Auditing Course Material

Part 52 of 61 (Chapters 5101-5200)

1. Introduction

First, let understand all use cases, where chi-square distribution is used.

1. Finding Confidence Interval for Estimation of Variance
Chi-square distribution is employed when estimating the confidence interval for population variance. The confidence interval is given by following formula:

$$\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

2. Hypothesis Testing for Variance of Population
When conducting hypothesis tests to assess whether the variance of a population is equal to a specific value, the chi-square distribution is used. The test statistics is given by following formula:

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma^2} \quad (\text{DoF} = n-1)$$

3. Comparing Proportions of 2 or More Populations (Chi Square Test of Difference)
The chi-square test is applied for comparing proportions in multiple populations. The observed value of chi-square is calculated using below formula:

$$\chi^2_{\text{CALCULATED}} = \sum \frac{(f_i - e_i)^2}{e_i}$$

e_i = Expected Frequency, when H_0 is True
f_i = Observed Frequency.

4. Compare Proportion of Multinomial Populations (Goodness of Fit Test)
In the context of multinomial populations, the chi-square distribution is employed in a goodness-of-fit test. This test assesses whether observed frequencies in different categories match the expected frequencies, providing insights into the distribution of categorical data.

5. Chi Square Test of Independence
The chi-square test is utilized to examine the independence between two categorical variables. This test helps determine whether there is a significant association between the variables or if they are independent of each other.

We have already learnt (1) and (2) in previous sections.

Regarding (3), we have also conducted the Chi Square Test of Difference, but for only 2 populations. The next steps involve extending our analysis to situations where we compare proportions for more than 2 populations.

Then we will learn the methods involved in the goodness-of-fit test (4) and the chi-square test of independence (5).

---

2. Chi Square Test of Difference

Let us understand the chi square process for comparing more than 2 populations. It is also called Chi Square Test of Difference.

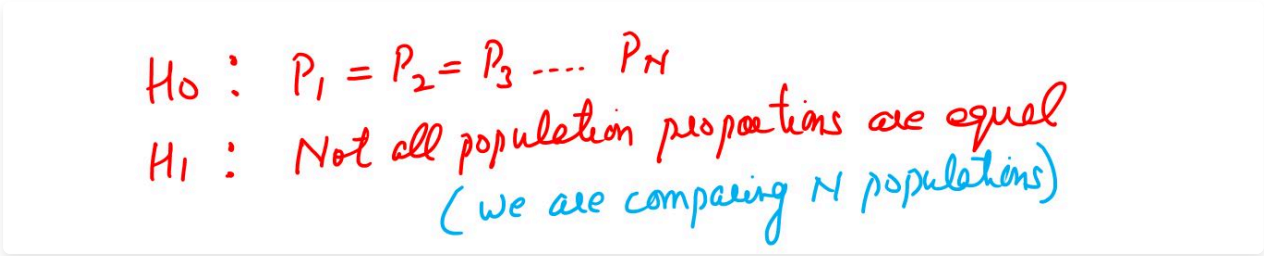
The process of comparing proportions of 3 or more populations is similar as the process of comparing propotions of 2 population.

The steps are writen below.

Step 1: Formulate Hypotheses

Null Hypothesis (H_0): There is no difference in proportions of populations.

Alternative Hypothesis (H_1): Not all proportions are same (at least one proportion differ)



Step 2: Set Significance Level

Choose a significance level (commonly denoted as α).

Step 3: Create Contingency Table

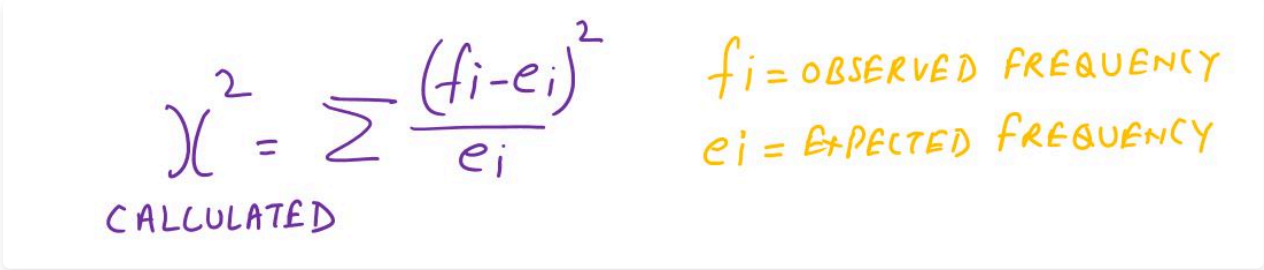
Organize your data into a contingency table. This table should have m rows (representing the m populations) and n columns (representing n categories being compared).

Step 4: Calculate Expected Frequencies

For each cell in the contingency table, calculate the expected frequency under the assumption that there is no difference in proportions of all populations. (with assumption that the null hypothesis is True).

Step 5: Calculate Chi-Square Statistic

Compute the chi-square statistic using the formula:



Step 6: Degrees of Freedom

Determine the degrees of freedom for the chi-square distribution. The degrees of freedom is (number of rows - 1) × (number of columns - 1).

Step 7: Compare Chi-Square Statistic with Critical Value

Look up the critical chi-square value for your chosen significance level and degrees of freedom in a chi-square distribution table. Compare the calculated chi-square statistic with the critical value.

Step 8: Make a Decision

If the calculated chi-square statistic is greater than the critical value, reject the null hypothesis and conclude that there is a significant difference in proportions between the two populations. If it is not, fail to reject the null hypothesis.

We can also use p-value approach.

Step 9: Interpret Results

Provide a conclusion based on the statistical analysis, considering the context of your study and the results of the test.

We will discuss these steps in the next illustration.

---

2. Chi Square Test of Difference

More shoppers do the majority of their shopping on Weekends than on Weekdays. However, is there a difference in the various age groups in the proportion of people who do the majority of their shopping on Weekends? A study showed the results for the different age groups.

	CHILDREN	ADULTS	OLD AGE
WEEKENDS	24%	28%	12%
WEEKDAYS	76%	72%	88%

Assume that 200 shoppers for each age group were surveyed. Is there evidence of a significant difference among the age groups with respect to major shopping days? (Use $\alpha = 0.05$)

Solution:

Let Children, Adults and Old Age be denoted by 1, 2 and 3.

Step 1: Formulate Hypotheses

1 = CHILDREN 2 = ADULTS 3 = OLD AGE

$H_0: P_1 = P_2 = P_3$

$H_1:$ At least one population proportion differ

Step 2: Set Significance Level

Significance Level $\alpha = 0.05$

Step 3: Create Contingency Table

	Weekends	Week days	
Children	48	152	200
Adult	56	144	200
old Age	24	176	200
	128	472	600

OBSERVED FREQUENCY TABLE

Step 4: Calculate Expected Frequencies

	Weekends	Week days	
Children	42.67	157.33	200
Adult	42.67	157.33	200
old Age	42.67	157.33	200
	128	472	600

$\frac{128}{600} \times 200$

$\frac{472}{600} \times 200$

EXPECTED FREQUENCY TABLE

Step 5: Calculate Chi-Square Statistic

CATEGORY	OBSERVED f_i	EXPECTED e_i	$(f_i - e_i)$	$\frac{(f_i-e_i)^2}{e_i}$
Children - Weekend	48	42.67	5.33	0.667
Children - Weekdays	152	157.33	-5.33	0.181
Adult - Weekend	56	42.67	13.33	4.166
Adult - Weekdays	144	157.33	-13.33	1.130
Old Age - Weekend	24	42.67	-18.67	8.167
Old Age - Weekdays	176	157.33	18.67	2.215
				16.526

