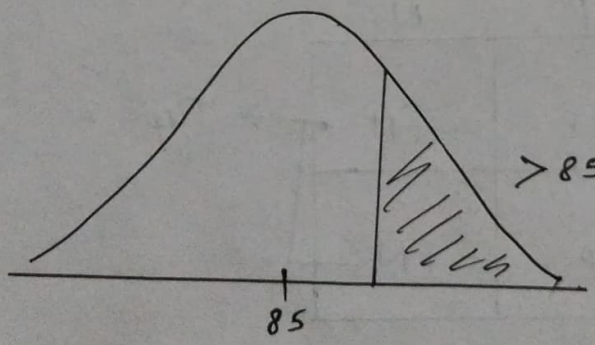
⇒ 2-Tailed test



2-Tail test

From the given question, we can have the placement rate either greater than 85 or less than 85, that is why it is a 2-tailed test.

Suppose, we have a question as greater than 85%, then it will be a 1-tail test.



1-Tail test

Point Estimate : The value of any **statistics** that estimates the value of a **parameter**.

$\bar{x}$ ⟶ μ

$\bar{x}$: Sample mean

μ : Population mean

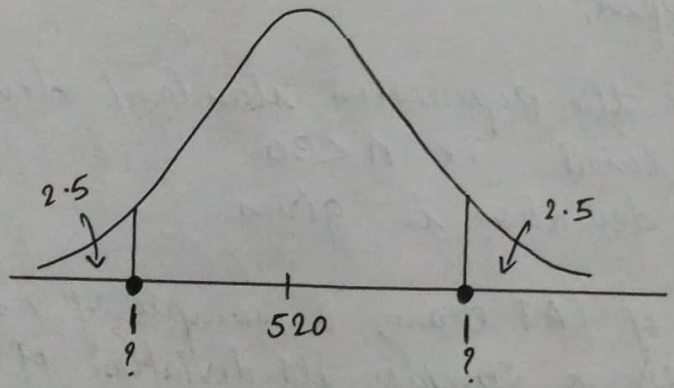
Through $\bar{x}$, we are estimating μ, this is called point estimate.

Confidence Interval:

Point Estimate ± Margin of Error.

Q. On the quant test of CAT exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean.

⟹ $\sigma = 100$, $n = 25$, $\alpha = 0.05$, $\bar{x} = 520$



① When population standard deviation is given, we apply z-test

Point Estimate ± Margin of error

$$\bar{x} \pm Z_{\alpha/2}\left[\frac{\sigma}{\sqrt{n}}\right] \rightarrow \text{Standard error.}$$

For a Z-test to happen,

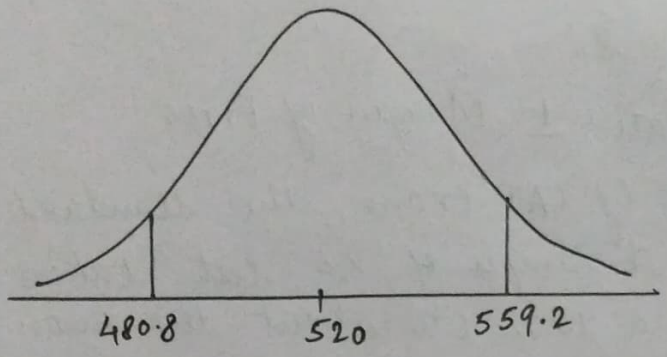
① We know the population standard deviation or
② We do not know our population standard deviation but our sample size ≥ 30.

Upper bound - $\bar{x} + \frac{Z_{0.05}}{2} \frac{100}{\sqrt{25}}$ $= 1 - 0.025 = 0.975 \xrightarrow{\text{Z-table}} 1.96$

$= 520 + (1.96)(20) = 559.2$

Lower bound - $\bar{x} - \frac{Z_{0.05}}{2} \frac{100}{\sqrt{25}}$ $= 520 - (1.96)(20) = 480.8$

→ CI Interval



For a t-test to happen,

① We do not know the population standard deviation.
② Sample size is small i.e. $n < 30$.
③ Sample standard deviation is given.

Q. On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a sample std. deviation of 80. Construct 95% confidence interval about the mean.

