* Hypothesis Testing :-

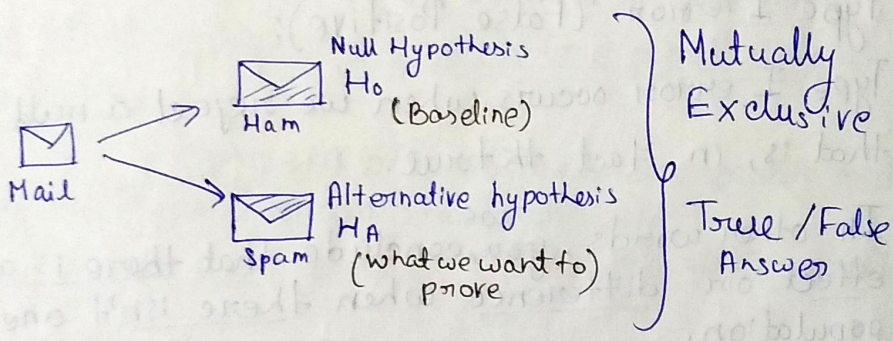
Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data.

Initially, a tentative assumption is made about the parameter or distribution. (H_0)

Also an alternative statement is also stated (H_A or H_1).

If we get enough evidence against H_0, we choose H_A.

Eg:-



→ To determine the result of test :-

Plenty of evidence against H_0 ⟶ Reject H_0 (and accept H_A)

* Type 1(I) & Type 2(II) error :-

Decision	Reality	
	H_0 True (Not spam)	H_0 False (Spam)
Reject H_0 (Decide spam)	Type I error	Correct
Don't Reject H_0 (Decide not spam)	Correct	Type II error

These are the 2 different ways in which errors can occur when conducting hypothesis testing.

① Type I error (False Positive):

→ Type I error occurs when we reject a null hypothesis that is, in fact, ~~th~~ true.

→ In other words, ~~you~~ we conclude that there is a significant effect or difference when there isn't one in the population.

→ The probability of making a Type I error is denoted by α (alpha), and is also called significance level.
Common choices of α are 0.05 (5%) or 0.01 (1%).

Setting α lower, we reduce chance of Type I error by increases risk of Type II error.
(There is a trade off)

(2) Type II error :- (False Negetive)

→ Type II error occurs when we fail to reject a null hypothesis that is actually false.

→ In this case, we conclude that there is no significant effect or difference when there is one in the population.

→ The probability of making a Type II error is denoted by β (beta)

* P-Value, Critical Value and Significance Level:

(1) Significance level :-

It is often denoted as α (alpha), is a predefined threshold that researchers set before conducting a hypothesis test.

It represents the maximum acceptable risk of making a Type I error (False poisitive).

(Mostly $\alpha = 0.05$ or 0.01)

(2) P-Value :-

The p-value, on the other hand, is a statistical calculated from the sample data, representing the stregth of evidence against the null hypothesis.

Researchers use the p-value to determine whether the null hypothesis should be rejected based on the chosen significance level.

→ While α is fixed and determined by the researchers, the p-value is data-dependent and can vary with different datasets.

③ Critical Value:-

A critical value, also known as a critical score or cutoff value, is a pre-determined ~~thes~~ threshold or specific value from a probability distribution that is used in statistical hypothesis testing. and corresponding ~~the~~ to the selected significance level (α).

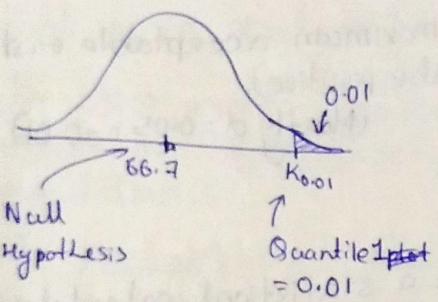
It represents the point beyond which we would reject the null hypothesis.

(It's usually found in statistical tables or can be calculated using software)

Eg:- $H_0: \mu = 66.7$ vs $H_0: \mu > 66.7$

$\alpha = 0.01$

Given, $n = 10$, $\sigma = 3$, $\bar{x} = 68.441$



$0.01 = P(\bar{x} > K_{0.01} \mid \mu = 66.7)$

If $\mu = 66.7$ $\bar{x} \sim N\left(66.7, \frac{3^2}{10}\right)$

$\boxed{K_{0.01} = 68.91}$

So, if $\bar{x} > 68.91$ reject H_0

→ Power of test:-

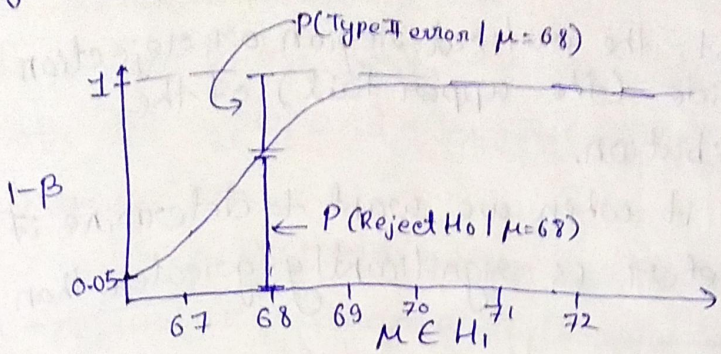
The power of test is a measure of the test's ability to detect a true effect or difference when it exist in the population.

Basically it quantifies the probability of correctly rejecting a false null hypothesis.

Type II error : $P(\text{Do not Reject } H_0 \mid \mu \in H_1) = \beta$

Power of test : $P(\text{Reject } H_0 \mid \mu \in H_1) = (1-\beta)$

Eg) Graph:



* Right-Tailed, Left-Tailed & Two-Tailed Tests:-

Note:-

Before starting a hypothesis test, we should ensure the reliability of data.

Reliability
- → Each sample has to be representive of the population.
- → The data needs to be completely randomized.
- → Sample size should be adequat. (Rule of thumb → 30 samples or more)

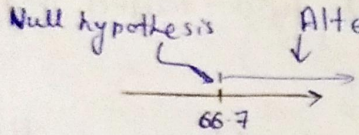
→ Right-tailed, Left-tailed and two-taild tests are the concepts that determins whether a test focuses on one side of distribution, the other side, or both sides.

• Right-Tailed Test:-

→ In a right tailed test, the critical region or rejection is on the right side (the upper tail) of the probability distribution.

We use it when we want to determine if a population parameter is significantly greater than a specified value.

Eg:- $H_0 : \mu = 66.7$ vs $H_1 : \mu > 66.7$

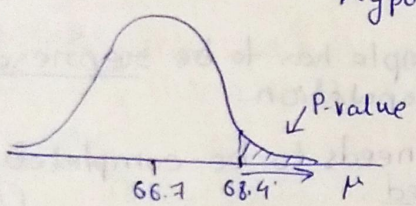


if $\bar{x} >> 66.7 \Rightarrow$ Reject Hypothesis

Type I error:
Determine $\mu > 66.7$, when μ did not change

Type II error:
Do not reject that $\mu = 66.7$, when in true $\mu > 66.7$

Eg:-



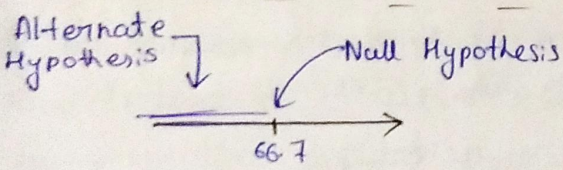
if p-value $< \alpha \rightarrow$ reject H_0

• ~~Left~~ Left-Tailed Test:-

→ In a left-tailed test, the critical region or rejection is on the left side (the lower tail) of the probability distribution.

We use it determine if a population parameter is significantly less than a specified value.

Eg:- $H_0 : \mu = 66.7$ vs ~~$H_1 : \mu > 66.7$~~ $H_1 : \mu < 66.7$

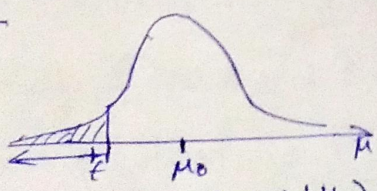


if $\bar{x} << 66.7 \Rightarrow$ Reject H_0

Type I error:
Determine $\mu < 66.7$, when population mean did not change.

Type II error:
Don't reject that $\mu = 66.7$. when in true $\mu < 66.7$

Eg:-



P-value : $P(T(X) < t \mid H_0)$

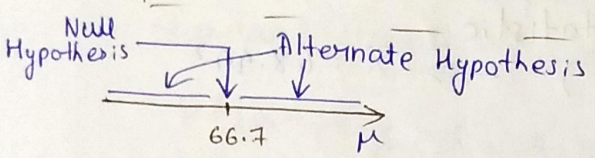
→ Two-Tailed Test :-

- In a Two-Tailed Test, the critical regian is split into both the upper and lower-tails of the probability distribution.