$\chi^2_{CALCULATED} = 16.526$

Step 6: Degrees of Freedom

$$\text{DEGREE OF FREEDOM} = (m-1)(n-1)$$
$$= (3-1)(2-1) = 2$$

Step 7: Calculate Critical Value of Chi-Square

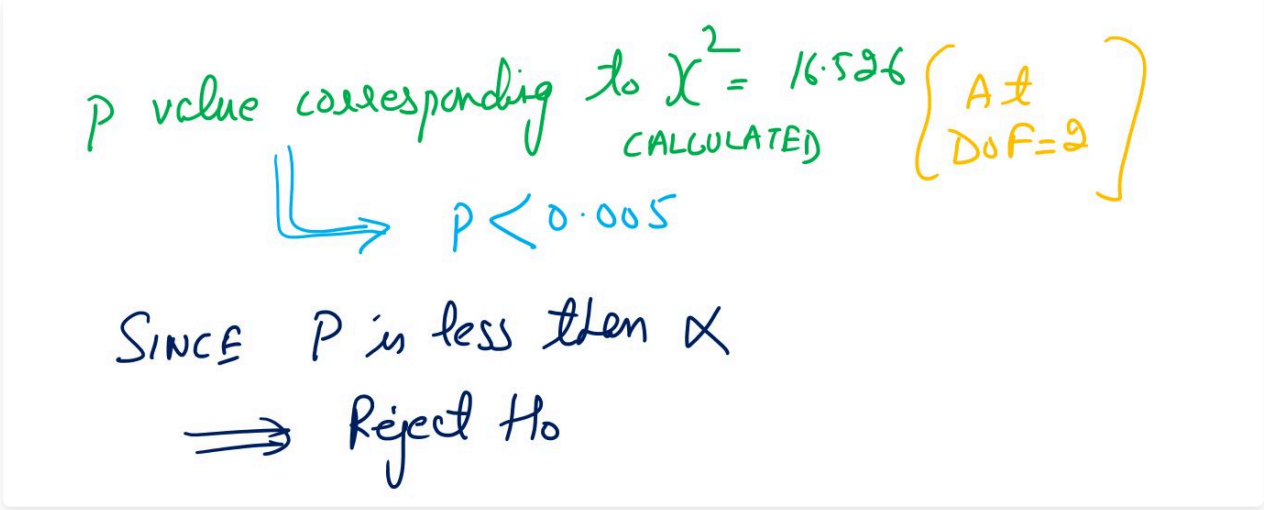
$$\text{DoF} = 2, \quad \alpha = 0.05 \Rightarrow \chi^2_{CRITICAL} = 5.991 \text{ (From TABLE)}$$

Step 8: Make a Decision

Decision with Critical Value Approach:

$$\text{SINCE } \chi^2_{CALCULATED} (16.526) > \chi^2_{CRITICAL} (5.991)$$
$$\Rightarrow \text{Reject } H_0$$

Decision with p Value Approach:



Step 9: Interpret Results

Since we rejected the null hypothesis, it means that there is enough statistical evidence to support the claim that there is significant difference among the age groups with respect to major shopping days.

However, from above analysis, we are not able to analyze, which age groups are different. We will use Marascuilo Procedure to identify, which age groups are different.

---

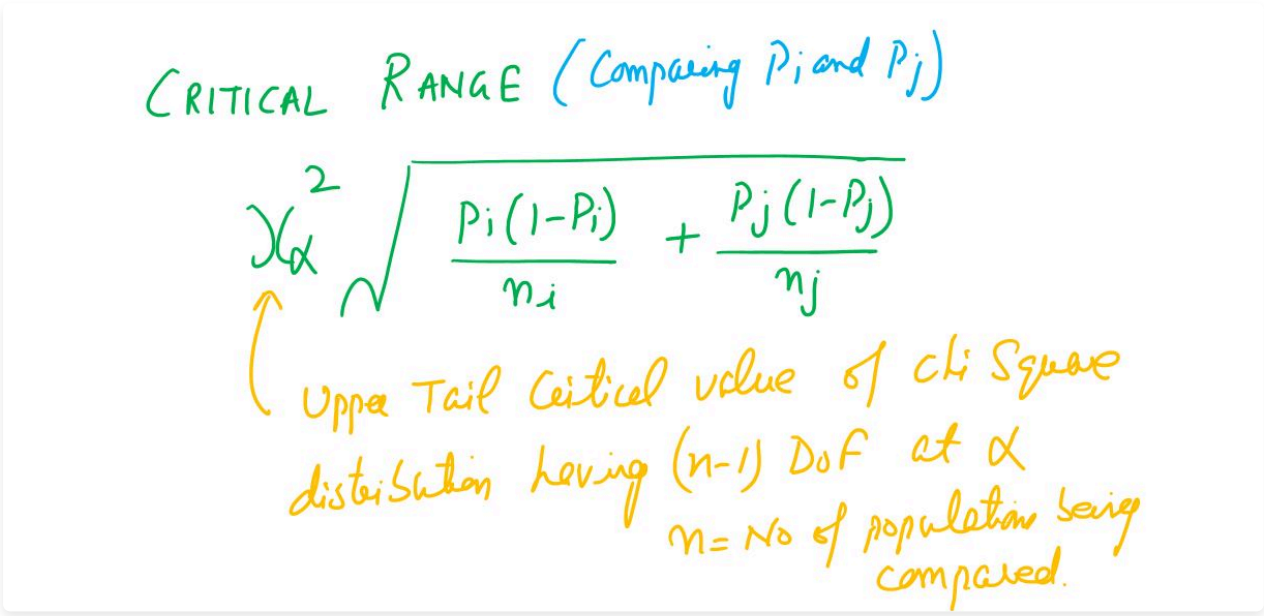
3. Marascuilo procedure

The Chi-square test for differences helps determine whether the proportions of populations are statistically similar or different. However, if the test indicates a significant difference, it does not specify which populations are distinct from each other.

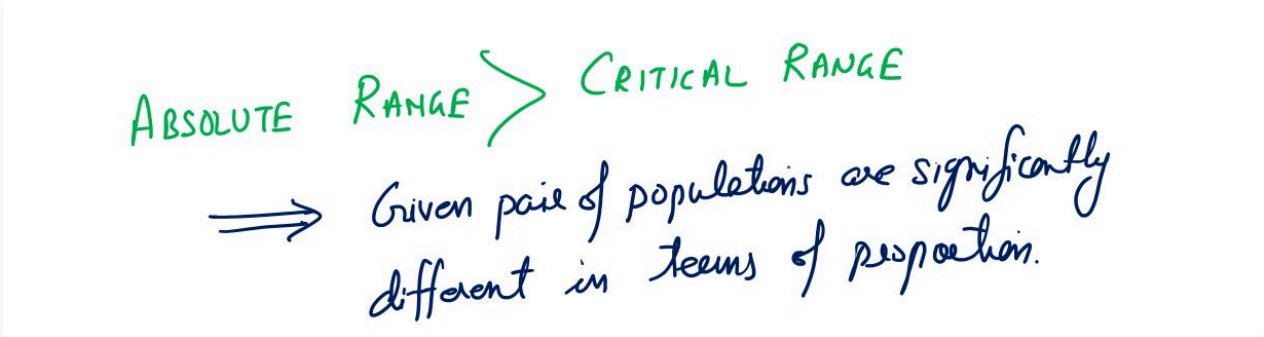
To identify the specific populations that differ, the Marascuilo procedure is employed.

This procedure involves the following steps:

1. Calculate the critical range for each pair of population proportions, for the desired level of significance. The Formula is given below:



2. Compare the absolute difference between each pair of proportions to the corresponding critical range value.



3. If the absolute difference between a pair of proportions exceeds the critical range value, conclude that the two populations are significantly different.

Let us understand this with help of one example.

---

3. Marascuilo procedure

Use the Marascuilo procedure and to determine which age groups are different in below question.

More shoppers do the majority of their shopping on Weekends than on Weekdays. However, is there a difference in the various age groups in the proportion of people who do the majority of their shopping on Weekends? A study showed the results for the different age groups.

	CHILDREN	ADULTS	OLD AGE
WEEKENDS	24%	28%	12%
WEEKDAYS	76%	72%	88%

Assume that 200 shoppers for each age group were surveyed.

Solution:

Comparing Population 1 (Children) and Population 2 (Adults):

$P_1 = 0.24$ $P_2 = 0.28$

CRITICAL VALUE $= \sqrt{\chi^2_\alpha} \sqrt{\frac{P_i(1-P_i)}{n_i} + \frac{P_j(1-P_j)}{n_j}}$

