**Statistics**
Statistics is the field of mathematics that deals with the data
Collect , Analyse …

**Data** is information
→ **Qualitative** (categorical data like id , name )
→ **Quantitative**(Numeric data ) ⟶ classified into Discrete and Continuous data

**Random** → each of the members in the population has an equal chance of being selected in the sample.

**Frequency distribution**
A list of values with corresponding frequencies.
A frequency distribution is a representation, either in a graphical or tabular format, **that displays the number of observations within a given interval**.

**Relative frequency** distribution is the percentage of ⇒ $frequency/n\ *\ 100$
**Cumulative frequency** distribution ⇒ Add sequential classes together or add frequencies one by one, final raw we will get $n$

**Measure of Central Tendency**

Mean $\mu(population\ mean) = \frac{1}{N}\sum_{i=1}^{n} x_i$ , sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Median ⇒data at $middle$ (the data must be ordered)
Mod ⇒ most repeated data

**Variation** : How the data is spread
**Ways to measure the variation**
1. **Range** = $Max - min$ , it's easy to find but doesn't consider all values.
2. **Standard deviation** , it is the average distance the data points are from the mean, never a negative value and zero unless all entries are the same.

   Sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{i=n}(x-\bar{x})^2}{n-1}}$ variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$

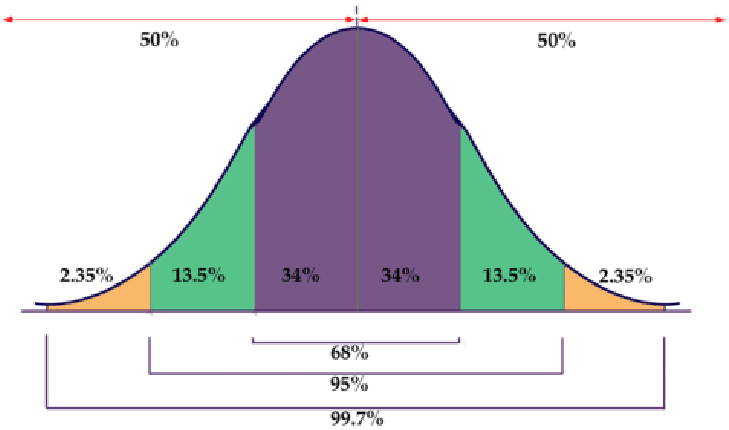
Population standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x-\mu)^2}{N}}$ variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x-\mu)^2$

Another equation for $s^2 = \frac{n(\sum_{i=1}^{n}(x^2)) - (\sum_{i=1}^{n}x)^2}{n(n-1)}$

Closely ground data will have small standard deviation

Spread-out data will have larger standard deviation

If the data is **normally distributed** we can use the **Empirical Rule**




68% of data will fall one standard deviation(1s) of the mean i.e, mean-s---mean---mean+s

95% of data will fall within 2s

99.7% of data fall will fall 3s

Coefficient of variation $c.v = \frac{s}{\bar{x}}$

Multiple variables(x,y): **covariance** $cov_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

**Measures of relative standing** :comparing measures between or within data set

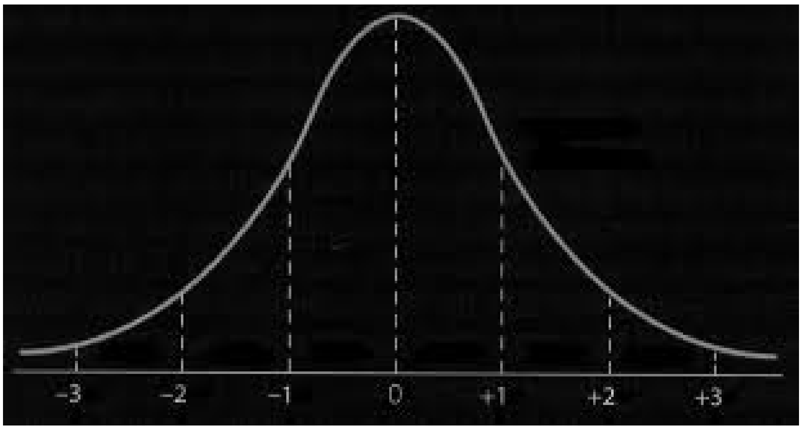
**Z-score** : The number of standard deviations that a specific data value(x) is away from the mean , *allows us to compare the variation in two different samples or population.*

