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}
⋮	⋮	...	
$y_{n_1 1}$	$y_{n_2 2}$	...	$y_{n_K K}$

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 $\bar{\bar{y}} = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{n_j} y_{ij}$

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

or n_K

The treatment sum of squares: (Between variance)

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

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

⇒ Treatment mean square (MST)

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

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

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

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

(SS_E)

→ 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 on rows in data

Then, $N-K = Kxn - K = K(n-1)$

→ 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	$MS_E = \frac{SSE}{K(n-1)}$		
Total	K-1+K(n-1) [Kn-1]	$SS_T = SS_B + SS_E$			

F_0 is F_c

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

Another way:-

~~Coore~~.

	A	B	C
	x_1	y_1	z_1
	⋮	⋮	⋮
	x_n	y_n	z_2
Total	X	Y	Z

$X + Y + Z = G_T$

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

$N = n \times K$

rows, # 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:- [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

H_0

H_A

Factor 2

H_0

H_A

Source of Variation	df	SS	MSS = SS/df	F-Ratio
Between ① Row	$r-1$	SSR	MSSR	$\frac{MSSR}{MSSE}$
Between ② Column	$c-1$	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	⋮	c	T_1
B	2	⋮	⋮	⋮	T_2
C	n	⋮	⋮	⋮	T_3
					G

Correction Term)

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

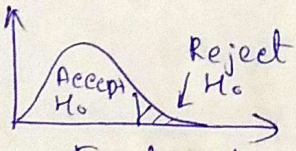
① ~~SSB~~ SSR (Between Row) $= \sum \frac{T_i^2}{n_i} - C$

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

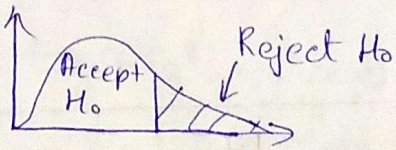
③ SST (Total) $= \sum\sum x_{ij}^2 - C$

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

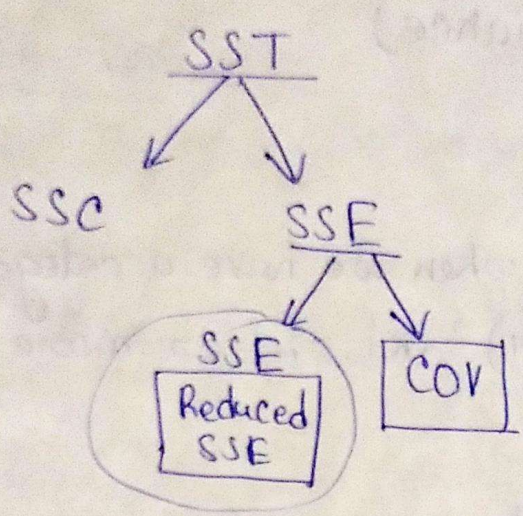
→ It increases the sensitivity of the test for main effects and interactions by reducing the error term; the error term is adjusted for, and hopefully reduced by, the relationship between the DV and the CV(s).
CVs are used to assess the "noise" where noise is the undesirable variance in the DV that is 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.

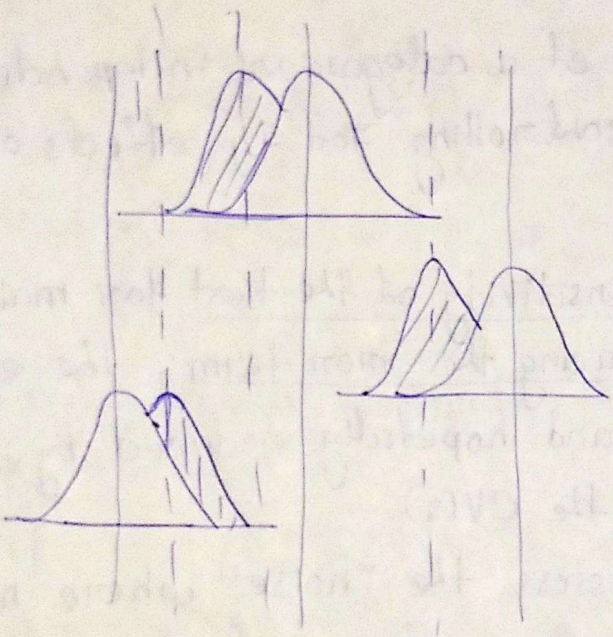
The CV(s) enhance prediction of the DV, but there is no implication of causality.