$\chi^2_\alpha = 2.447$

χ^2 at DOf = 2 and $\alpha = 0.05$ is 5.991

$= 0.1072$

ABSOLUTE VALUE, $|P_1 - P_2|$

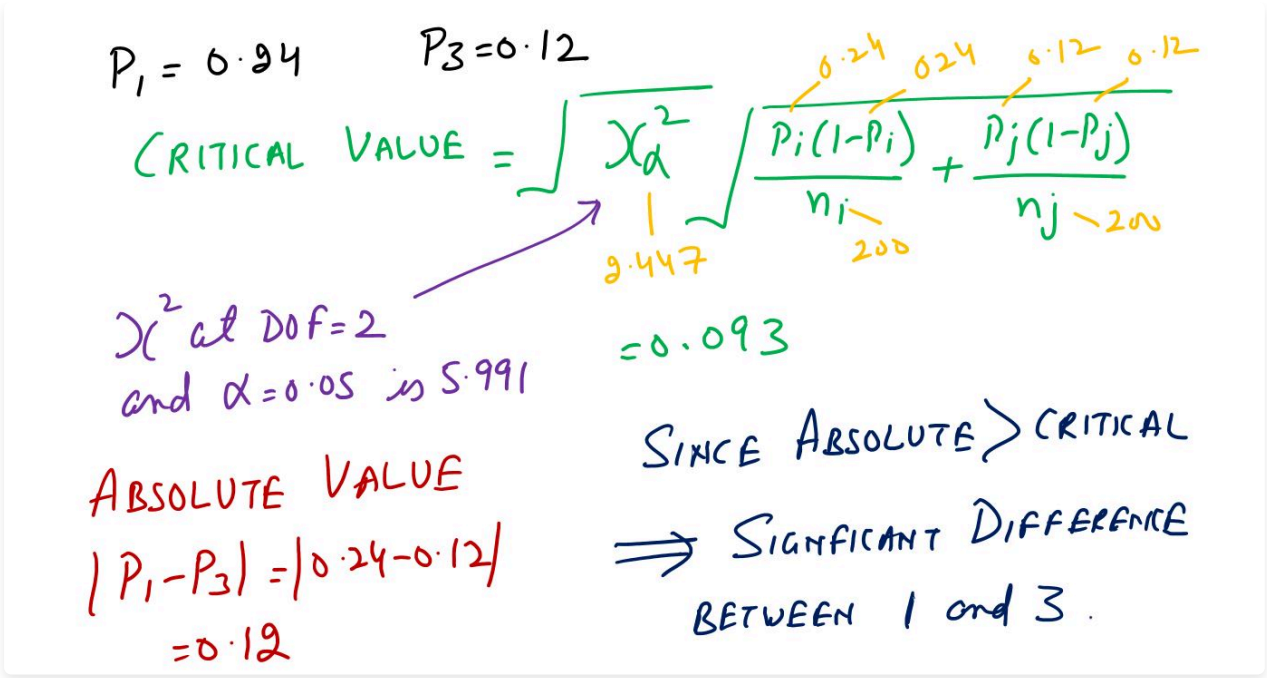
$= |0.24 - 0.28|$

$= 0.04$

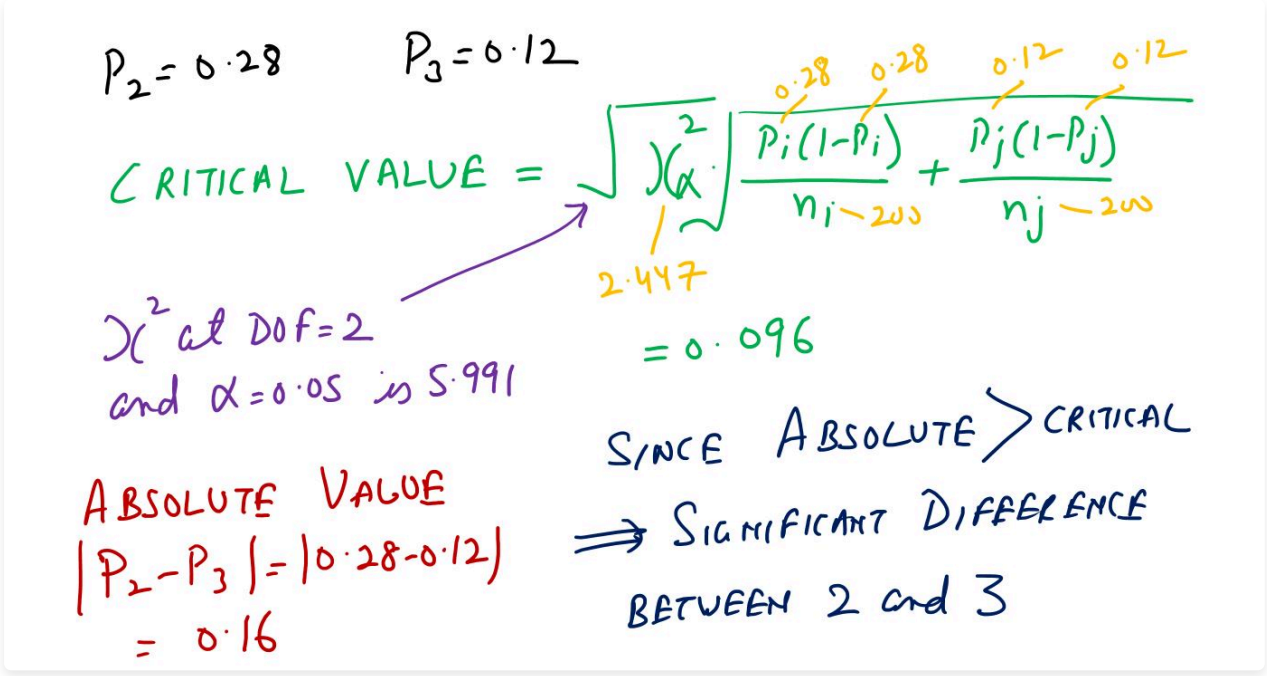
SINCE ABSOLUTE < CRITCAL

$\Rightarrow$ NO DIFFERENCE BETWEEN 1 and 2

Comparing Population 1 (Children) and Population 3 (Old Age):



Comparing Population 2 (Adults) and Population 3 (Old Age):



4. Chi-Square Goodness of Fit Test

The Chi-Square Goodness of Fit Test is utilized to compare the observed distribution of categorical data across different categories or levels against an expected distribution. This is particularly applicable when dealing with multinomial populations, where data falls into more than two categories. The test helps determine whether there are significant differences between the observed and expected frequencies.

Some of the use cases are listed below:

1. Retail Store Inventory Management
A retail store is struggling to optimize its inventory management across multiple product categories. The store wants to determine if there is a significant difference between the observed and expected sales distribution for various product types (e.g., electronics, clothing, and home goods). The goal is to identify areas where inventory adjustments may be needed to better match consumer demand.

2. Employee Training Program Effectiveness
A company has implemented different training programs for its employees across various departments. The HR department is interested in assessing whether there is a significant difference in the observed completion rates for these programs compared to the expected distribution. This information will help the company tailor future training initiatives based on departmental needs.

3. Online Learning Platform User Engagement
An online learning platform offers courses in diverse subjects and formats. The platform administrators want to investigate whether there is a significant difference in user engagement (measured by course completion rates) across different course categories such as technology, business, and humanities. This analysis will guide content curation and platform improvements.

4. Hospital Patient Satisfaction Across Departments
A hospital aims to enhance overall patient satisfaction by evaluating the experiences in various departments (e.g., emergency room, maternity, surgery). The hospital administration wants to determine if there is a statistically significant difference in the observed satisfaction scores compared to the expected distribution. This information will assist in pinpointing areas for improvement in patient care and services.

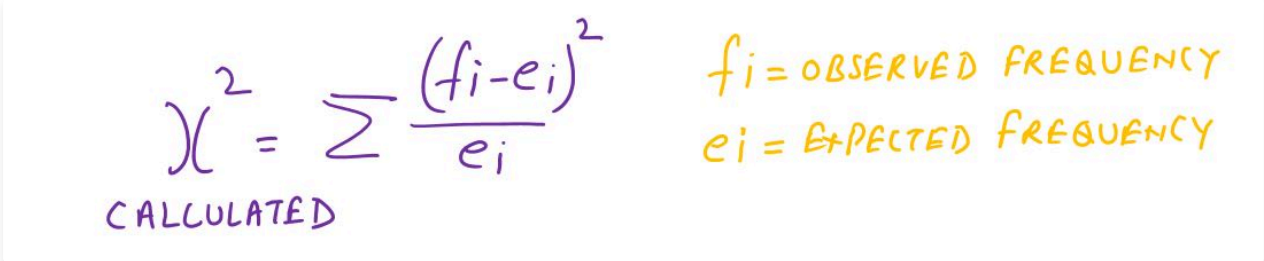
The steps of the process are given below:

1. Formulate Hypotheses
Null Hypothesis (H_0): The observed distribution is consistent with the expected distribution.
Alternative Hypothesis (H_1): There is a significant difference between the observed and expected distributions.

2. Determine Significance Level (α)
Choose a significance level, which represents the probability of rejecting the null hypothesis when it is true.

3. Calculate Observed and Expected Frequencies.
Gather observed frequencies for each category and determine the expected frequencies based on assumptions or prior knowledge.

4. Calculate Test Statistic
Use the Chi-Square formula to calculate the test statistic:



5. Determine Degrees of Freedom
Degrees of freedom (DoF) are calculated as the number of categories minus one (k-1).

6. Calcutae Test Statistic for Critical Value

Calculate Chi-Square statistic at critical value from the Chi-Square distribution table at the chosen significance level.