(sample) $Z = (x - \bar{x})/s$ s→ standard deviation , $\bar{x}$→ mean

(population) $Z = (x - \mu)/\sigma$ σ→ standard deviation,μ→ mean

(**how many standard deviation that x is away from the mean**)

The **Z-score between -2 and 2 is considered "usual"**(that is within 95%(2s)) ; outside of -1 and 2 is "unusual".




**Quartiles**

**Q1**⇒ bottom 25% of sorted data,

**Q2**(Median M)⇒ bottom 50% of sorted data,

**Q3**⇒ bottom 75% of sorted data

Percentile of $x = \frac{number\ of\ elements\ less\ than\ x}{total\ number\ of\ values} * 100$

**Probability 0 <= P(x) <= 1**

Event : A collection of outcomes of a procedure.

Simple event : A single outcome.S={all possible outcomes }.

Sample space : All the single events(Every possible outcome).

**Probability** : The likelihood of an event occuring

(classical probability) $p(A) = preferred\ outcome\ /\ sample\ space$

**Experimental Probability/(**observed**)** $p(A) = successful\ trails\ /\ all\ trails.$

Subjective Probability : is an educated guess

**Conditional probability** $p(A|B) = \frac{p(A \cap B)}{p(B)}$ , $p(A|B) \rightarrow$ (probability A given B).

**Additive rule** $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

**Multiplication rule** $p(A|B) * p(B) = p(A \cap B)$

Probability of complementary events "At least one"

$P(A) + p(A') = 1, \quad P(A) = 1 - P(A')$

E.g probability of at least one head when flipping 3 coins

$S = \{HHH, HHT, HTH, HTT, TTT, TTH, THT, THH\}$

$p(at\ least\ 1H) = \frac{7}{8}$(from sample space only TTT is without H)

According to $P(A) = 1 - P(A')$

$$P(at\ least\ 1H) = 1 - P(not\ getting\ heads) = 1 - P(1/8) = 1 - 1/8 = 7/8$$

**Bayes Theorem** $\Rightarrow$ $P(A|B) = P(B|A) \,.\, P(A) \,/\, p(B)$

**Permutation and Combination**

Permutation : How to Arrange

$nP_r = n!\,/(n - r)!$

n$\longrightarrow$total number of objects, r$\longrightarrow$ number of objects selected

Non-distinct elements $\Rightarrow$ $nP_r = n!\,/(n_1! * n_{2!} * ...)$

$n_1, n_{2!}, n_{3..}$ $\longrightarrow$ count of non distinct items (means repeating)

Combination:How to pick

$nC_r = n!\,/(n - r)!\,r!$

**Discrete Probability:** Discrete means means countable

**Random Variable** : A variable x that has a value for each outcome of a procedure that is determined by chance.

**Probability Distribution** : A table that gives the probability for each value of a random variable.

E.g probability distribution table for rolling a die

| x | P(x) |
|---|---|

| | |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |

Mean $\mu = \frac{1}{N}\sum(x.f) \Rightarrow \sum(x \,.\, f \,/\, N) = \sum(x.P(x))$ (because $f/N$ is $P(x)$ )

Mean $\mu = Expected\ value\ E(x) = \sum_{all\ x}(x.P(x))$

Variance $\sigma^2 = \sum(x^2.P(x)) - \mu^2$ or $\sigma^2 = E[(x - \mu)^2] = \sum((x - \mu)^2 \,.\, P(x))$

Standard deviation $\sigma = \sqrt{\sigma^2}$: The avg distance from the mean

**Binomial Probability Distribution** : two outcomes → success or failure

1. Must be fixed number of trails
2. Trails must be independent(outcomes does not dependent )
3. Each trails have only 2 outcomes
4. The probability of success remains the same in all trails

n→ # of trails

p→ The probability of a successful outcome in a single trail(one single success)

q→ The probability of a failing outcome in a single trail

x→ The # of success that occur in the n trails(number of success you looking for)

$$P(x) = nCx \,.\, p^x \,.\, q^{n-x} \Rightarrow P(x) = nCx \,.\, p(x)^x \,.\, (1 - p)^{n-x}$$