⇒ $n = 25$, $\bar{x} = 520$, $S = 80$, $\sigma = 0.05$

Population std. dev is not given ⟶ t-test

$$\bar{x} \pm t_{\alpha/2} \boxed{\left(\frac{S}{\sqrt{n}}\right)} \rightarrow \text{Standard Error}$$

Degree of freedom = $n - 1 = 25 - 1 = 24$

Upper bound = $520 + 2.064\left(\frac{80}{5}\right) = 553.024$

Lower Bound = $520 - 2.064\left(\frac{80}{5}\right) = 486.97$

$[486.97 \rightarrow 553.024]$

Standard Error: It indicates how different the population mean is likely to be from a sample mean.

Chi Square Test: It claims about population proportions.

It is a Non-parametric test that is performed on categorical (Nominal or Ordinal) data.

Parametric vs Non-Parametric:

Parametric statistics are based on assumptions about the distribution of population from which the sample was taken. Non-Parametric statistics are not based on assumptions, that is, the data can be collected from a sample that does not follow a specific distribution.

Q. In the 2000 Indian Census, the age of the individuals in a small town were found to be the following:

Less than 18	18-35	>35
20%	30%	50%

In 2010, age of n=500 individuals were sampled. Below are the results.

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the population distribution of ages has changed in the last 10 yrs?

⇒

<18	18-35	>35	
121	288	91	→ Observed.] Huge difference
500×0.2 = 100	500×0.3 = 150	500×0.5 = 250	→ Expected

H_0 = The data meets the distribution of 2000 census.

H_1 = The data does not meet the distribution of 2000 census.

$\alpha = 0.05$ (95% CI)

Degree of Freedom = $n - 1 = 3 - 1 = 2$ (This is a 2-Tail test)

↓ No of categories

Chi-square table.

If χ^2 is greater than 5.99, then reject H_0.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.494 > 5.99 \rightarrow \text{Reject } H_0.$$

∴ Conclusion is, the data does not meet the distribution of 2000 Census.

If P-value < significance value

Reject the Null Hypothesis

Covariance :

X	Y
Weight	Height
50	160
60	170
70	180
75	181

$x \uparrow \; y \uparrow$

$x \downarrow \; y \downarrow$

Quantify relationship between X & Y ⟶ Covariance.

Covariance

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

N ⟶ In case of sample this is (n-1)

$x \uparrow y \uparrow$, $x \downarrow y \downarrow$] → +ve correlation

$x \uparrow y \downarrow$, $x \downarrow y \uparrow$] → -ve correlation

0] → No correlation

Covariance: No particular magnitude. We have direction but magnitude wise, it is not a good choice. So, we move to Pearson's correlation.

Pearson Correlation Coefficient: It restricts the value between (-1, 1). The more towards +1, the more it is positively correlated and the more towards -1, the more it is negatively correlated.

$$\rho(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

This captures the linear properties well.

Spearman Rank correlation:

$$Spear(x, y) = \frac{Cov(R(x), R(y))}{R\sigma_x R\sigma_y}$$

Height (x)	Weight (y)	R(X)	R(Y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

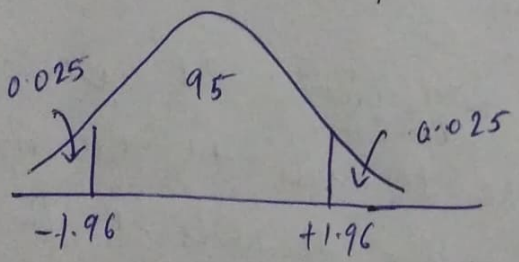
We use this correlation because it ~~has~~ supports Non-Linear properties.

Q. The average weight of all residents in Bangalore city is 168 pounds with a standard deviation of 3.9. We take a sample of 36 individuals and the mean is 169.5 pounds. C.I = 95%.

$\Rightarrow$ $H_0 = \mu = 168$

$H_1 = \mu \neq 168$

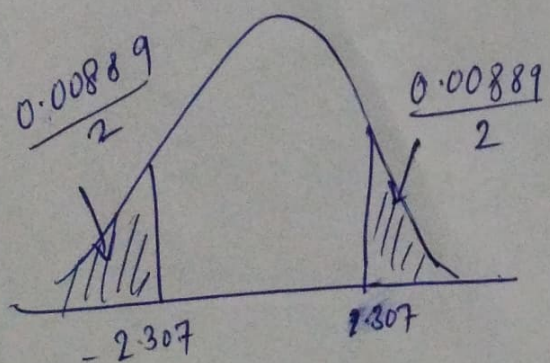
$\alpha = 0.05$



We use Z-test here.

$$Z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}} = \underline{\underline{2.307}}$$

$2.307 > 1.96 \longrightarrow$ We reject the null hypothesis.