7. Make a Decision

If the calculated Chi-Square statistic is greater than the critical value, reject the null hypothesis, indicating a significant difference in the distributions.

8. Draw Conclusions

Conclude whether there is enough evidence to suggest a significant difference between the observed and expected distributions.

---

4. Chi-Square Goodness of Fit Test

In the dynamic landscape of digital payments, significant policy changes, including demonetization and adjustments to Merchant Discount Rates (MDR), have been implemented. The initial market share distribution among three prominent digital payment companies, namely Paytm, Phonepe, and Bharatpe, was reported as 30%, 50%, and 20%, respectively.

A survey of 200 customers engaged in digital payments revealed that 48, 98, and 54 customers preferred Paytm, Phonepe, and Bharatpe, respectively. The objective of this study is to ascertain whether the observed preferences of digital payment customers align with the originally reported market share distribution. The investigation seeks to determine if recent policy changes have had a significant impact on the market share structure of these three companies. The hypotheses is to be tested with a significance level of $\alpha = 0.05$.

Solution:

Let us denote Paytm, Phonepe, and Bharatpe with 1, 2 and 3 respectively.

1. Formulate Hypotheses

$H_0 : P_1 = 0.30 \quad P_2 = 0.50 \quad P_3 = 0.20$

H_a : Population proportion changes for at least one of P_1, P_2 or P_3.

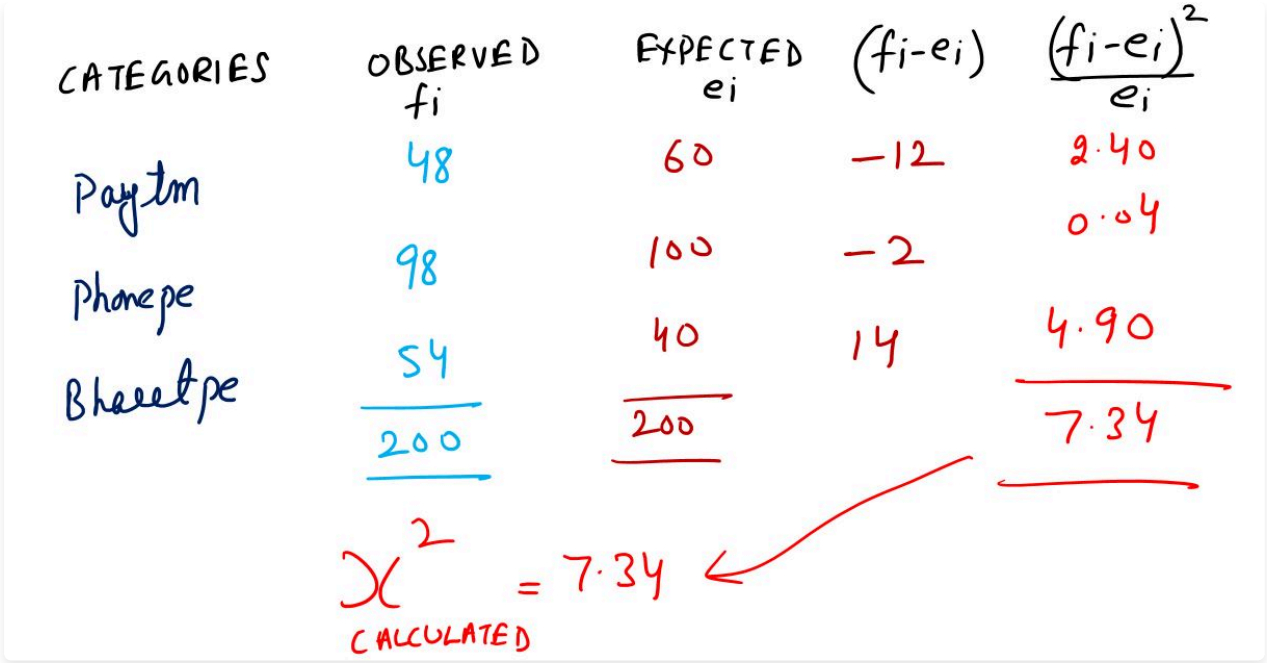
2. Determine Significance Level (α)

Level of Significance, $\alpha = 0.05$

3. Calculate Observed and Expected Frequencies

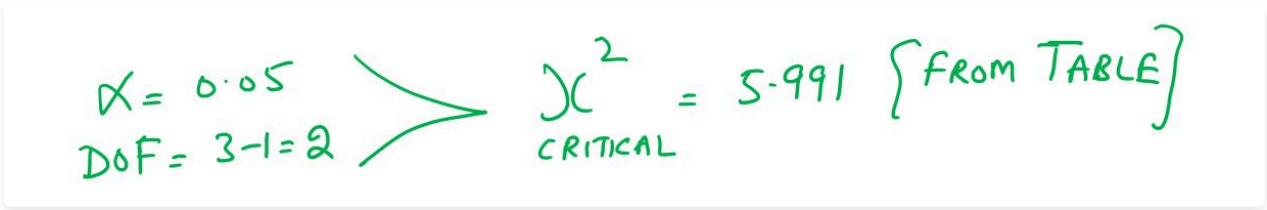
OBSERVED FREQUENCIES		EXPECTED FREQUENCIES	
Paytm	48	Paytm	$0.30 \times 200 = 60$
Phonepe	98	Phonepe	$0.50 \times 200 = 100$
Bharatpe	54	Bharatpe	$0.20 \times 200 = 40$

4. Calculate Test Statistic



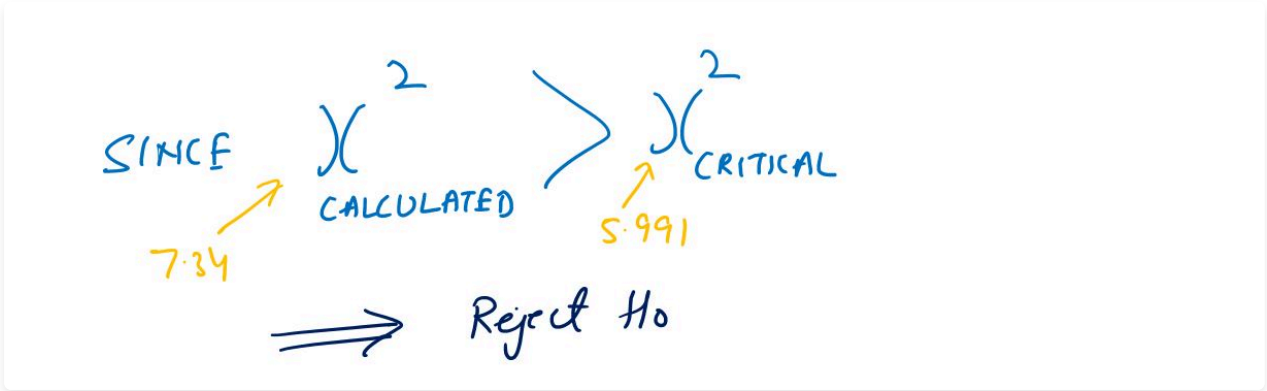
5. Determine Degrees of Freedom

6. Calculate Critical Value of Chi Square

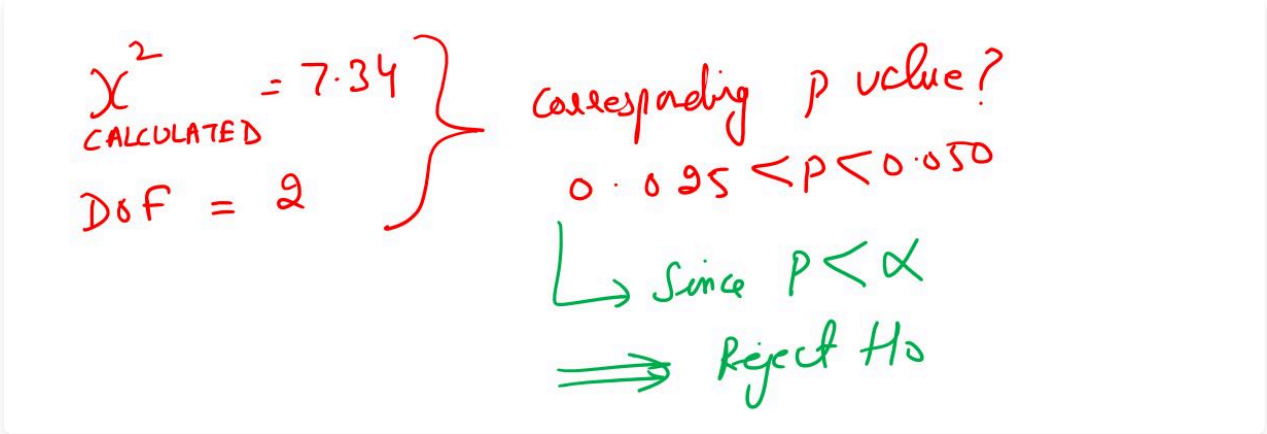


7. Make a Decision

Critical Value Approach:



p Value Approach:



8. Draw Conclusions

Since the null hypothesis is rejected with a significance level of α = 0.05, it implies that there is a significant difference between the observed preferences of digital payment customers and the initially reported market share distribution of Paytm, Phonepe, and Bharatpe.