**Mean , variance and standard deviation of binomial distribution**

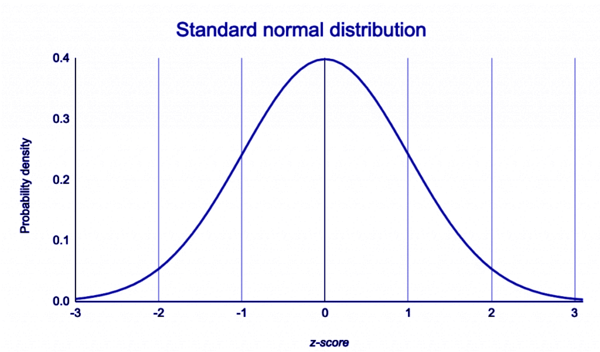
**Mean** : number of success you expected to occur from your procedure

$\mu = n \,.\, p$ (product of total number of trails and probability of success for each one)

**Variance** : $\sigma^2 = n \,.\, p \,.\, q \Rightarrow n \,.\, p \,.\, (1 - p)$

**Standard Deviation** : $\sigma = \sqrt{n \,.\, p \,.\, q}$

**Continuous probability distribution** : measurements , can't countable




**Standard Normal distribution**

$\mu = 0$ , $\sigma = 1$ , $Area\ under\ curve = 1$

$$Z = \frac{x - \mu}{\sigma}$$

**x→ continuous random variable**
**Z is the normal distribution , it's also z-score**
**Area under the curve is = 1**

The probability of a continuous random variable x at that point is always zero , because a single dimensional line has no area , the probability of less than x is area under the curve till x , i.e we can find probability between two points of less than or greater than . if a and b are two points then the probability is area under the curve i.e, $\int_a^b f(x)\,dx$ $\quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

**Less than x** ⇒ area between -10 and x
(-10 just a negative value theoretically it is -∞)
**Greater than x** ⇒ 1- area between -10 and x (because complement)

Points

- Z -score is the distance from the mean (how much $\sigma$ away from the mean )
- Area is a probability (cannot be -ve)
- Z -score can be negative

**Sampling Distribution** : Use a sample to estimate a population.
Assume you take all the possible samples of size 'n' and find the required statistic for each sample , if you organize all of those statistics of each different sample into a table this is a sampling distribution .

P → population proportion
$\hat{p}$→ sample proportion( p hat)

**Central Limit Theorem**

1. n>30 then the sampling distribution of sampling means is normally distributed, $\mu_{\bar{x}} = \mu \, and \ \sigma_{\bar{x}} = \sigma/\sqrt{n}$

   (the avg of sample mean must equal to population mean)
   (standard deviation of sample means is standard error)

2. n<=30 and the population is normally distributed, then the sampling distribution of sample mean is also normal distribution,

   $\mu_{\bar{x}} = \mu \, and \ \sigma_{\bar{x}} = \sigma/\sqrt{n}$

   $\sigma_{\bar{x}} = \sigma/\sqrt{n} \rightarrow$ is also known as standard error

   $Z = \frac{x-\mu}{\sigma} \Rightarrow Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ (this is useful for calculate

   z-score 1 ,2 , we use t-score for the 3rd condition )

3. n<=30 and you don't know nothing about the population, then Student's T-statistic

   **This is central limit theorem**

**Confidence interval :** estimating proportion with sample proportion
p^-E < P < p^+E.
(where E is the margin of error, the difference between p and p^)
Conditions : 1 random variable , 2 condition for binomial → fixed # of trails ,trails are independent ,two outcomes success/failure , np>=5,nq>=5
**Point Estimate:** A single value used to approximate a population.
P=population proportion of success.
p^=sample proportion of success = x/n
q^=sample proportion of failure = 1-p^

p^ is a point estimate for P

**Range** : is used to estimate a population parameter

**Confidence level($1-\alpha$):** How confident you are that the actual value of the population parameter will be inside the interval .