  It is used when we want to determine if a population parameter is ~~split into both t~~ significantly different from a specific value (not just greater or less)

- If the test statistic falls in either tail (greater than the upper critical value or smaller than the lower critical value) then we reject H_0 in favor of H_A.

Eg:- $H_0 : \mu = 66.7$ vs $H_1 : \mu \neq 66.7$



if $\bar{x} \gg 66.7$ or $\bar{x} \ll 66.7$ $\Rightarrow$ Reject H_0

Type I error:
Determine $\mu \neq 66.7$, when population mean did not change.

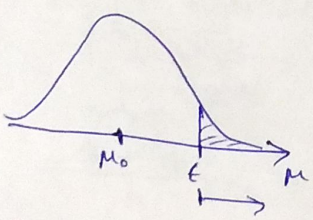
Type II error:
Don't reject that $\mu = 66.7$ when true $\mu \neq 66.7$

Eg:-

$T(X)$: Test Statistics

t : observed statistic

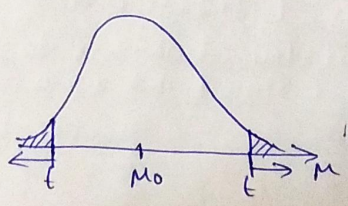
$H_0 : \mu = \mu_0$



Right-Tailed Test

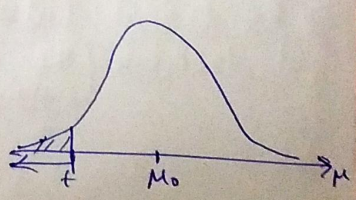
$P(T(X) > t \mid H_0)$

P-value



Two-Tailed Test

$P(|T(X) - \mu_0| > |t - \mu_0| \mid H_0)$



Left-Tailed Test

$P(T(X) < t \mid H_0)$

* Steps for Performing Hypothesis Testing :-

① State the hypotheses.

- Null hypothesis : the baseline $\rightarrow H_0 : \mu = 66.7$
- Alternative hypothesis: the statement we want to prove $\rightarrow H_1 : \mu > 66.7$

② Design the test.

- Decide the test statistic to work with. $\rightarrow \bar{X}$
- Decide the significance level. $\rightarrow \alpha = 0.05$

③ Compute the observed statistic (based on your sample) $\rightarrow \bar{X} = 68.442$

④ Reach a conclusion:

- If the p-value is less than the significance level reject H_0

$\hookrightarrow P(\bar{X} > 68.442 \mid \mu = 66.7) \overset{?}{>} 0.05$

* Chebyshev's Inequality :-

Chebyshev's inequality is a fundamental theorem in probability and statistics that provides a bound on the probability that a random variable takes values within a certain distance from its mean.

It is a useful tool for understanding the spread or variability of data, especially when the probability distribution of the random variable is unknown.

- Chebyshev's inequality states that, for any finite mean (μ) and a finite non-zero standard deviation (σ), the probability that the random variable deviates from its mean by more than K standard deviations is at most $1/K^2$, where $K>1$.

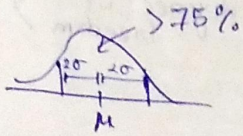
$$P(|X-\mu| \geq K\sigma) \leq \frac{1}{K^2}$$

- Further, It also upholds that $1-\left(\frac{1}{K^2}\right)$ of a distribution's values must be within, but not including, K ~~σ away~~ Standard deviations away from the mean of the distribution

$$P(\mu - K\sigma < x < \mu + K\sigma) > 1 - \frac{1}{K^2}$$

Eg:- $K=2$

$$P(\mu - 2\sigma < x < \mu + 2\sigma) > \underbrace{1 - \frac{1}{2^2}}_{\frac{3}{4} = 75\%}$$



* Z-test & t-test :-

→ Z-test:-

Z-test is used when you have a Known population standard deviation (σ).

It is generally used for relatively large sample size (typically, $n > 30$). However, for smaller sample sizes, it can still be used if the data is normally distributed.

Formula: $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

→ z-test follows a standard normal distribution. ($\mu = 0$, $\sigma = 1$)

→ T-test:-

The t-test is used when we don't know the population standard deviation (σ), and we estimate using the sample standard deviation (s).

It can be used for smaller sizes making it more versatile than the z-test.

Formula: $$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

→ The t-test follows t-distribution with $n-1$ degrees of freedom. (The t-distribution has heavier tails than the normal distribution & varies based on degrees of freedom)

Eg:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{10}} \sim t_g$$

$n = 10$

$\nu = 10 - 1$

Degrees of freedom

⇒ Variations in T-Tests :-

① **One-Sample T-Test :-**

Used to compare the mean of a single sample to a known population mean.

② **Independent Two-Sample T-Test :-**

Used to compare the means of two independent groups (Eg. treatment vs control) to determine if they are significantly different.

③ **Paired (Dependent) T-Test :-**

Used to compare means of paired or matched data points (Eg. before and after measurements) within the same group.

→ Independent Two-Sample, t-Test: Two Tailed Test:-

An independent two-sample ~~t-test~~ t-test, also known as a two-sample t-test or an unpaired t-test, is a statistical difference b/w the means of two independent groups.

⇒ Assumptions :-

- Each person in both the samples are independent.
- All people in the sample from the two groups are different (Random sampling)
- Populations are approximately normally distributed.
- Populations having equal variance (Homoscedasticity).

Eg:-

$X \sim N(\mu_{US}, \sigma^2_{US})$ $\quad$ $Y \sim N(\mu_{Aug}, \sigma^2_{Aug})$

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i \quad ; \quad \bar{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i$$

Then,

$$\bar{X} - \bar{Y} \sim N\left(\mu_{US} - \mu_{Arg}, \frac{\sigma_{US}^2}{10} + \frac{\sigma_{Arg}^2}{9}\right)$$

$$\Downarrow$$

$$\frac{(\bar{X}-\bar{Y}) - (\mu_{US} - \mu_{Arg})}{\sqrt{\frac{\sigma_{US}^2}{10} + \frac{\sigma_{Arg}^2}{9}}} \sim N(0,1)$$

When, we don't know, σ_{US} & σ_{Arg} we replace them with S_X^2 & S_Y^2

So,

$$T = \frac{(\bar{X}-\bar{Y}) - (\mu_{US} - \mu_{Arg})}{\underbrace{\sqrt{\frac{S_X^2}{10} + \frac{S_Y^2}{9}}}_{SE(\hat{p}_X - \hat{p}_Y)}} \sim t_\nu$$

$$\nu(\text{Degrees of freedom}) = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

→ Pooled standard deviation is used in the case of independent two-sample t-tests when we are comparing the means of two groups. It is used to estimate a common standard deviation for both groups.

$$S.D. = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

For computing common proportion we can estimate it by ~~using~~ pooling the samples.

Eg:-

$$\boxed{\begin{array}{l} \overset{p_1}{0.55} \times \overset{n_1}{1000} = 550 \text{ voters.} \\ \overset{p_2}{0.58} \times \overset{n_2}{1500} = 870 \text{ voters.} \end{array}}$$

Then,

$$\frac{550+870}{1000+1500} = \frac{1420}{2500} = 56.8\%$$

So, we estimate $SE(\hat{p}_2 - \hat{p}_1) \doteq \sqrt{\frac{0.568(1-0.568)}{1000} + \frac{0.568(1-0.568)}{1500}}$

$$= 0.02022$$

→ Paired (Dependent) T-test:-

The paired sample t-test also sometimes called the dependent t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero.

In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations.

It's common application would be when, we try to observe "how well a training program was?" or "how effective was a drug".

→ Assumptions:-

- The dependent variable must be continuous (interval/ordinal)
- The observations are independent of one another.
- The dependent variable should be approximately normally distributed.
- The dependent variable should not contain any outliers.

→ Method:-

x	y
X_1	Y_1
X_2	Y_2
X_3	Y_3
$\vdots$	$\vdots$
X_n	Y_n

x → People ~~after~~ before a drug treatment

y → People after a drug treatment

(X_1, Y_1) are the same people before and after drug treatment.

Then,

$$\frac{(X_1-Y_1)+(X_2-Y_2)+(X_3-Y_3)+\dots+(X_n-Y_n)}{n}$$

$$\Rightarrow \bar{D} = \frac{D_1 + D_2 + D_3 + \dots + D_n}{n}$$

⇒ If X_i, Y_i are ~~gauss~~ gaussian ⇒ $D_i = X_i - Y_i$ is gaussian.

$$D_i \overset{i.i.d.}{\sim} N(\mu_D, \sigma_D^2)$$

Then, $\dfrac{\bar{D} - \mu_D}{\sigma_D/\sqrt{10}} \sim N(0, 1)$ But σ_D is unknown

So, we consider S_D

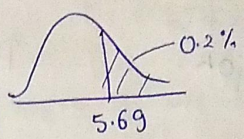
$$\sigma_D \rightarrow S_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}}$$

Then,

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{10}} \sim t_{10-1} = t_g$$

This will consider the p-value

Eg:- $T = 5.69$



---

* Parametric & non-parametric test :-

→ Parametric test is a statistical test which assumes parameters and the distributions about the population are known.