In other words, the survey results suggest a deviation from the expected distribution, and it is reasonable to conclude that recent policy changes, such as demonetization and adjustments to Merchant Discount Rates (MDR), have had a notable impact on the market share structure of these three digital payment companies.

4. Chi-Square Goodness of Fit Test

The RBI collects data on the use of credit cards, debit cards, mobile (UPI, wallets etc), and cash when consumers pay for in-store purchases. In 2018, the reported usages is given in the table.
A sample taken in 2022 found that for 220 in-stores purchases, 46 used a credit card, 67 used a debit card, 33 used mobile, and 74 used cash.

MODE	CREDIT	DEBIT	MOBILE	CASH
PERCENTAGE	22%	21%	18%	39%

(i) At α = .01, can we conclude that a change occurred in how customers paid for in-store purchases over the four-year period from 2018 to 2022? What is the p-value?
(ii) Compute the percentage of use for each method of payment using the 2022 sample data. What appears to have been the major change or changes over the four-year period?
(iii) In 2022, what percentage of payments was made using plastic (credit card or debit card)?

Solution:

Let us denote credit cards, debit cards, mobile (UPI, wallets etc), and cash with 1, 2, 3 and 4 respectively.

1. Formulate Hypotheses

H_0: $P_1 = 0.22$, $P_2 = 0.21$, $P_3 = 0.18$, $P_4 = 0.39$
H_a: Population Proportions are not equal to these values in null hypothesis

2. Determine Significance Level (α)

SIGNIFICANCE LEVEL, $\alpha = 0.01$

3. Calculate Observed and Expected Frequencies

	OBSERVED FREQUENCIES	EXPECTED FREQUENCIES
CREDIT	46	$220 \times 0.22 = 48.4$
DEBIT	67	$220 \times 0.21 = 46.2$
MOBILE	33	$220 \times 0.18 = 39.6$
CASH	74	$220 \times 0.39 = 85.8$

4. Calculate Test Statistic

CATEGORIES	OBSERVED f_i	EXPECTED e_i	$(f_i - e_i)$	$\frac{(f_i - e_i)^2}{e_i}$
CREDIT	46	48.4	−2.4	0.12
DEBIT	67	46.2	20.8	9.36
MOBILE	33	39.6	−6.6	1.10
CASH	74	85.8	−11.8	1.62
				12.21

$\chi^2_{CALCULATED} = 12.21$

5. Determine Degrees of Freedom

Degree of Freedom = $4 - 1 = 3$

6. Calculate Critical Value of Chi Square

$\alpha = 0.01 \implies \chi^2_{CRITICAL} = 11.345$

DoF = $4 - 1 = 3$

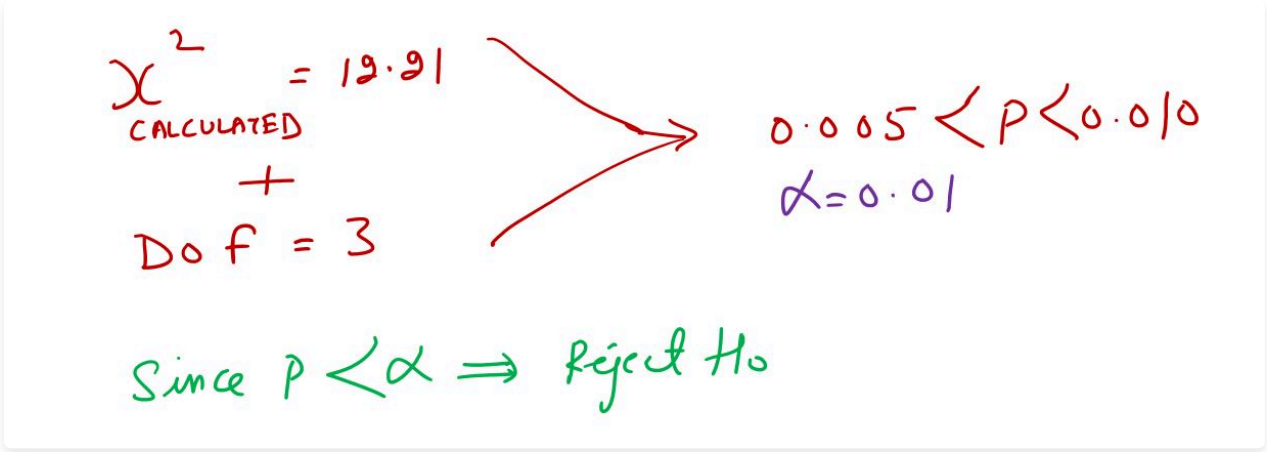
7. Make a Decision

Critical Value Approach:

SINCE $\chi^2_{CALCULATED} > \chi^2_{CRITICAL} \implies$ Reject H_0

$\chi^2_{CALCULATED} = 12.21$, $\chi^2_{CRITICAL} = 11.345$

p Value Approach:

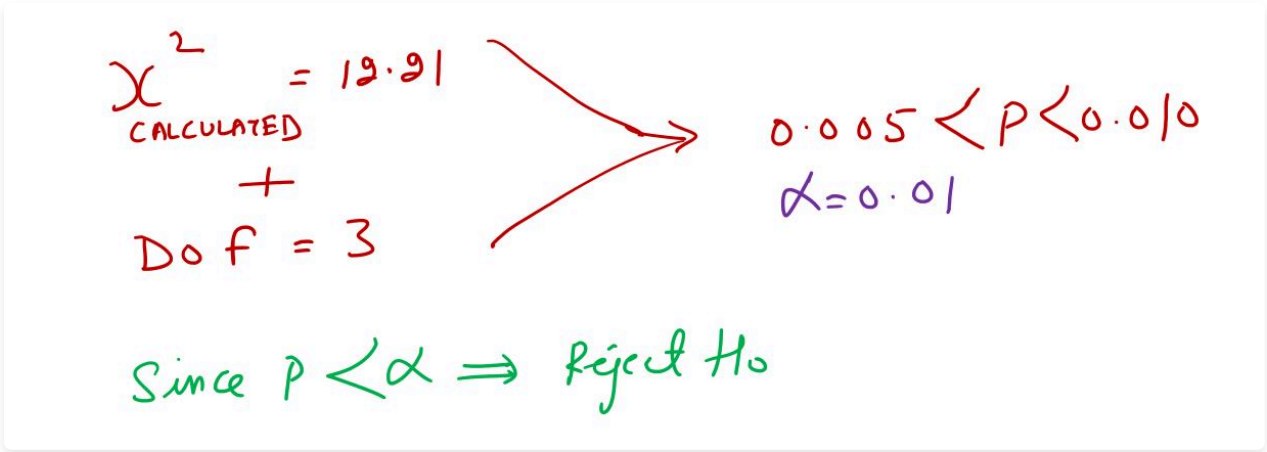


8. Draw Conclusions

Since we decided that the null hypothesis is rejected, it implies that there is statistical evidence to support the claim that there has been a significant change in how customers paid for in-store purchases from 2018 to 2022. The rejection suggests that the distribution of payment methods has shifted, and the observed data is not consistent with the idea that there is no change. The specific nature of this change would need to be explored further by examining the percentages and patterns of payment methods.

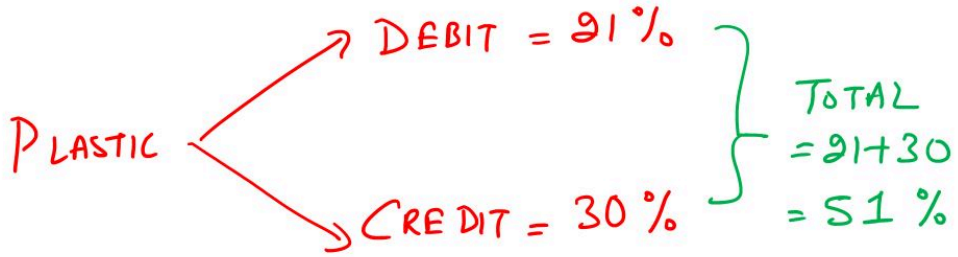
(i) At α = .01, can we conclude that a change occurred in how customers paid for in-store purchases over the four-year period from 2018 to 2022? What is the p-value?