$1 - \alpha$ ⇒ $\alpha$ is the complement of a confidence level.

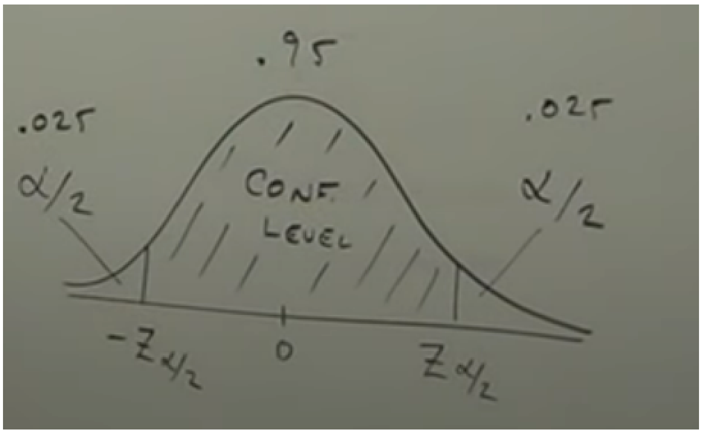
Most common confidence level : 90 , 95 , 99 ; most used confidence level : 95.

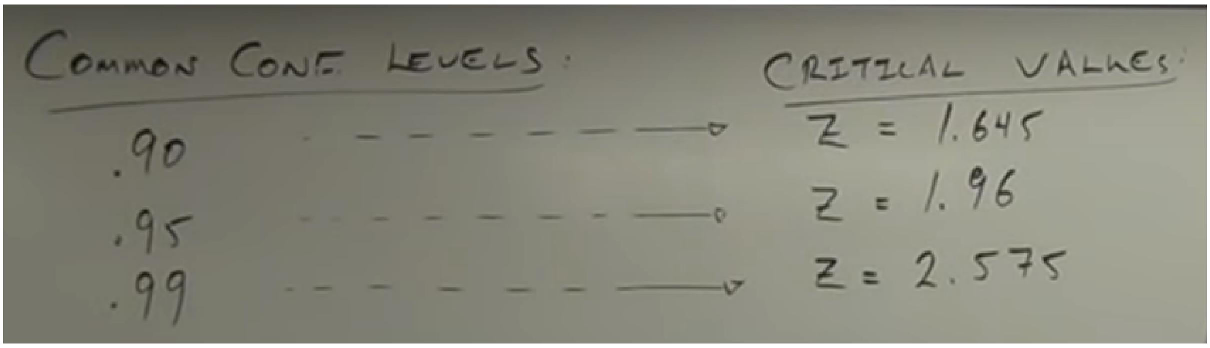
$\alpha$ For the 90 % = 0.10

$\alpha$ For the 95 % = 0.05

$\alpha$ For the 99 % = 0.01

**Critical value** $Z_{\alpha/2}$ is a Z - score that separates the likely region or unlikely region




**Margin of error E** : The maximum difference between p^ and P

$$E = Z_{\alpha/2}\sqrt{\frac{(\hat{p}\,.\,\hat{q})}{n}} \quad , \quad Z_{\alpha/2}$$ → critical value

**Confidence interval** p^-E < P < p^+E , or $P = \hat{p} \pm E$

**Finding required sample size:** Given an E , you can find the sample size needed to get the E.

$$E = Z_{\alpha/2}\sqrt{\frac{(\hat{p}\,.\,\hat{q})}{n}} \Rightarrow n = \frac{(Z_{\alpha/2})^2\,.\,\hat{p}\,.\,\hat{q}}{E^2}$$

The worst cause $n = \frac{(Z_{\alpha/2})^2\,(.25)}{E^2}$(if you don't know anything about proportion of success or proportion of failure then , assuming 50% for success and failure.

There for the p= 0.5 and q=0.5 and p.q=0.5*0.5=0.25 ).

Example 1) We want to determine the % of US people who use e-mail @ 95% confidence level . what does the sample size need to be to ensure an error of 4%.