It uses a mean value to measure the central tendency.

→ Non-parametric test does not make any assumptions and measures the central tendency with the median values.

→ Although non-parametric tests requires less assumptions than parametric tests.

Parametric tests are more powerful than non-parametric tests.

Meaning parametric test gives a better guarantee when we reject a null hypothesis than in case for non-parametric test.

Properties	Parametric	Non-Parametric
Assumptions	Yes / many	No / few
Central tendency values	Mean value	Median value
Correlation	Pearson	Spearman
Probability distribution	Normal	Arbitrary.
Population Knowledge	Requires	Does not require
Used for	Interval Data	Nominal Data
Applicability	Variables	Attributes & Variables
Eg:-	z-test, t-test	Kruskal-Wallis, Man-Whitney

	Parametric tests	Non-Parametric Tests
One sample →	sampt t-test	Wilcoxon test for one sample
Two dependent samples →	Paired sample t-test	Wilcoxon test
Two ~~ide~~ independent samples →	Unpaired sample t-test	Mann-Whitney U Test
More than two independent samples →	One factorial ANOVA	Krusskal-Wallis Test
More than two dependent samples →	Repeated Measures ANOVA	Friedman Test
Correlation between two samples. →	Pearson-Korrelation	Spearman-Korrelation

* Sign Test :-

→ The sign test is a non-parametric test is used to test whether or not groups are equally sized.

The sign test is used when dependent samples are ordered in pairs, where the bivariate random variables are mutually independent.

It is based on the direction of the plus and minus sign of the observation, and not on their numerical magnitude. It is a binomial sign test, with $p = 0.5$...

→ The sign test is considered a weaker test, because it tests the pair value below or above the median and it does not measure the pair difference.

→ Assumptions :-

- Data distributions :- The sign test is a non-parametric test, (distribution free)

  so we do not assume that the data is normally distributed.

- Two sample :-

  Data should be from two samples. The population may differ for the ~~following~~ two samples.

- Dependent sample :-

  Dependent samples should be a paired sample or matched. Also known as 'before-after' sample.

Suppose we want to test the hypothesis that median (η) of a population has a specified value, say η_0, i.e.,

$$H_0 : \eta = \eta_0$$

Vs $H_1 : \eta \neq \eta_0$ (Two-tailed) or $H_1 : \eta > \eta_0$ (Right-tailed)

or $H_1 : \eta < \eta_0$ (Left-tailed)

→ Procedure:-

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size n from the given population with median $\eta = \eta_0$ (under H_0).

→ Subtract, η_0 from each X_i's and write or Replace each observation with,

- exceeding η_0 → with plus (+) sign.
- less than η_0 → with minus (+) sign,
- equal to η_0 with zero (0).

→ By the definition of Median, we have

$$P(X > \text{Median}) = P(X < \text{Median}) = \frac{1}{2}$$

Thus, under H_0 ($\eta = \eta_0$):, we have

$$P(X > \eta_0) = P(X < \eta_0) = \frac{1}{2}$$

Hence, if H_0 is true, then the number of + signs should be approximately equals to the − signs.

If the difference in the number of plus (+) and minus (−) signs is due to chance variations (or fluctuations of sampling), then we fails to reject the H_0.

(Accept H_1)

Notations:- (After discarding zeros)

- T^+ = Number of positive sign
- T^- = Number of negative sign
- $T = \min(T^+, T^-)$
- $n = T^+ + T^-$

Main Procedure:-

The following steps are summarized:-

→ **Step 1:** Set the Hypothesis:-

Null Hypothesis:-	Alternative Hypothesis:-
$H_0 : \eta = \eta_0$	$H_1 : \eta < \eta_0$ or $H_1 : \eta > \eta_0$ or $H_1 : \eta \neq \eta_0$

→ **Step 2:** Compute T^+, T^-

Subtract η_0 from each observations
Discard zeros and hence
compute T^+, T^- values.

→ **Step 3:** Test statistics:

$T = \min(T^+, T^-)$

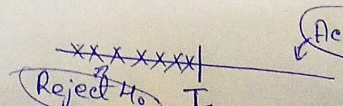
[for $n < 25$]

→ **Step 4:** Critical Region:

Define the critical region as $T \leq T_c$

Where, T_c is the critical value of T at given level of significance for one-tailed or two-tailed

[T_c is found from the table]



→ For Large samples, $n > 25$

Under H_0: $P(T > n_0) = P(T < n_0) = 0.5$

p = Probability of + signs $= 0.5 = \frac{1}{2}$
which is constant for each trial.

Hence, under H_0, the variable T has <u>Binomial distribution</u> with parameter n and $p = 0.5$, i.e., $T \sim Bino(n, 0.5)$

Therefore,

$$\text{Mean} = \underline{E(T) = np} = n \times 0.5 = \underline{\frac{n}{2}}$$

$$\sigma = \sqrt{var(T)} = \sqrt{npq} = \sqrt{\frac{n}{4}} = \underline{\frac{\sqrt{n}}{2}}$$

$[\because var(T) = \underline{npq}]$

[So, the 3rd step of the procedure will change for <u>large samples</u>:-]

→ Computing test statistics:-

If $T^+ < \frac{n}{2}$ $Z = \dfrac{(T^+ + 0.5) - \frac{n}{2}}{\frac{1}{2}\sqrt{n}} \sim N(0,1)$	If $T^+ > \frac{n}{2}$ $Z = \dfrac{(T^+ - 0.5) - \frac{n}{2}}{\frac{1}{2}\sqrt{n}} \sim N(0,1)$

⇒ Paired Sample Sign Test :-

It is a variation of sign test used in case of dependent or paired sample data

Let (X_i, Y_i), $i = 1, 2, \ldots, n$ be a collection of paired observations from two continuous population.

X	Y
⋮	⋮
n_1	n_2

To perform a paired test n_1 and n_2 must be $\boxed{n_1 = n_2}$

and $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$ be the paired differences.

Based on D_i we assign the signs.

Every step after it is identical to normal sign test.

Eg:-

X	Y	D	sign
4	5	-1	-
7	3	+4	+
9	9	0	~~+~~
1	4	-3	-
11	14	-3	-

Then, $T^+ = 1$

$T^- = 3$

$n = T^+ + T^- = 4$

$T = \min(T^+, T^-)$

$= 1$

Here, when $n \geq 10 \rightarrow$ Large sample

when $n < 10 \rightarrow$ small sample.

* Wilcoxon Signed-Rank Test:

→ The wilcoxon Signed-Rank test is used in similar situations as the Mann-Whitney U-test.

The main difference is that the Mann-Whitney U-test tests two independent samples, whereas the Wilcoxon sign test tests two depent samples

→ The Wilcoxon signed rank test pools all differences, ranks them and applies a negative sign to all the ranks where the difference between the two observations is negative. ~~and~~

This is called the signed rank.

→ The Wilcoxon test can be good alternative to the t-test when population means are not of interest;

(Rather interest is the median).

→ The sign test is not very sensitive to the amount of variation of a data value from the conjectured value of the median.

It uses only the information whether the value is above or below the median value.

To overcome it, a new test named as Wilcoxon Signed-Rank test is used, in addition to the signs, also uses the amount of variation of data values from the conjectured values and Hence is considered to be better than the sign test

Here the procedure is identical as for signed test the only changes are:

① Step 2: Computing T^+ & T^-

② Step 3: Test statistics (Only for large samples $n \geq 30$)

1) Step 2: Complete T^+, T^-

Subtract η_0 from each observation Discard zeroes and hence compute T^+, T^- values

Eg:-

Median = 3

X	Y = X - Median	|Y|	Sign	Rank of |Y|
6	3	3	+	8
5	2	2	+	5.5
1	-2	2	-	5.5
2	-1	1	-	2
2	-1	1	-	2
5	2	2	+	5.5
7	4	4	+	10
5	2	2	+	5.5
3	0	0		
7	4	4	+	10
4	1	1	+	2
7	4	4	+	10

When, we have multiple |Y| same value

Like,

→ 1, 1, 1 for 1st, 2nd, 3rd rank, we assign these |Y| the avg ranks, $\frac{1+2+3}{3} = 2$

→ 2, 2, 2, 2 for 4th, 5th, 6th, 7th, $\frac{4+5+6+7}{4} = 5.5$

$T^+ =$ ~~Pos~~ Sum of positive Ranks

$T^- =$ sum of Negetive Ranks

$T^+ = 8 + 5.5 + 5.5 + 10 + 5.5 + 10 + 2 + 10$

$= 56.5$

$T^- = 5.5 + 2 + 2$

$= 9.5$

$T = \min(T^+, T^-) = 9.5$

---

2) Step 3: Test statistics:-

When, $n \geq 30$, we use the <u>normal test</u>

$$Z = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

where $T = \min(T^+, T^-)$

$$\mu_T = \frac{n(n+1)}{4}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Then, check the value of observed z with the critical value of z derived from the question.