Yes, we can conclude that a change occurred in how customers paid for in-store purchases over the four-year period from 2018 to 2022. The p value is calculated below.



(ii) Compute the percentage of use for each method of payment using the 2022 sample data. What appears to have been the major change or changes over the four-year period?

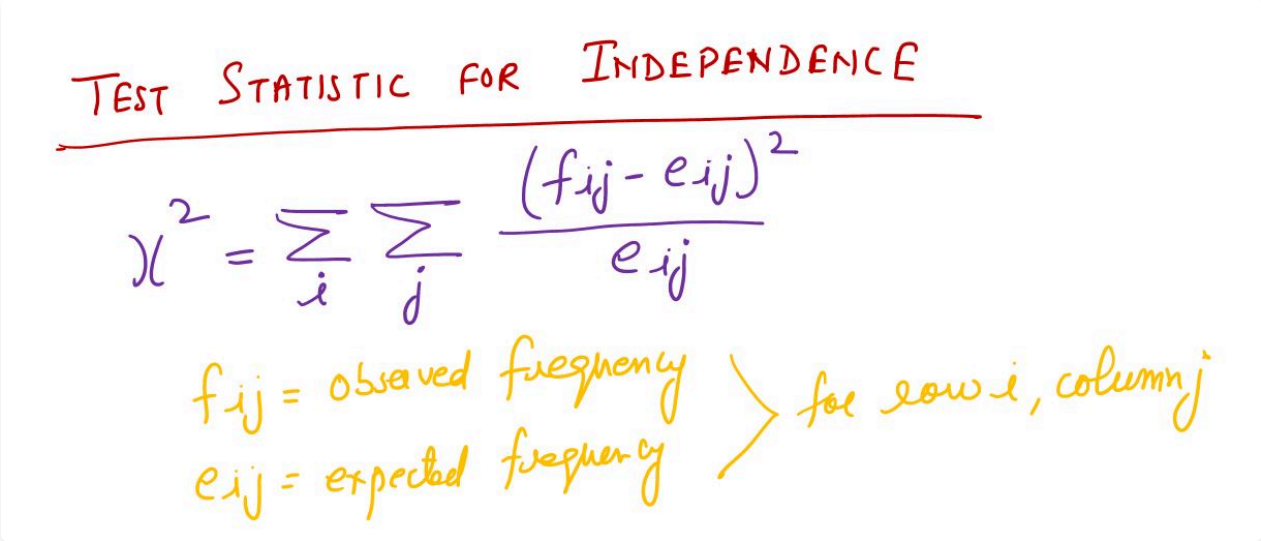
(iii) In 2022, what percentage of payments was made using plastic (credit card or debit card)?


PLASTIC
DEBIT = 21 %
CREDIT = 30 %
TOTAL
= 21 + 30
= 51 %

5. Chi-Square Test for Independence

The chi-square distribution is not only useful for testing the goodness of fit but also plays a crucial role in examining the independence of two variables in sample data. This test aims to determine whether there is a significant relationship between two variables. The process involves selecting a sample and creating a cross-tabulation to simultaneously summarize data for both variables.

The test for independence is a one-tailed test, and the rejection region is located in the upper tail of the chi-square distribution. This means that statistical significance is assessed based on the right side of the distribution.



This type of test is often referred to as a contingency table test because it utilizes a contingency table format to organize the data. The number of degrees of freedom for the chi-square distribution used in this test is calculated by multiplying the number of rows minus 1 by the number of columns minus 1.

Unlike the goodness of fit test, which focuses on a single variable, the test of independence delves into the relationship between two variables. The examination is conducted through a contingency table, and the objective is to determine whether the observed pattern in the table is strong enough to indicate a dependence between the two variables.

Let us look at a few examples to understand the difference between these tests.

Scenario 1
Conducting a survey to analyze whether there is an association between age group and preferred shopping channels (online, in-store, or both). An independence test is used to examine if age group and shopping channel preference are related or independent, providing insights for targeted marketing strategies.

Scenario 2
A manufacturing plant produces bolts, and the expected distribution of lengths is known. The goal is to check if the observed lengths of a sample match the expected distribution. We use a goodness of fit test to determine if the observed frequencies of bolt lengths align with the expected distribution, ensuring quality control.

Scenario 3
Investigating whether there is a relationship between students' educational performance (e.g., grades) and their socioeconomic status (low, middle, high). An independence test is employed to understand if there is a significant association between educational performance and socioeconomic status.

Scenario 4
In a genetics study, researchers expect a specific ratio of genotypes based on Mendelian inheritance. They want to verify if the observed genotypes from a population follow the expected ratios. Genetic studies often involve comparing observed genotypes to expected ratios to assess adherence to Mendelian laws, and a goodness of fit test is suitable for this purpose.

Scenario 5
Analyzing survey responses to determine if there is a relationship between customer satisfaction levels and preferences for

specific product features. An independence test helps assess whether customer satisfaction and product feature preferences are independent or if changes in one are associated with changes in the other.

Key Considerations:

Goodness of Fit Tests: Applied when assessing how well observed data fits an expected distribution or model for a single categorical variable.

Independence Tests: Used to examine whether two categorical variables are associated or independent, providing insights into relationships between variables in a contingency table.

---

5. Chi-Square Test for Independence

Bira manufactures and distributes 3 types of beer: light, regular, and dark. In an analysis of the market segments for the 3 beers, the firm's market research group raised the question of whether preferences for the three beers differ among male and female beer drinkers.

If beer preference is independent of the gender of the beer drinker, one advertising campaign will be initiated for all of Bira's beers. However, if beer preference depends on the gender of the beer drinker, the firm will tailor its promotions to different target markets.

A simple random sample of 150 beer drinkers is selected. After tasting each beer, the individuals in the sample are asked to state their preference or first choice. The table summarizes the responses for the study. Of the 150 individuals in the sample, 20 were men who favored light beer, 40 were men who favored regular beer, 20 were men who favored dark beer.

Solution:

1. Formulate Hypotheses

H_0 : Beer Preference is independent of gender

H_1 : Beer Preference is not independent of gender

2. Determine Significance Level (α)

Level of Significance, $\alpha = 0.05$

3. Calculate Observed and Expected Frequencies

OBSERVED FREQUENCY

	LIGHT	REGULAR	DARK	
MALE	20	40	20	80
FEMALE	30	30	10	70
	50	70	30	150

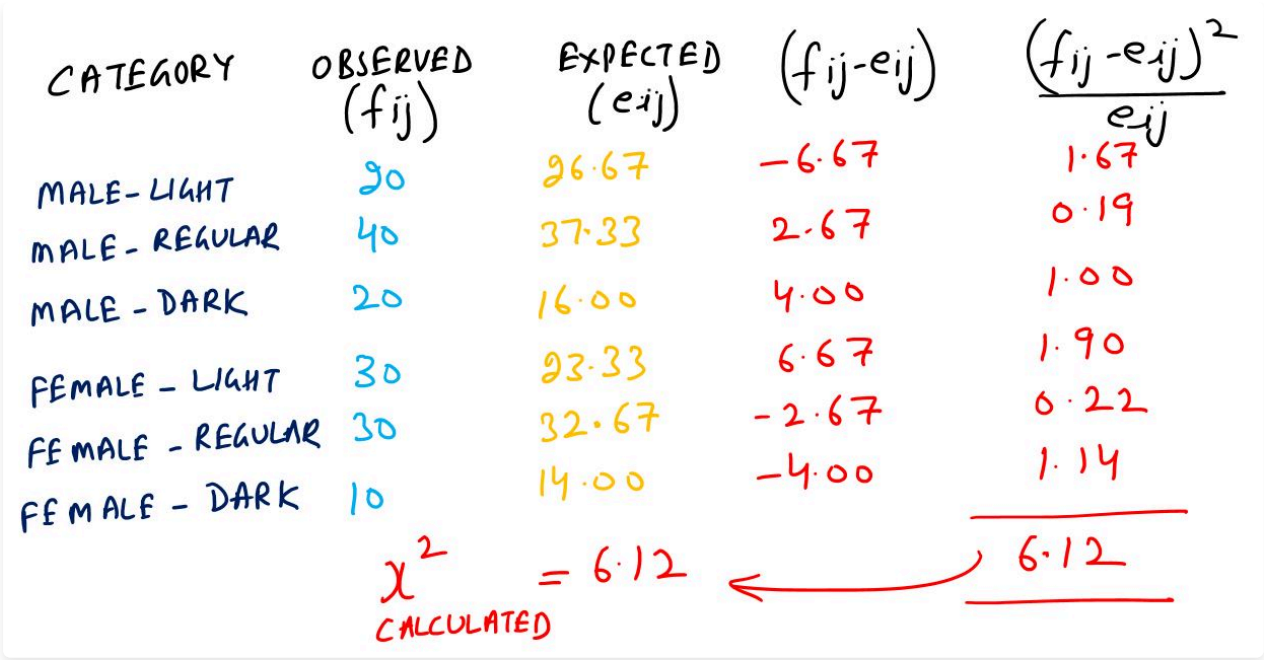
$\frac{50}{150} \times 80$ $\frac{70}{150} \times 80$ $\frac{30}{150} \times 80$