- 16.9% of people used email in 1997

  n=? → want to find,

  p^=0.169 → sample proportion ,

  q^= (1 - p^ ) = 0.831 ,

  $Z_{\alpha/2} = 1.96$ (*because* 95% *confident* ) , $E = 0.4$(*because* 4% *error of margin* )

$$n = \frac{(Z_{\alpha/2})^2\,.\,\hat{p}\,.\,\hat{q}}{E^2} \Rightarrow$$

$$n = \frac{(1.96)^2.(0.169).(831)}{(0.04)^2} = 337.19 \simeq 338$$

**Given A C.I. Find p^ and E**

$$\hat{p} = \frac{(\hat{p} - E) + (\hat{p} + E)}{2} = \frac{upper+lower}{2}$$

$$E = \frac{(\hat{p} + E) - (\hat{p} - E)}{2} = \frac{upper-lower}{2}$$

**Estimate Population mean** μ

$\bar{x}$ is the point estimate for μ

$E = Z_{\alpha/2}\,.\,\frac{\sigma}{\sqrt{n}}$ ⟶ maximum difference between $\bar{x}$ & μ(Margin of error)

$\bar{x} - E < \mu < \bar{x} + E$ or $\mu = \bar{x} \pm E$(confidence interval)

$$E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{(Z_{\alpha/2}) \cdot \sigma}{E}\right)^2$$

**Estimate Population** $\mu$ , *unknown* $\sigma$

If You don't know $\sigma$, you can't use **Z-score** .Instead , we use a **t-score**

**t-Score**

1. Random sample
2. n>30 or sample is from a population that is normally distributed

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \Rightarrow t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$ s→ sample standard deviation

**Degrees of freedom** = n - 1 (degrees of freedom is sample size - 1)

**Critical value** $t_{\alpha/2}$

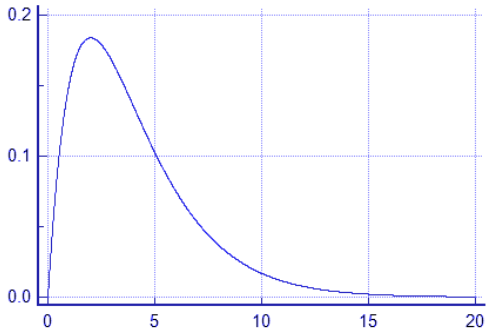
**Margin of Error** $E = t_{\alpha/2} \frac{s}{\sqrt{n}}$

Confidence interval : $\bar{x} - E < \mu < \bar{x} + E$ or $\mu = \bar{x} \pm E$

**Estimate Confidence interval with population variance**

$\chi^2$ **Distribution**

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$ n→ sample size , $s^2$→ sample variance ,$\sigma^2$→ population variance




- Values are not negative.
- As degrees of freedom go up , the distribution becomes more symmetric.
- Give critical values for the area to the right

$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \Rightarrow \sigma^2 = \frac{(n-1)s^2}{\chi^2}$

Critical value : $\chi^2_R, \chi^2_L$

Confidence interval : $\frac{(n-1)\, s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)\, s^2}{\chi^2_L}$

Standard deviation : $\sqrt{\frac{(n-1)\, s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)\, s^2}{\chi^2_L}}$

**Hypothesis Testing :** Testing whether or not a claim is valid

**Rare event rule** : If the probability of an assumption occurring is very small then the assumption is probably incorrect .

**Null Hypothesis & Alternative Hypothesis**

In statistics you can't prove anything right , but can prove it is false . That means you can't prove he is innocent , but can prove he is not guilty.

**Null Hypothesis** $H_0$**:**State that the population parameter (mean$\mu$,proportion P) is equal to some value.

Note : how to test a hypothesis → Start by assuming the $H_0$ is true .Then ,use evidence to reach a conclusion .

- **Reject** $H_0$: I have enough evidence to prove $H_0$is wrong
- **Fail to reject** $H_0$**:** I don't have enough evidence to prove $H_0$is wrong

**Alternative Hypothesis** $H_1$**:**State that the parameter (mean$\mu$,proportion P) has a value different than $H_0$. i.e, ($<$,$>$,$\neq$)

- If you want to support a claim , you must be state it as $H_1$(not $H_0$).

**How to identify** $H_0$ & $H_1$

- State the original claim symbolically i.e, equal to or > , <
- State the opposite of the original claim as well.