Z-table for 2.307 $\rightarrow$ 0.099111

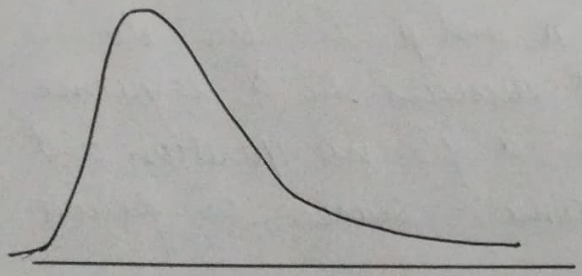
$1 - 0.99111 = 0.00889$

P-value = $0.0044 + 0.0044$

$= \underline{\underline{0.0088}}$

$0.0088 < 0.05 \longrightarrow$ We reject null hypothesis.

Log Normal Distribution:



Eg: 1) Wealth Distribution
2) People writing big comments.

If $X \approx$ Log Normal distribution

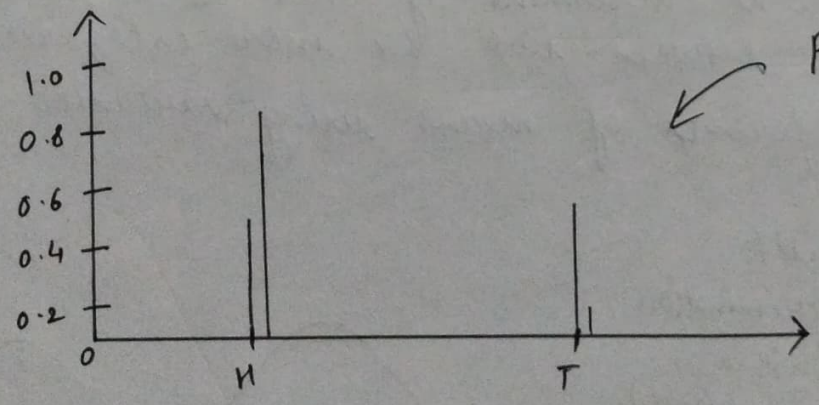
then

$Y \approx \ln(X) \Rightarrow$ Normal distribution

If a random variable X is log-normally distributed, then $Y = \ln(X)$ has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, $X = \exp(Y)$, has a log-normal distribution.

Bernoulli's distribution: It is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$.

It has only 2 outcomes → Either 0 or 1.



Probability Mass function (pmf)

For categorical variables.

PMF $q = 1 - p$ if $k = 0$

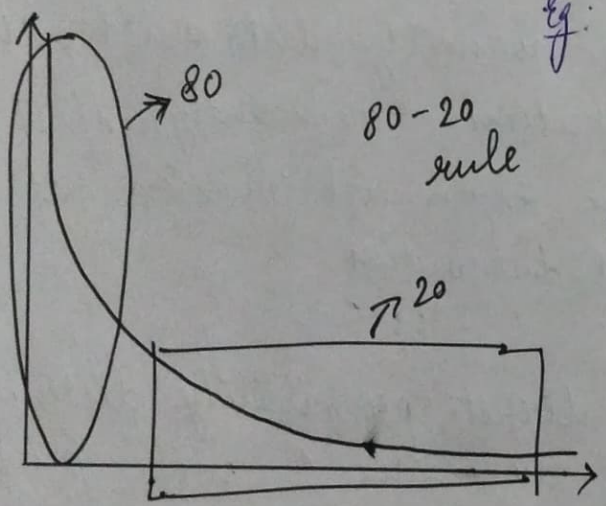
p if $k = 1$

$p^k (1-p)^{1-k}$

Binomial Distribution :

The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question and each with its own Boolean-valued outcome : success or failure.

Pareto Distribution: → Also called as power laws.



Eg: 1) 80% of wealth is distributed with 20% of people.

2) 80% of sales is done by 20% of most famous product.

Anova Test (Analysis of Variance): An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.