(Either given in question or find from z-table)

* Mann-Whitney U test :-

(Also called Mann-Whitney-Wilcoxon (MWU/MWW), Wilcoxon rank-sum test, or Wilcoxon-Mann-Whitney test)

It is a non-parametric test of the null hypothesis that, for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X.

→ **Assumptions:-**

1. All the observations from both groups are independent of each other.
2. The responses are atleast ordinal.
3. Under the H_0, ~~both~~ the distribution of both ~~procedures~~ populations are identical.
4. The alternative H_1, the distributions are not identical.

Let $(X_1, X_2, \ldots X_{n_1})$ and $(Y_1, Y_2, \ldots, Y_{n_2})$ be independent random samples of size n_1 & n_2 from two populations with ~~p.d.f's~~ p.d.f.'s $f_1(x)$ & $f_2(y)$ respectively.

(We want to test if the samples have been drawn from the same population, that is $f_1(x) = f_2(x)$ for all x)

The Mann-Whitney test uses the sum of ranks (or rank-sum) for each sample.

→ Procedure :-

① Step 1 : (Similar to signed test)
Set the Hypothesis

② Step 2 :
Compute U_x, U_y

③ Step 3 :
Test statistics - U

④ Step 4 : (Find from Mann-Whitney U table)
Critical Region :

Changes :-

② Step 2 : Compute U_x, U_y

→ T_x : Sum of the ranks of the observation in X sample
→ T_y : " " " " " " " Y sample

→ $$\boxed{U_x = T_x - \frac{n_1(n_1+1)}{2}}$$

& → $$\boxed{U_y = T_y - \frac{n_2(n_2+1)}{2}}$$

→ n_1 = sample size of X
→ n_2 = " " " Y

→ $U = \min(U_x, U_y)$

The following equalities always satisfy :-

→ $U_x + U_y = n_1 n_2$

→ $T_x + T_y = \frac{(n_1+n_2)(n_1+n_2-1)}{2}$

Eg:-

X : 60 45 23 32

Y: 10 25 20 54 32 65 8

$\Downarrow$

Sample : Y Y Y X Y X Y X Y X Y

Data : 8 10 20 23 25 32 32 45 54 60 65

Rank : 1 2 3 4 5 6.5 6.5 8 9 10 11

$T_x = 4 + 6.5 + 8 + 10$
$= 28.5$

$T_y = 1 + 2 + 3 + 5 + 6.5 + 9 + 11$
$= 37.5$

Then,

$U_x = T_x - \frac{n_1(n_1+1)}{2}$
$= 28.5 - \frac{4(4+1)}{2}$
$= 18.5$

$U_y = T_y - \frac{n_2(n_2+1)}{2}$
$= 37.5 - \frac{7(7+1)}{2}$
$= 9.5$

Therefore,

$U = \min(U_x, U_y)$
$= 9.5$

Now, compare it with critical value of U to see if we may accept H_0 or reject it.

When, $n_1 \leq 10$ & $n_2 \leq 10$, it is small sample.
otherwise large sample

③ Step 3: Compute the test statistics:-

$$z = \frac{U - \mu_U}{\sigma_U} \sim N(0,1)$$

where, $U = \min(U_x, U_y)$;

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Then, we use this z value to compare with critical value of z derived from the question.

* Bootstrap Hypothesis Test:-

It implements the concept of bootstraping:-

→ Elements of Hypothesis Test:-

① Specify H_0, H_A

② Choose Test-stat → $\text{Test-stat}_1 = |\bar{Y}_c - \bar{Y}_M|$ (mean)
→ $\text{Test-stat}_2 = |\hat{Med}_c - \hat{Med}_M|$ (median)

③ Determine distribution of test stat

④ Convert Test-stat → P-value.

$$\text{P-value} = \frac{\text{No. of Bootstrap-Test-stats} \geq \text{Observed Test-stat}}{\text{No. of Bootstraps}}$$

Reasons for using Bootstrap:-

→ Smaller sample size

→ Standard Error and /or sampling Distribution for test ~~stast:~~ statistic is difficult to work out

* Permutation Tests:-

A permutation test (also called re-randomized test or shuffle test) is an exact statistical hypothesis test making use of the proof by contradiction.

The method by which treatments are allocated to subjects in an experimental design is mirrored in the analysis of that design.

They are based on the idea of repeatedly shuffling or permuting the data to generate a null distribution under the assumption of no effect.

⇒ These tests are particularly useful when the assumptions of traditional parametric tests cannot be met or when the sample size is small.

⇒ Procedure :-

① Formulate the hypothesis:-

② Choose a test statistic :- (Mean diff, Median diff, etc)

④ Calculate the test statistic for each permutation.

⑤ Compare the observed Test statistic to the Permutation Distribution.

⑥ Calculate p-value

⇒ Permutation tests may be computationally intensive, especially with large datasets, as all possible permutations need to be considered.

---

* ANOVA :-

→ It stands for "Analysis of Variance" is an extremely important tool for analysis of data (both One way and Two way ANOVA as used).

It is ~~the~~ a statistical method to compare the population means of two or more groups by analyzing variance.

The variance would differ only when the means are significantly different.

→ ANOVA test is the way to find out if survey or experiments are significant.

We are testing groups to see if there's a difference between them.

→ One-way ANOVA:- (CRD → Completely Randomized Design)

⇒ **Purpose:**

Used when comparing means across more than two groups (treatments or levels) for a single independent variable.

Eg:- Comparing the mean scores of three different teaching methods.

→ It deals with one independent variable.

[Number ~~of~~ of Independent Variables]

→ ~~It has a one-dimensional design matrix.~~ [~~Design Matrix~~]

→ It provides information about differences in means across different levels of single factor.

[Interpretation]

→ It compares three or more levels of one factor.

→ Basic Terms in ANOVA

① Experimental unit – The object on which a measurement (or measurements) is taken.

② Factor – It is an independent variable whose values are controlled and varied by the experimenter.

③ Level – It is the intensity setting of a factor

④ Treatment – It is a specific combination of factor levels.

→ Assumption for ANOVA :-

① Samples follow normal distribution.

② Samples have been selected randomly and independently.

③ Each group should have common variance

④ Data are independent.

⑤ Additivity of variance : Total variance = Between variance + Within Variance [One-way]

Note:-

→ $ANOVA = \frac{\text{Variance Between}}{\text{Variance Within}}$

→ Total variance = Variance Between + Variance Within

ANOVA > 1 [Reject H_0]

ANOVA ≤ 1 [Fail to Reject H_0]

→ Three Kinds of variations in ANOVA:-

① Between Groups: Variation from one group to another

$$\sum n_i (\bar{x}_i - \bar{x})^2$$

② Within Groups: Variation among the observations of each specific group.

$$\sum\sum (x_{ij} - \bar{x}_i)^2$$

③ Total: Variations among all the observations (which is nothing but the sum of Between Groups & within Groups)

$$\sum\sum (x_{ij} - \bar{x})^2$$

Eg:-

Metods:	A	B	C
	10	8	9
	9	9	8
	8	10	7
	7.5	8	10
	8.5	8.5	9
	9	7	8
	10	9.5	7
	8	9	10
	8	7	9
	9	10	8
Mean:	8.7	8.6	8.5

Between Group Variation

$$= 10(8.7-8.6)^2 + 10(8.6-8.6)^2 + 10(8.5-8.6)^2$$

$$= \underline{\underline{0.2}}$$

Within group variation:-

A: $(10-8.7)^2 + (9-8.7)^2 + \dots (9-8.7)^2 = 6.6$ ($\bar{y}_A$)

B: $(8-8.6)^2 + (9-8.6)^2 + \dots + (10-8.6)^2 = 10.9$ ($\bar{y}_B$)

C: $(9-8.5)^2 + (8-8.5)^2 + \dots + (8-8.5)^2 = 10.5$ ($\bar{y}_C$)

$$6.6 + 10.9 + 10.5 = \underline{\underline{28}}$$

Mean $\equiv$ 8.6 ($\bar{\bar{y}}$)

Method 1

$$\frac{\text{Variance Between}}{\text{Variance Within}} = \frac{0.2}{28} = \underline{0.0071} < 1$$

[Fail to reject H_0]

Method 2

$F_{stat} = \underline{0.0071}$

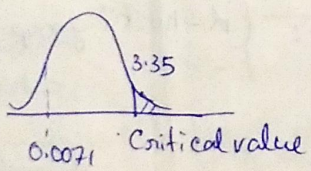
$$F_{critical} = \frac{\text{Numerator Degree of Freedom}}{\text{Denominator Degree of Freedom}}$$

→ Numerator Degree of Freedom = No. of samples − 1
$= 3 - 1 = \underline{2}$

→ Denominator Degree of Freedom $= \Sigma(n_j - 1) = n_T - K = 30 - 3$
$= \underline{27}$

$F_{critical} = F_{(2,27)} = 3.35$

∴ $F_{critical} > F_{stat}$; we fail to reject H_0.