- Note : The original claim could be $H_0$ $or$ $H_1$,Depending on where the equality is.

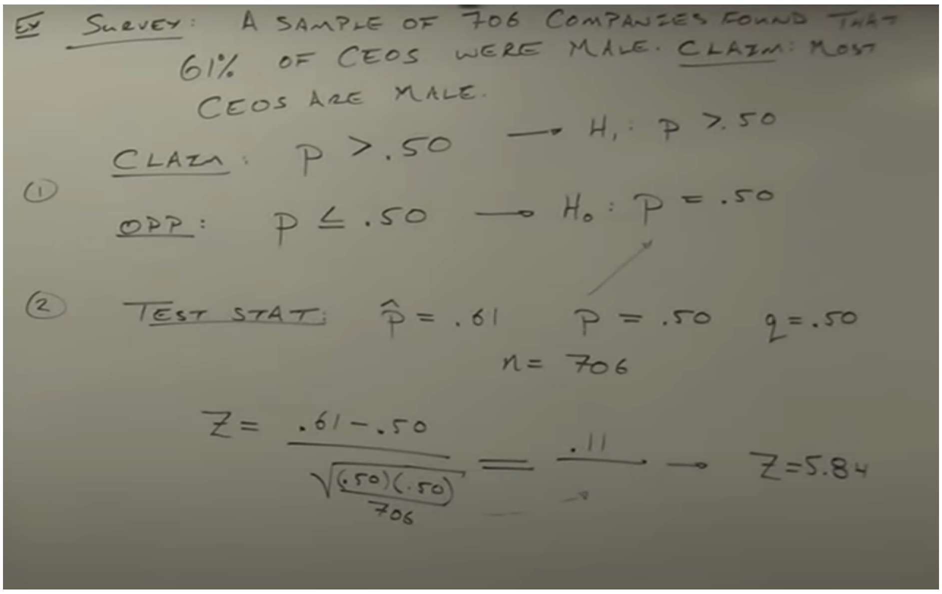
First read the question and state claim and opposite claim , which clime have an equal sign that is consider as $H_0$ .

**Test Statistic**

Use Z statistic for the proportion ; if question is about the mean then use Z for known population standard deviation , use t for unknown population standard deviation

Proportion P: $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{P.q}{n}}}$

Mean μ : $Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$  $t = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$