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 group(s) 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 Scheffe's Test
⑥ Bonferroni Correction

→ It ~~adj~~ 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 $\alpha = 0.05$, $K = 1$

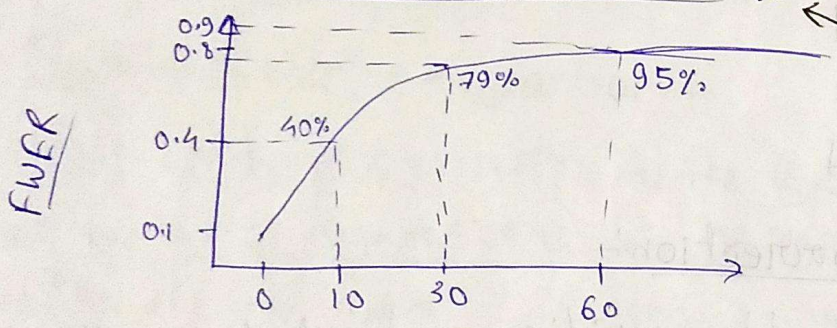
$FWER = 1 - (1-0.05)^1 = 0.05$ (Not much change)

Let $K = 3$,

$FWER = 1 - (1-0.05)^3 = 0.143 \leftarrow$, 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,

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

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 i^{th} 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 i^{th} 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 i^{th} 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

(K → 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.

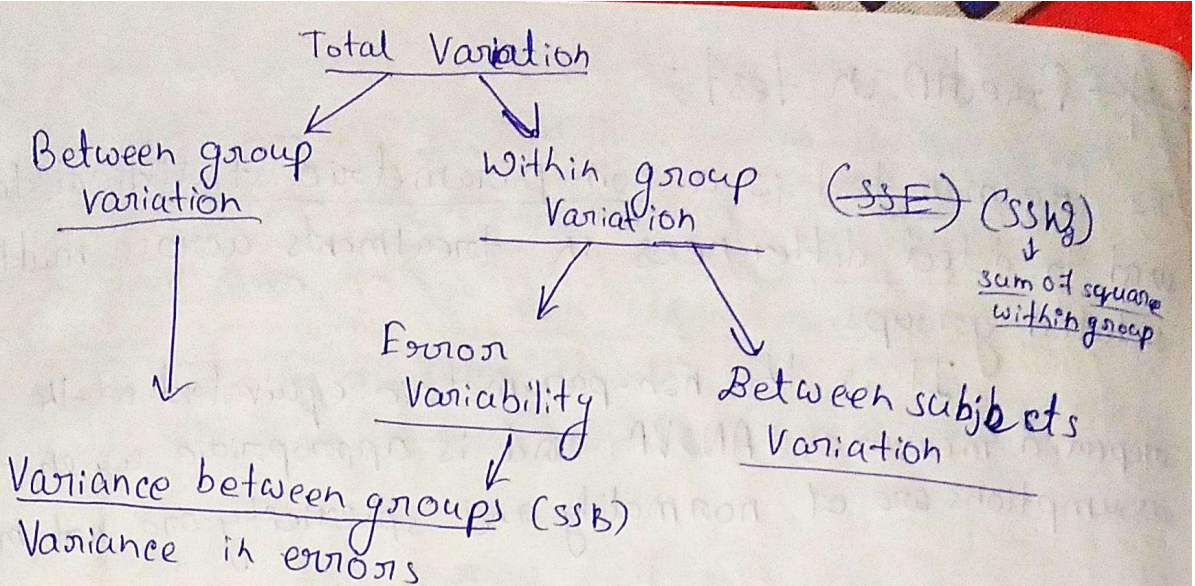
Eg:- Readings of patient before a therapy and after 2 days and 2 weeks of therapy.

⇒ 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.



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

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.

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

→ 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 in.~~

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 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.

⇒ It 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 (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 $\sum O = \sum E$ follows chi-square 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. $\sum O = \sum E$

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~~ ~~for~~ 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 $\frac{713}{2229} = 32\%$

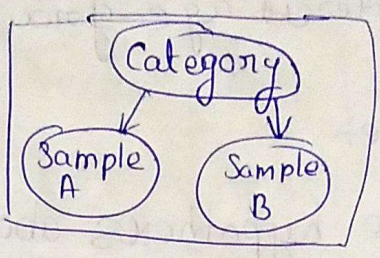
So, the expected number of surviving first class passengers is $32\% \times 325 = 104$

	First	Second	Third	Crew	Total
Survived	104	91.2	225.8	292.1	713
Died	221	193.8	480.1	620.8	1516
					2229

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 :-

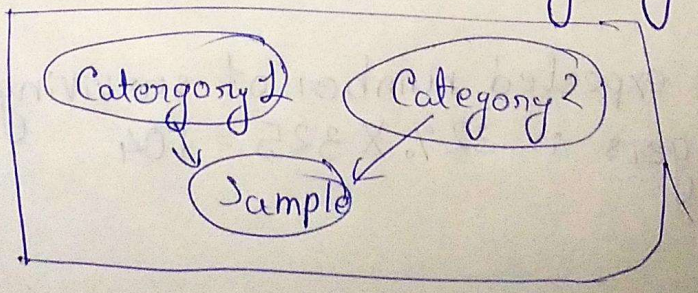
→ Purpose :-

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 ~~the~~ χ^2 test for ~~in~~ independence is same as ~~the~~ 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 categorical 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.

→ Calculations:-

	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 ~~not~~ health outcome (controls) with respect to their exposure to potential risk factors.

→ 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

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

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

→ 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) $= P(D/E) - P(D/E^c) = \frac{0.30}{} - 0.20$

$= \frac{0.10}{} = 10\%$

⇒ 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	Has Cancer: 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,

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

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]

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

Note :-

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

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