Another example

We have K groups and the jth group has n_j observations

group 1	group 2	. . .	group K
y_{11}	y_{12}	. . .	y_{1K}
$\vdots$	$\vdots$		
$y_{n_1 1}$	$y_{n_2 2}$	. . .	$y_{n_K K}$

In total there are $N = n_1 + \dots + n_K$ observations

The sample mean of the jth group is $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$.

The overall sample mean is $\bar{\bar{y}} = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{n_j} y_{ij}$

or

$$\bar{\bar{y}} = \frac{1}{K} \sum_{j=1}^{K} \bar{y}_j$$

or n_K

The treatment sum of squares: (~~Total~~ Between variance)

$$\underset{(SS_B)}{SST} = \sum_j \sum_i (\bar{y}_j - \bar{\bar{y}})^2$$

→ It has $K-1$ degrees of freedom: (K = no. of groups)

→ Treatment mean square (MST)

$$\underset{(MS_B)}{MST} = \frac{SST}{K-1}$$

[It measures the variability of the treatment means $\bar{y}_j$ [variance b/w groups]]

→ The error sum of squares: (SSE) [Within group Variation]

$$\underset{(SS_E)}{SSE} = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$$

→ SSE has N-K degrees of freedom.

The error mean square, $MSE = \frac{SSE}{N-K} = MS_E$

It measures the variability within the groups.

Here, $N = K \times n$

K → no. of groups

n → No. of elements or rows in data

Then, $N-K = K \times n - K = K(n-1)$

→ F ratio : $F = \boxed{\frac{MS_B}{MS_E}} = \frac{MST}{MSE}$

→ p-value : $P(F > F_0)$

→ Total sum square:

$$TSS = SST + SSE$$

One way ANOVA table

Source	df Degree of Freedom	SS (Sum of squares) [Variation]	MS (Mean squares) [Variance]	F (F ratio)	p-value
Factor (Between)	$K-1$	SS_B	$MS_B = \frac{SS_B}{K-1}$	$F = \frac{MS_B}{MS_E}$	$P(F > F_0)$
Error (Within)	$K(n-1)$	SS_E	$MS_E = \frac{SS_E}{K(n-1)}$		
Total	$K-1+K(n-1)$ [$Kn-1$]	$SS_T = SS_B + SS_E$			

F_0 is F_e

$= \frac{K-1}{K(n-1)}$

Another way:-

~~Corre.~~

	A	B	C
	x_1	y_1	z_1
	⋮	⋮	⋮
	x_n	y_n	z_n
Total	X	Y	Z

$X + Y + Z = \underline{\underline{G}}$

→ Correction Factor: $C = \frac{G^2}{N}$

$N = n \times K$

n → # rows, K → # group

→ $SST = \sum\sum x_{ij}^2 - C$

→ $SSB = \sum_{i=1}^{K} \frac{T_i^2}{n_i} - C$ [T → Total of a group]

→ $SSE = SST - SSB$

→ Two-way ANOVA:- [RBD → Randomised Block Design]

⇒ Purpose:-

Used when there are two independent variables (factors) and you want to study their main effects and interactions on a dependent variable.

Eg:- Investigating the effects of two factors, such as drug dosage and gender on blood pressure.

→ Two way analysis ~~have~~ deals with two independent variables.

→ Two way analysis provides information about the main effects of two factors and their interaction.

→ Extra assumption in a two-way ANOVA :

The effects of one factor is the same at all levels of the other factor.

Meaning that there is no interaction between the two factors.

⇒ Interaction exists when the effect of one factor depends upon the level of the other factor.

[Interaction can only be assessed if more than one measurement is taken at each combination of the factor levels]

Here we make 4 hypothesis; 2 null hypothesis and 2 H_A for respective factors:

Factor 1	Factor 2
H_0	H_0
H_A	H_A

Source of Variation	df	SS	MSS = SS/df	F-Ratio
Between Row ①	$r-1$	SSR	MSSR	$\frac{MSSR}{MSSE}$
Between Column ②	$c-1$	SSC	MSSC	$\frac{MSSC}{MSSE}$
Within Error	$(r-1)(c-1)$	SSE	MSSE	
Total ③	$rc-1$	SST		

Within Error = ③ − ① − ②

r = # rows
c = # columns

Here,

	P	Q	R	S	Total
A	1	2	⋮	a	T_1
B	2 ⋮				T_2
C	n				T_3
					G

Correction Term)

$$C = \frac{G^2}{N}$$

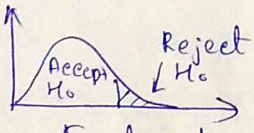
① ~~SSB~~ SSR (Between Row) $= \sum \frac{T_i^2}{n_i} - C$

② SSC (Between columns) $= \sum \frac{T_j^2}{n_j} - C$

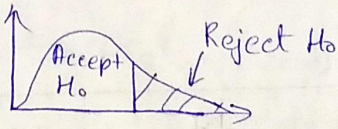
③ SST (Total) $= \sum\sum x_{ij}^2 - C$

Within Error = (3) - (1) - (2)

Then, we would finally compute two F-ratios for the two factors



Factor 1
(Row Factor)



Factor 2
(Column Factor)

* ANCOVA

(~~An~~ Analysis of Covariance)

⇒ Purpose:

~~ANCova~~ → ANCOVA is used when we have a categorical independent variable (factor) and one or more continuous covariates.

→ It assess whether population means of a dependent variable (DV) are equal across levels of a categorical independent variable, while statistically controlling for the effects of other continuous variables

→ It increases the sensitivity of the test for main effects and interactions by reducing the error term; the error term is adjusted for, and hopefully reduced by, the relationship between the DV and the CV(s).

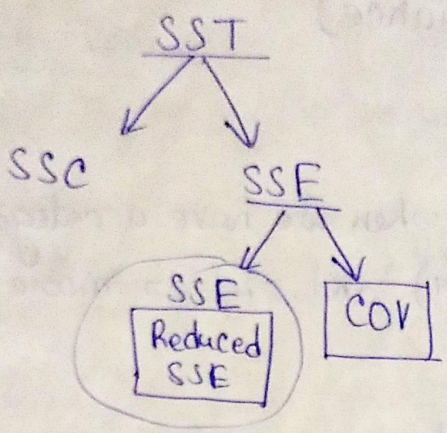
CVs are used to assess the "noise" where noise is the undesirable variance in the DV that is estimated by scores on the CV.

→ To adjust the means on the levels of DV itself to what they would be if all subjects scored equally on the CV(s). Differences b/w subjects on CV(s) are removed so the presumably, the only differences that remain are related to the effects.

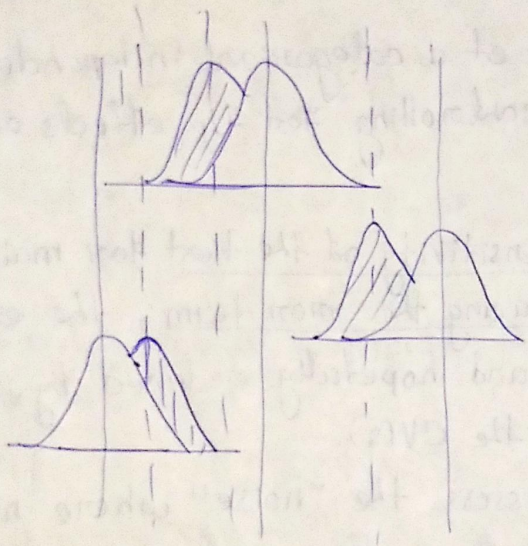
The CV(s) enhance prediction of the DV, but there is no implication of causality.

Eg:- Examining the effect of a teaching method (categorical variable) on student performance while controlling for pre-test scores (covariate)

→ ANCOVA increases power of F.



→



→ Assumptions :-

Similar to ANOVA, including homogeneity of variance and normality of residuals.

* MANOVA (Multivariate Analysis of Variance)

⇒ Purpose:

→ MANOVA is an extension of ANOVA when there are two or more DVs that are related or conceptually similar.