	LIGHT	REGULAR	DARK	
MALE	26.67	37.33	16.00	80
FEMALE	23.33	32.67	14.00	70
	50	70	30	150

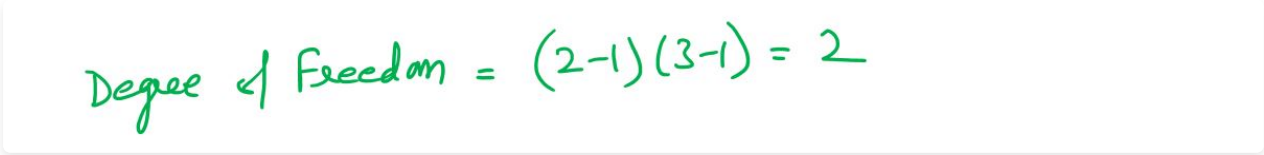
EXPECTED FREQUENCY

$\frac{50}{150} \times 70$ $\frac{70}{150} \times 70$ $\frac{30}{150} \times 70$

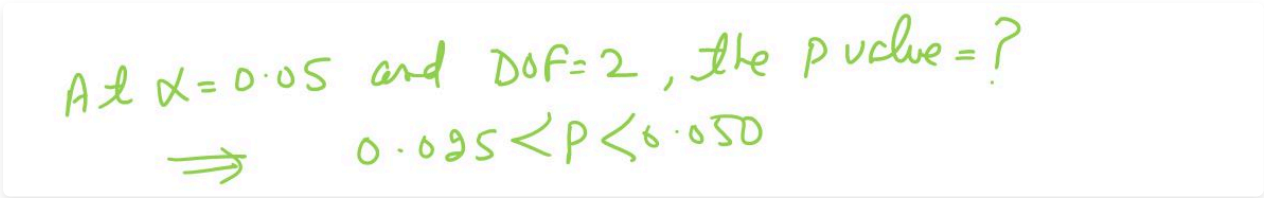
4. Calculate Test Statistic



5. Determine Degrees of Freedom



6. Calculate p Value



7. Make a Decision

8. Draw Conclusions

Since we rejected the null hypothesis of independence and conclude that beer preference is not independent of the gender of the beer drinker.

Although no further conclusions can be made as a result of the test, we can compare the observed and expected frequencies informally to obtain an idea about the dependence between beer preference and gender. We see that male beer drinkers have higher observed than expected frequencies for both regular and dark beers, whereas female beer drinkers have a higher observed than expected frequency only for light beer. These observations give us insight about the beer preference differences between male and female beer drinkers.

---

5. Chi-Square Test for Independence

A sample of 200 persons were interviewed to find out, if they are happy in life. Conduct a test of independence to determine whether the effect on happiness is independent of the age of the person. Use α = 0.05. What is the p-value, and what is your conclusion? Use row percentages to learn more about how age affects happiness. What is your conclusion?

	Happy	Neutral	Sad	Total
0-30 years	26	50	14	90
30-60 years	16	27	17	60
above 60 years	11	19	20	50
Total	53	96	51	**200**

Solution:

1. Formulate Hypotheses

H_0: Happiness is independent of age

H_a: Happiness is not independent of age

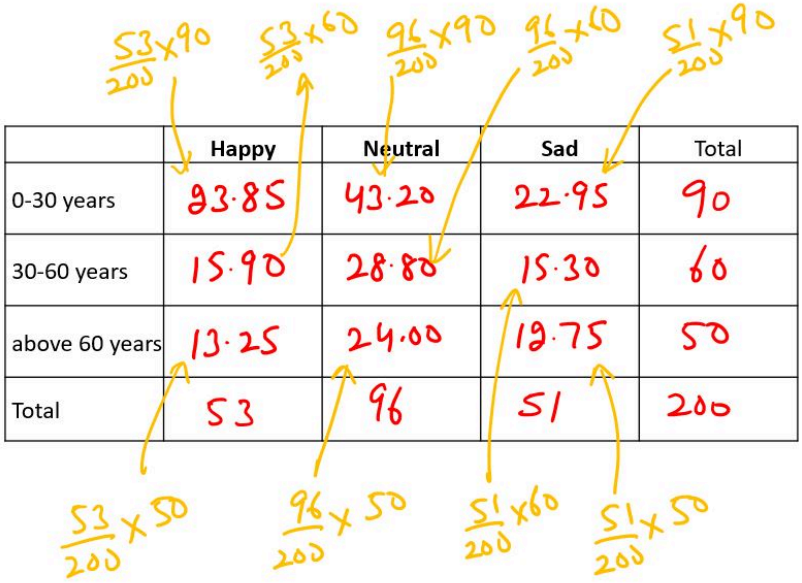
2. Determine Significance Level (α)

Level of significance, $\alpha = 0.05$

3. Calculate Observed and Expected Frequencies

	Happy	Neutral	Sad	Total
0-30 years	26	50	14	90
30-60 years	16	27	17	60
above 60 years	11	19	20	50
Total	53	96	51	200

OBSERVED FREQUENCIES



	Happy	Neutral	Sad	Total
0-30 years	93.85	43.20	22.95	90
30-60 years	15.90	28.80	15.30	60
above 60 years	13.25	24.00	12.75	50
Total	53	96	51	200

EXPECTED FREQUENCIES

4. Calculate Test Statistic

f_i e_i $\sum \frac{(f_i - e_i)^2}{e_i}$

	Observed Frequency	Expected Frequency	
0-30 years- Happy	26	23.85	0.194
0-30 years- Neutral	50	43.20	1.070
0-30 years- Sad	14	22.95	3.490
30-60 years- Happy	16	15.90	0.001
30-60 years- Neutral	27	28.80	0.112
30-60 years- Sad	17	15.30	0.189
above 60 years- Happy	11	13.25	0.382
above 60 years- Neutral	19	24.00	1.041
above 60 years- Sad	20	12.75	4.123

$10.602 = \chi^2$ CALCULATED

5. Determine Degrees of Freedom

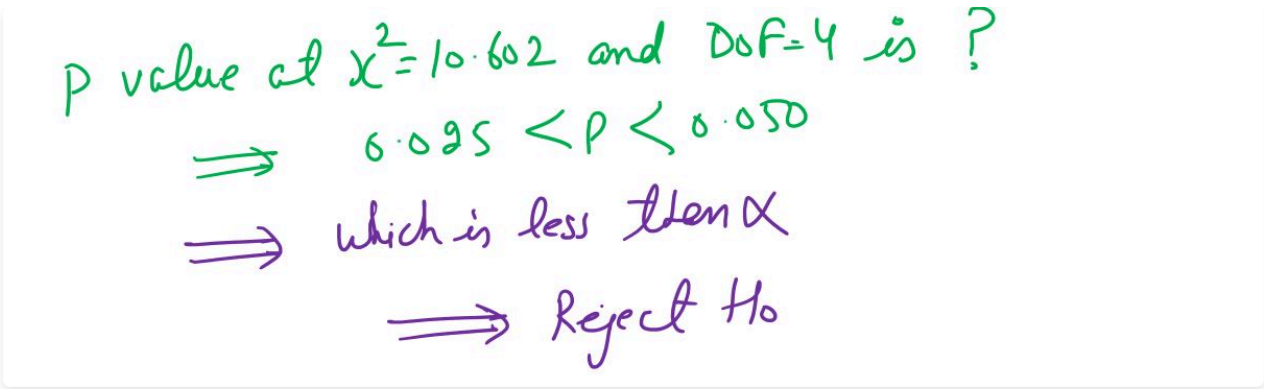
Degree of Freedom $= (3-1)(3-1) = 4$

6. Calculate p Value

p value at $\chi^2 = 10.602$ and DoF = 4 is ?