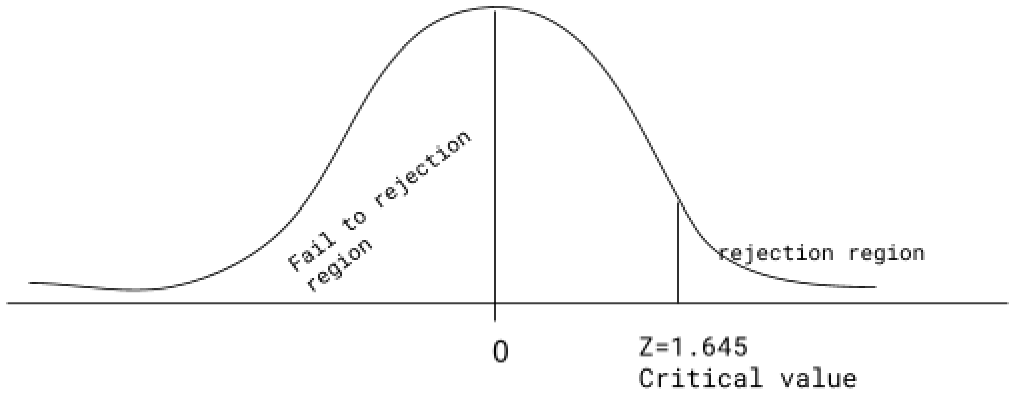

the answer is unusual , -2 to 2 is usual value ,$H_0$ $is\ wrong$

**How to make decision**

**Significance level** $\alpha$: $0.1, 0.05, 0.01$($most\ common\ significance\ level$)

**Critical values** : Separate rejection region from the fail to rejection region.

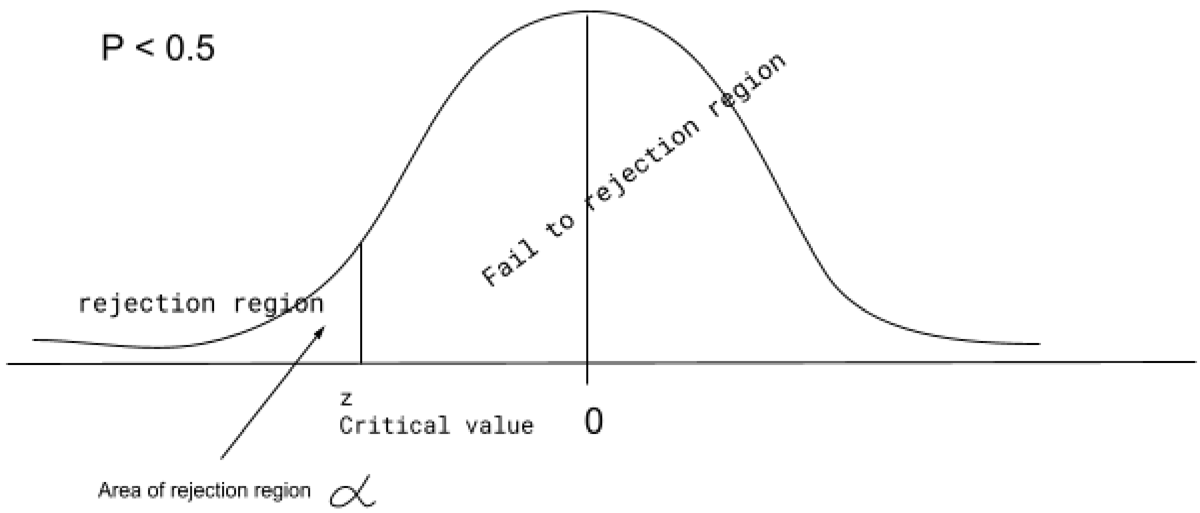
**Rejection region:** If our test statistic falls into these region reject $H_0$.