Eg:- Investigating the effect of a treatment on multiple outcome measures simultaneously, such as blood pressure, cholesterol levels, and heart rate.

⇒ Procedure:-

→ Use a statistical software package to conduct MANOVA, specifying the independent and dependent variables.

→ Choose a test statistic (Eg:- Wilk's Lambda, Pillai's Trace) and examine sig significance level.

⇒ Interpretation:-

If the MANOVA test is statistically significant, then we reject H_0; otherwise may accept H_0

* MANCOVA (Multivariate Analysis of Covariance):

Purpose :-

→ MANCOVA is an extension of both ANCOVA and MANOVA.

→ It combines the features of ANCOVA and MANOVA, allowing for the analysis of group differences on multiple dependent variables while controlling for the effects of covariates.

⇒ Test statistics:-

Similar to MANOVA

* ~~Post~~ Multiple comparisons:-

→ Multiple comparisons refer to the problem of conducting several pairwise comparisons b/w group means after detecting a significant difference among the groups.

→ It helps to identify groups that differ from other groups significantly.

* Post-hoc test:-

Post hoc test → "Post hoc analysis"
→ "Post hoc comparison"

→ It is conducted after a main analysis (like ANOVA or regression) to ~~sure~~ make specific pairwise comparisons between groups.

→ These tests are used when the initial analysis indicates a significant difference but does not provide information about which specific groups are different from each other.

⇒ Various approaches to conduct post hoc analysis and multiple comparisons:-

1. Tukey's Honestly Significant Difference (HSD)
2. Fisher's Least Significant Difference (LSD)
3. Holm's Method
4. Dunnett's Test
5. Schettie's Test
6. Bonferroni Correction

→ It ~~ab~~ adjusts the significance level for each comparision to maintain an overall level of Type I error

→ Simple but can be conservative, potentially leading to increased Type II errors.

Familywise Error Rate (FWER):-

→ The FWER is the probability of making one or more Type I errors in a set of comparisons or tests.

→ In the context of multiple comparisons or tests, conducting several tests simultaneously increases the likelihood of at least one Type I error occuring by chance.

$$FWER = 1-(1-\alpha)^K$$

Let $\alpha = 0.05$, $K = 1$

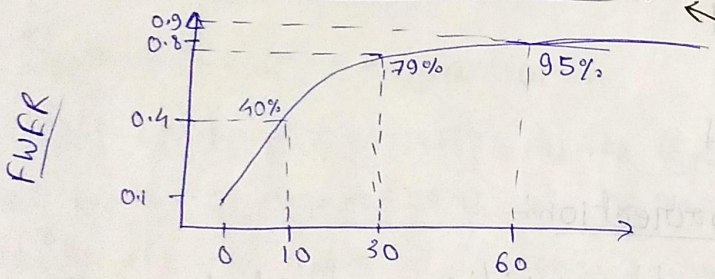
$FWER = 1-(1-0.05)^1 = 0.05$ (Not much change)

Let $K = 3$,

$FWER = 1-(1-0.05)^3 = 0.143$ ← meaning there is 14% chance of having Type I error

Let $K = 6$,

$FWER = 1-(1-0.05)^6 = 0.265 = 26.5\%$ ← Multiple test Fallacy



→ To account for this error, we can use a multiple comparision correction method, (Eg:- Bonferroni)

$$\text{Bonferroni-corrected } \alpha = \frac{\alpha}{K}$$

α → original α

K → number of tests

Let $\alpha = 0.05$ & $K = 6$

Then,

$$B-C\ \alpha = \frac{0.05}{6} = 0.008$$

Now, $FWER = 1-(1-\alpha)^6 = 1-0.992^6$

$= 0.047$

$= 4.7\%$

Using Bonferroni correction the FWER went down from 26.5% to 4.7%

* Kruskal-Wallis Test or H-Test :-

It is the non-parametric and valuable alternative to one-way ANOVA test (F test) when the normality and equality of variance assumptions are violated.

⇒ Assumptions:-

→ There are at least three independently drawn random samples.

→ Each sample has at least 5 observations, $n_i \geq 5$.

Kruskal-Wallis Test is used to test the Null Hypothesis if K independent samples, $K \geq 3$, have been drawn from the population which have identical distributions, and does not require the conditional of normality of the population

⇒ Stating Hypothesis:-

H_0:

$\mu_1 = \mu_2 = \ldots = \mu_K$

H_A:

At least one of the μ_i's is different from others.

Kruskal-Wallis Test is an extension of the Mann-Whitney U-test to the situation where more than two populations are considered and is based on the ranks of the sample observations.

⇒ Notations:-

→ K : number of samples or groups

→ n_i : Size of the ith sample ; $i = 1, 2, \ldots, K$

→ $n = n_1 + n_2 + \ldots + n_K$

(Total number of observation in K samples)

→ T_i : Sum of the ranks of the ith sample.

Rank from smallest (1) to the largest (n).
If there is a tie, then we take average.

⇒ Procedures:-

→ Step 1 :
Define H_0 & H_A

→ Step 2 :
Rank the sample observations in the combined series and.
Compute T_i : sum of the ranks of the ith sample.

→ Step 3 :

Compute Test statistics, H-value :

Under H_0,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{T_i^2}{n_i} - 3(n+1)$$

and is approximately distributed as chi-square variate with $(K-1)$ degree of freedom

No. of samples.

Thus, $H \sim \chi^2_{(K-1)}$ at α level of significance

→ Step 4:

Take the tabulated value of $\chi^2_{(K-1)}(\alpha)$.

(and compare it with the computed H from step 3)

Note:-

Kruskal-Wallis Test is a Right-Tailed Test.

* Repeated Measures ANOVA:-

Repeated measures ANOVA is a statistical technique used to analyse the mean differences within a single group that has been measured at multiple time points. or under different conditions.

This design is often called a "within subjects". or "within-groups" design.

It is an extension of the one-way ANOVA, allowing for the analysis of repeated measurements. on the same subject.

Eg:- Readings of patient before a therapy and after 2 days and 2 weeks of therapy.

⇒ Assumptions:-

(Mauchly's Test of sphericity)

→ Sphericity:-
The variances of the different differences between all possible pairs of conditions are equal.

→ The observations within each group are independent, and the dependent variable is measured on an interval or ratio scale.

Total Variation

Between group variation

Within group Variation (~~SSE~~) (SSWg)

sum of square within group

Error Variability

Between subjects Variation

$F = \frac{\text{Variance between groups (SSB)}}{\text{Variance in errors}}$

Most of the computations are similar to the one-way ANOVA,

Here, the Variance within groups are further splitted as Variance in errors (SSE) and Between subjects (SSS) Variance.

sum of square Errors

Sum of square of subjects.

$SSE = SSWg - SSS$

→ ~~F~~ =

degree of freedom for Between grp. $(df_B) = K - 1$

degree of freedom for Within grp $(df_W) = N - K$

degree of freedom for subjects $(df_S) = N_{sub} - 1$

degree of freedom for errors $(df_{error}) = df_W - df_S$

So, $MSSB = \frac{SSB}{df_B}$, $MSSE = \frac{SSE}{df_W - df_S}$

→ $\boxed{\text{F-ratio} : \frac{MSSB}{MSSE}}$

* Friedman test:-

The friedman test is a non-parametric statistical test used to detect differences in treatments across multiple related groups.

It is the non-parametric equivalent of the repeated measures ANOVA and is appropriate when the assumptions ~~are~~ of normality and sphericity are not met.

⇒ Assumptions:-

→ The dependent variable is measured on an ordinal or continuous scale.

→ Observations within each group are independent

→ The dependent variable is not assumed to be normally distributed.

⇒ Test statistic:-

The friedman test uses the chi-square (χ^2) distribution.

Here, we rank elements within each Row.

Eg:-

~~34~~	34 → 3	45 → 1	36 → 2
	33 → 2	36 → 1	31 → 1

The rest of the procedure is similar to the Kruskal-Wallis test.

Note:-

During Multiple comparisons and analyzing large amounts of data, it is easy to fall into the trap of multiple test fallacy or look-elsewhere effect. because there are so many potential ~~ste~~ relationships to explore, ~~When analyzing large amounts of data is easy to fall int.~~

which leads to data snooping (= data dredging)

* P-hacking :-

→ P-hacking, short for "probability hacking" or "data dredging", is a questionable research practice, wherein researchers manipulate or exploit statistical analyses, often subconsciously to achieve statistical significance or to find significant results.

→ P-hacking involves selectively analyzing data or cherry-picking results, conducting multiple statistical tests, or adjusting analytical choices until a desired level of statistical significance (usually $p < 0.05$) is obtained.

→ P-hacking can contribute to publication bias, where only statistically significant results are more likely to be published.

This can distort the scientific literature and result to misleading conclusions.