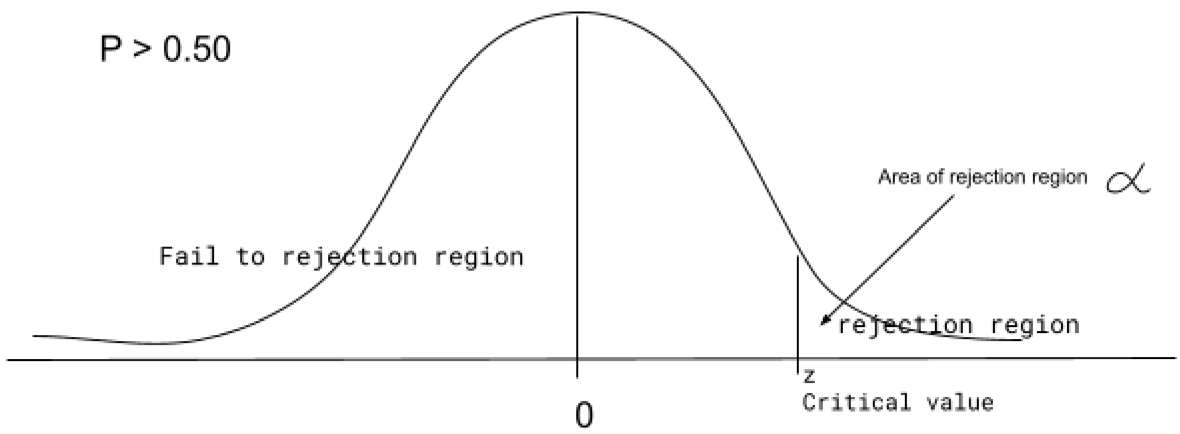

Note : You have 2 Z-score s ⇒ 1)critical value z 2)test statistic Z

**3 Type of test**

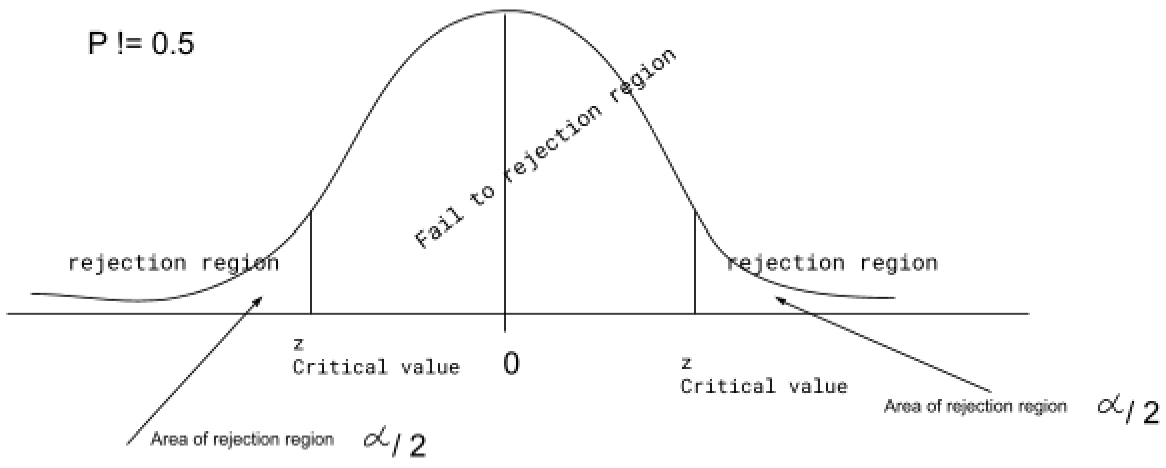
**If** $H_1 has\ <, then\ the\ rejection\ region\ is\ in\ left - tail$




**If** $H_1 has\ >, then\ the\ rejection\ region\ is\ in\ right - tail$

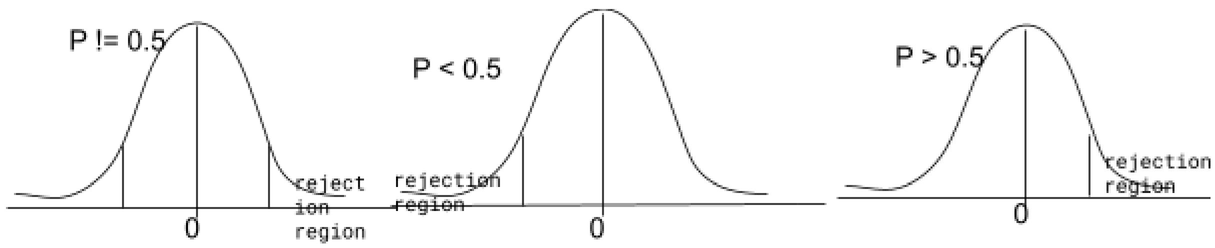



**If** $H_1 has\ \neq, then\ the\ rejection\ region\ is\ in\ both\ tails$




**Note:** 2 ways for hypothesis testing , the first way is the traditional method → using critical values and significance level (above methode ) and Second method is using P value ,in P-Value don't worry about the critical value here we use the z-statistic , significance level.

1)

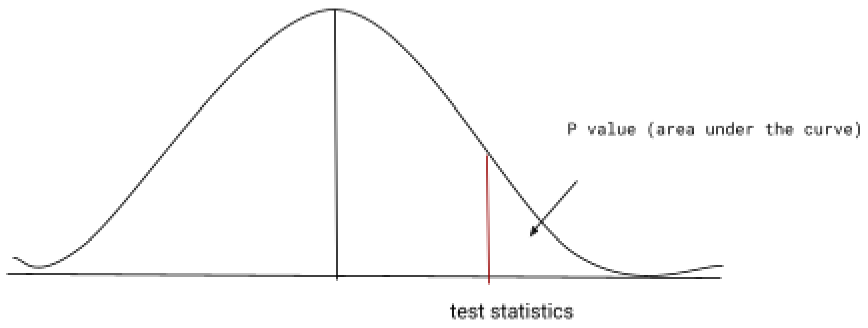



**2)**

**P-value : Probability value**

Probability associated with test statistics

P-value is the area of rejection region , α value is the significance level




Reject ⇒ $H_0$ If P-value <= α

Fail to Reject ⇒ $H_0$ If P-value >α