[We can use correction methods like Bonferroni correction, False Discovery Rate control for multiple testing]

→ 'Data dredging' or 'p-hacking' and other problems have lead to crisis with regard to replicability (getting similar conclusions with different samples, procedure and data analysis methods) and reproducibility (getting the same results when using the same data and methods of analysis)

* False Discovery Proportion (FDP):

Accounting for multiple testing, we can also try to control the False Discovery Proportion (FDP)

$$FDP = \frac{\text{number of false discoveries}}{\text{Total number of discoveries}}$$

where, 'discovery' occurs when a test rejects the null hypothesis.

* FDR:

False Discovery Rate (FDR): It controls the expected proportion of discoveries that are false.

For a multiple testing threshold T, (significance)

$$FDR = E(FDP(T))$$

$$= E\left(\frac{FP}{TP+FP}\right)$$

[Usually, we ~~tike~~ want to control, FDR = 5%]

⇒ Benjamini-Hochberg procedure to control the FDR at level $\alpha = 5\%$ (say)

① Sort the p-values : $P_{(1)} \leq \dots \leq P_{(m)}$

② Find the largest K such that $P_{(K)} \leq \frac{K}{m}\alpha$

③ Declare discoveries for all tests i from 1 to K.

---

* Chi-Square Test :-

The chi-square test is a statistical test that is used to determine if there is a significant association between two categorical variables.

It is a non-parametric test, meaning it makes no assumptions about the distribution of the data.

The chi-square test is widely used in various fields, including statistics, biology, social sciences.

⇒ In enables us to find the deviation of the experiment from theory is

- just by chance or
- It really due to the inadequacy of the theory to fit the observed data.

→ Computation:

If $O_i (i = 1, 2, \ldots n)$ be the set of observed (experimental) frequencies and E_i is the corresponding set of expected ~~set of~~ (theoretical or Hypothetical) frequencies, i.e.,

Events	1	2	...	n
Observed freq.	O_1	O_2	...	O_n
Expected freq.	E_1	E_2	...	E_n

then Karl Pearson's chi-square, given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where $\Sigma O = \Sigma E$ follows chi-square ~~dis~~ distribution with $(n-1)$ degree of freedom (d.f.)

→ Conditions for the Validity of χ^2-test

① The sample observation should be independent

② Constraints on the cell frequencies, if any, should be linear, eg. $\Sigma O = \Sigma E$

③ The total frequency should be greater than 50.

④ No theoretical cell frequency should be less than 5

[In this case we pool the cell with preceding or succeeding frequency]

→ Procedure :-

① Step-1 : Define the Null Hypothesis

② Step-2 : Calculate E and χ^2

③ Step-3 : Conclude the result.

→ Choices of tests in χ^2- test :-

① Test ~~for~~ ~~Go~~ for Goodness of Fit :-

→ Purpose :-

It is used when we want to assess whether the distribution of observed categorical data fits a specific expected distribution or model.

Eg :- Checking if the observed distribution of colors of candies in a bag matches the expected distribution or model.

→ Procedure :-

① Formulate hypotheses about the expected distribution.

② Collect obseved frequencies for each category.

③ Calculate expected frequencies based on the hypothesized distribution.

④ Use the chi-square test to compare observed and expected frequencies.

eg:- M&Ms company claim the color distribution of M&Ms are as follows.

Blue	Orange	Green	Yellow	Red	Brown
~~240~~ 24%	20%	16%	14%	13%	13%

We calculated the colors of M&Ms in one bag of milk chocolate M&Ms. We get-

	Blue	Orange	Green	Yellow	Red	Brown	Total
Observed (O)	85	79	56	64	58	68	410
Expected (E)	98.4	82	65.6	27.4	53.3	53.3	410

Based on the given proportions of color distribution by M&M company, we find the expected values for color distribution count for these 410 M&Ms ($\Sigma O = \Sigma E$)

Now, $$\chi^2 = \sum \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$= \frac{(85-98.4)^2}{98.4} + \frac{(79-82)^2}{82} + \dots + \frac{(68-53.3)^2}{53.3} = 8.57$$

Here, (d.f.) degree of freedom $= (6-1) = 5$

② Test for Homogenity:-

→ Purpose:-

Used when you want to compare the distributions of categorical variables across two or more independent groups.

Eg:- Comparing the distribution of categorical ~~variables across two or more~~ independent groups. Comparing the distribution of voting preferences among different age groups in two cities.

→ Procedure:-

① Formulate hypotheses about the homogeneity of distributions across groups.

② Create a contingency table with observed frequencies for each combination of variables and groups.

③ Use the chi-square test to assess whether the distributions are homogeneous across groups.

[Observed Cross Tabulations]

Eg:-

	First	Second	Third	Crew	Total
Survived	202	118	178	215	713
Died	123	167	528	698	1,516
					2,229

Then,

~~Expected~~ Estimated probability as $\frac{713}{2229} = 32\%$

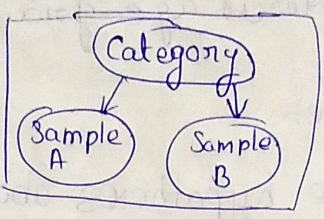
So, the expected number of surviving first class passengers is $32\% \times 325 = 104$

	First	Second	Third	Crew	Total
Survived	104	91.2	225.8	292.1	713
Died	221	193.8	480.1	620.8	1516
					2,229

Now,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 192.2$$

$$d.f. = (4-1) \times (2-1) = 3$$



③ Test for independence :-

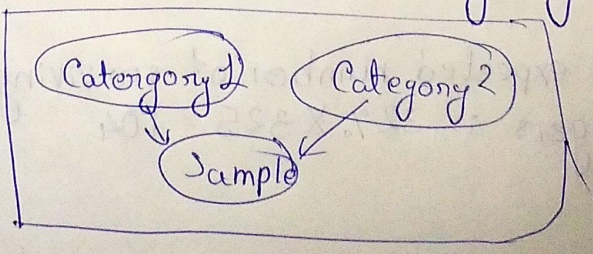
→ Purpose :-

It is used when we want to examine whether there is a significant association between two categorical variables.

Eg :- Investigating whether there is an association between gender and voting preference in a political survey.

→ Procedure :-

The procedure for χ^2 test for independence is same as for χ^2 test for homogeneity.



Summary:-

⇒ χ^2 test for Goodness of Fit:-

→ When we have one categorical variable.

→ Want to compare its distribution to an expected distribution.

⇒ χ^2 test for Homogenity:-

→ When we have two or more independent groups.

→ Want to compare the distributions of a categorical variable across these groups.

⇒ χ^2 test for Independence:-

→ When we ~~want~~ have two categorical variables

→ Want to investigate whether there is a significant association or relationship between them.

	Sample	Research Question
χ^2-test of Homogeneity	Single catergorical variable measured on several samples	Are the groups homogeneous? (have the same distribution of the categorical variable)
χ^2-test of Independence	Two categorical variable meased on one sample	Are the two categorical variables independent?

* F-test :-

→ The F-test is a versatile statistical tool used in various contexts to assess differences in variability or means among multiple groups. It is also an essential component of ANOVA and regression analysis.

→ It is use to check whether the variances of two or more groups are equal.

→ It follows F-Distribution.

Formula:

$$F = \frac{\text{Between-group Variance}}{\text{Within-group Variance}}$$

* Fisher's Exact Test :-

→ Purpose :-

→ It is a statistical test used to determine if there are nonrandom associations between two categorical variables.

→ It is partially useful when dealing with small sample size and when the conditions for using χ^2-test ~~the conditions for using the conditions for using~~ are not met, especially when expected cell frequencies are low.

→ Assumptions:- (In addition to assumptions of χ^2 test)

⇒ Total number of patients (N) < 20 with a 2X2 matrix

OR

→ N > 20, but expected cell count is 5 or greater is less than 80% of cells.

⇒ In this test we don't need to find the degrees of freedom or the critical χ^2 value.

⇒ Here we directly find the p-value.

→ Calulations:-

Response to vaccine

	Susceptible	Resistant	Total
Male	a	b	a+b
Female	c	d	c+d
	a+c	b+d	N = a+b+c+d

Then,

$$p = \frac{{}^{a+b}C_{a} \times {}^{c+d}C_{c}}{{}^{N}C_{a+c}}$$

That's how we get a p-value in this method.

* A-B (A/B) Testing :-

→ A-B Testing, also known as split testing, is a method used in marketing, product development, and other fields to compare two versions of a variable (A & B) to determine which one performs better.

→ This testing approach involves exposing two groups (from random sampling) to different versions of a product, ~~web~~ webpage, email or other elements, and then analyzing the outcomes to make data-driven decisions.

→ Here we can use χ^2 test.

→ For performance metrics we can use conversion rates (No. of people buying) and No. of clicks.

Eg :-

		Purchase Flag (or Conversion)	
		Yes	No
Suggestion Type	With Cover	130	206
	With Phone	117	253

Recommending screen guard with Phone Cover (A) and with Phone itself (B)

* Risk difference, Relative Risk, Odds Ratio & log of odds Ratio in case control study:-

→ A case-control study is a type of observational study design used in epidemiology to investigate the causes of a particular disease or condition.

→ In case-control study, individuals with specific health ~~co~~ outcome (cases) are compared to those without the ~~hat~~ health outcome (controls) with respect to their exposure to potential risk factors.

~~The~~ g

→ The goal is to identify associations between exposure and the occurence of the disease.

⇒ To understand risk difference, relative risk, odds ratio and log of odds ratio we will use the following data:-

Eg:- X = Exposed (Y/N), Y = Disease (Y/N)

[Cross-Table]

		Disease		
		Y	N	Total
Exposed	Y	30	70	100
	N	20	80	100
	Total	50	150	200

Y → Yes
N → No

$P(D/E) = \frac{30}{100} = 0.3$ ≡ [Exposed = Y, Disease = Y]

$P(D/E^c) = \frac{20}{100} = 0.2$ ≡ [Exposed = N, Disease = Y]

→ Relative Risk :-

Relative risk, also known as risk ratio, is a measure used in epidemiology and statistics to quantify the strength of the association between the occurrence of an event in two groups.

It compares the probability of an event happening in two groups.

• **Formula :-**

Relative Risk (RR) =

$$\frac{\text{Risk in Group A}}{\text{Risk in Group B}}$$

Eg :- $RR = \frac{P(D/E)}{P(D/E^c)} = \frac{0.30}{0.20} = 1.5$

∴ Risk of getting disease ~~has~~ is 1.5 times more likely if we get exposed to it compared to not getting exposed

• **Interpretation :-**

→ RR = 1 : Events occur at the same rate in both groups

→ RR > 1 : The event is more likely to occur in Group A

→ RR < 1 : The event is less likely to occur in Group A

→ Risk Difference :-

→ Risk difference, also known as absolute risk reduction, is measure that quantifies the absolute difference in the risk of an event occuring between two groups.

→ It provides a straight forward understanding of the impact of an intervention or exposure on the occurrence of an event.

Formula :- Risk Difference (RD) = Risk in Group A − Risk in Group B

eg:- Risk Difference (RD) $= P(D/E) - P(D/E^c) = 0.30 - 0.20$

$= 0.10 = 10\%$

⇒ Interpretation:-

→ RD = 0: The event occurs at the same rate in both groups

→ RD > 0: The event occurs more frequently in Group A

→ RD < 0: The event occurs less frequently in Group B

→ Odds Ratio:-

→ The odds ratio is a measure of the association between an exposure and an outcome in epidemiology and statistics.

→ It compares the odds of an event happening in one group to the odds of the same event happening in other group.

• Formula:-

$$\text{Odds Ratio (OR)} = \frac{\text{Odds of Exposure in Group A}}{\text{Odds of Exposure in Group B}}$$

eg:-

		Has Cancer	
		Yes	No
Has Mutated genes	Yes	23	117
	No	6	210

Note:-

Odds are just,

$\frac{\text{the ratio of something happening}}{\text{the ratio of something not happening}}$

→ Odds of having cancer given that the person has mutated genes $= \frac{23}{117} = 0.2$ (A)

→ Odds of having cancer given that the person has no mutated genes $= \frac{6}{216} = 0.03$ (B)

Now,

$$\text{Odds Ratio (OR)} = \frac{A}{B} = \frac{0.2}{0.03} = \underline{6.88}$$

This Odds Ratio tells us that the odds are 6.88 times greater that someone with the mutated gene will also have cancer.

→ Log of Odds Ratio (Log(OR)) :-

- To facilitate statistical analysis and interpretation, the natural logarithm of the odds ratio (logarithm to the base e) is often used.
- The log transformed odds ratio is denoted as Log(OR).

- Formula:-

$$\text{Log (Odds Ratio)} = \log(\text{odds Ratio})$$

Eg:- [Extending on the example of Odds Ratio]

$$\text{Log(odds Ratio)} = \log(6.88)$$
$$= \underline{1.93}$$

Note:-

→ Larger Magnitude in Log(OR) (farther from 0) : Strong Association

→ Smaller Magnitude in Log(OR) (Closer to 0) : Weaker Association.

* Q-Q plot:- ~~and~~

→ A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular distribution.

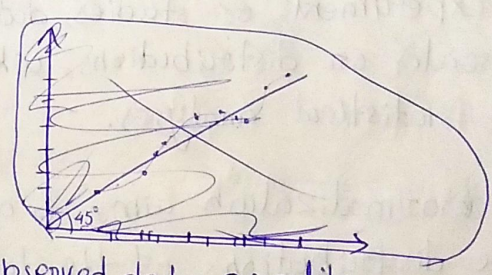
(Mostly we check for normal distribution)

→ It compares the quantiles of the observed data to the quantiles of a specified theoretical distribution, (typically the normal distribution)

→ X-axis represents the quantiles from the theoretical distribution.

→ Y-axis represents the quantiles from the observed dataset.

Eg:-



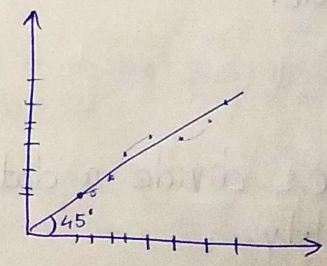
$y: [y_1, y_2, \ldots, y_n]$ → Observed data quantiles

$\underbrace{x: [x_1, x_2, \ldots, x_n]}$

Standard Normal distribution of the observed data

(Normalized quantiles of the observed data)

Observed data Quantiles (y)



Normal distribution Quantiles (x)

⇒ A ~~45°~~ 45° diagonal line is drawn on the plot.

→ If the points on the Q-Q plot closely follow:-

Yes → Then the observed data follows the specified theoretical distribution.

---

* Quantile Normalization:-

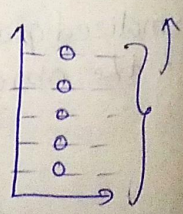
→ Quantile normalization is a statistical technique used to normalize the distribution of two or more datasets by ensuring that their quantiles are the same or very similar.

→ In some experiments or studies, datasets may have different scales or distributions, which can impact the validity of statistical analyses.

→ Quantile normalization aims to address this issue by making the distributions of the datasets more comparable.

Note:-

Quantiles divide a dataset into intervals of equal probability.



⇒ Procedure:-

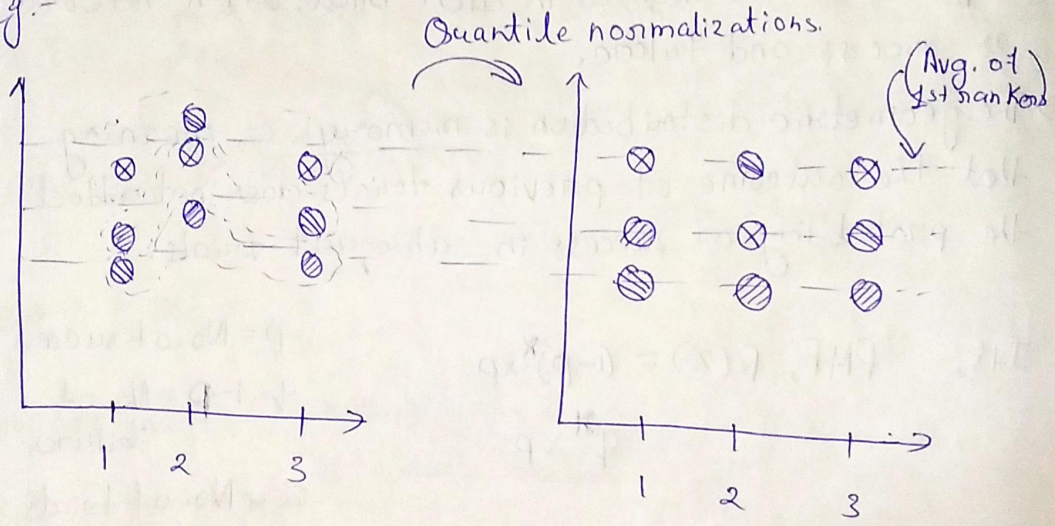
The quantile normalization process involves aligning the quantiles of the datasets. It works as follows:

↳ ① Sort each dataset indepently.

↳ ② Replace the values in each dataset with the corresponding quantiles from a reference distribution (Eg. the avg. of Quantiles)

↳ ③ After quantile normalization, the datasets share similar quantile, making them comparable.

Eg:-



→ It is a robust method that does not assume a specific underlying distribution.

→ After quantile normalization, the values for each sample are the same.

But the order of the sample is ~~preve~~ preserved.