$\Rightarrow 0.025 < p < 0.050$

7. Make a Decision



8. Draw Conclusion

The rejection of the null hypothesis implies that there is a significant association between the effect on happiness and the age of the person. In other words, age and happiness are not independent of each other.

We can also conclude from row percentages that as age increases, the sadness increases and happiness decreases.

---

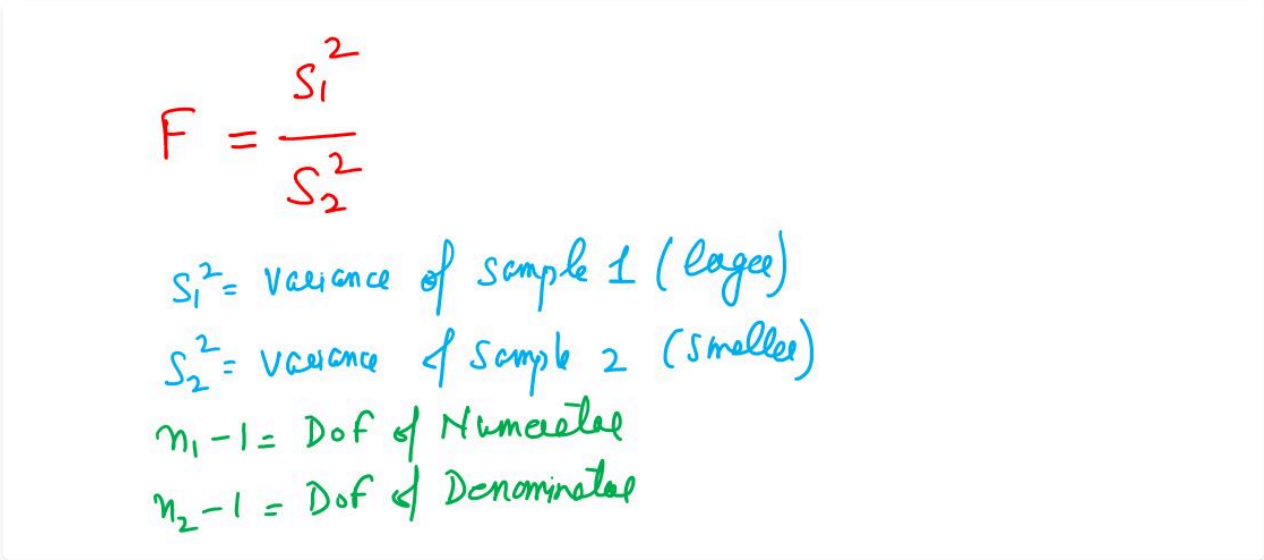
1. Variances of Two Populations

In some statistical applications, we may want to compare two population variances, for example, product quality resulting from two different production processes, the variances in assembly times for two assembly methods, or the variances in temperatures for two heating devices.

One use case of testing the variance difference is to decide whether to employ the pooled-variance t-test (assuming equal variances) or the separate-variance t-test (not assuming equal variances) when comparing the means of these populations.

The test for the difference between the variances of two independent populations is based on the ratio of the two sample variances.

If you assume that each population is normally distributed, then the ratio \(\frac{s_1^2}{s_2^2} \) follows the F distribution.



The critical values of the F distribution depend on the degrees of freedom in the two samples. The degrees of freedom in the numerator of the ratio are for the first sample, and the degrees of freedom in the denominator are for the second sample.

The first sample (denoted as 1) taken from the first population is defined as the sample that has the larger sample variance. The second sample (denoted as 2) taken from the second population is the sample with the smaller sample variance.

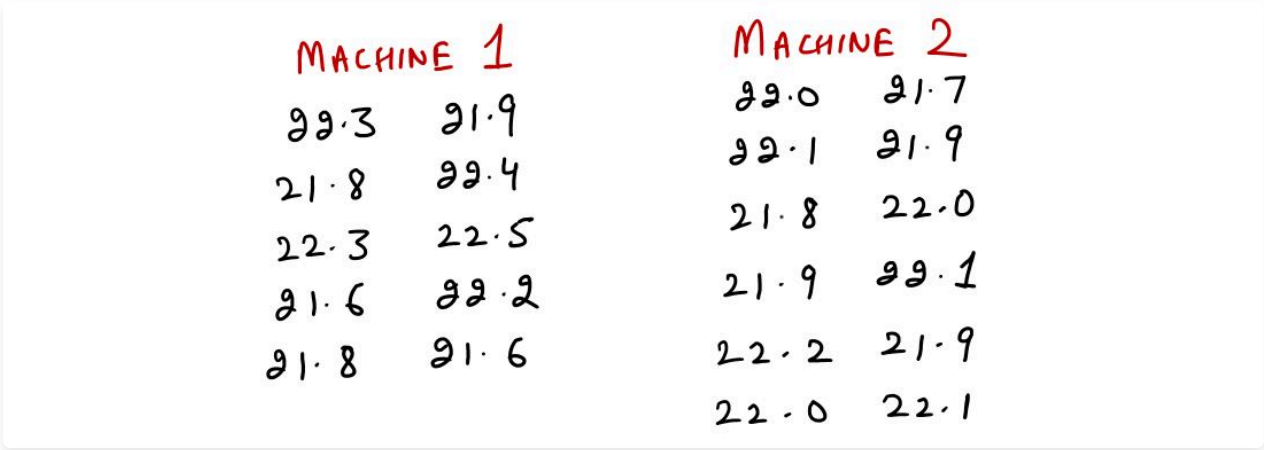
Let us understand this method, with a few illustrations.

---

1. Variances of Two Populations

Suppose a machine produces metal sheets that are specified to be 22 millimeters thick. Because of the machine, the operator, the raw material, the manufacturing environment, and other factors, there is variability in the thickness. Two machines produce these sheets.

Operators are concerned about the consistency of the two machines. To test consistency, they randomly sample 10 sheets produced by machine 1 and 12 sheets produced by machine 2. The thickness measurements of sheets from each machine are given in the table.



Assume sheet thickness is normally distributed in the population.
How can we test to determine whether the variance from each sample comes from the same population variance (population variances are equal) or from different population variances (population variances are not equal)? Let the level of significance to be 0.05.

Solution:

The test for the difference between the variances of two independent populations, often known as the F-test for variances, involves following steps.

1. Formulate Hypotheses

2. Choose Level of significance

3. From sample data calculate 'Observed' value of F

MACHINE 1	
22.3	21.9
21.8	22.4
22.3	22.5
21.6	22.2
21.8	21.6

$n_1 = 10$

$S_1^2 = 0.11378$

MACHINE 2	
22.0	21.7
22.1	21.9
21.8	22.0
21.9	22.1
22.2	21.9
22.0	22.1

$n_2 = 12$

$S_2^2 = 0.02023$

$$F = \frac{S_1^2}{S_2^2} = \frac{0.11378}{0.02023} = 5.62 \Leftarrow \text{Observed value of } F$$

4. Identify 'Critical' value of F from table

SINCE TWO TAILED →

$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

DoF (Numerator) = 10 − 1 = 9

DoF (Denominator) = 12 − 1 = 11

$F(1-\frac{\alpha}{2}, 9, 11) = 0.28$

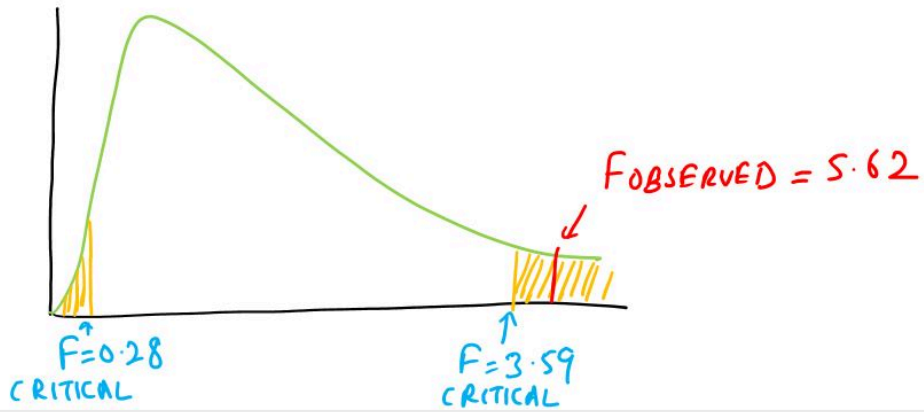
$F_{CRITICAL}$ → less than 0.28 / More than 3.59

$F(\frac{\alpha}{2}, 9, 11) = 3.59$

5. Make Decision

'Observed F' is greater than 'Upper Critical F'

⟹ Reject H_0



6. Interpret Decision

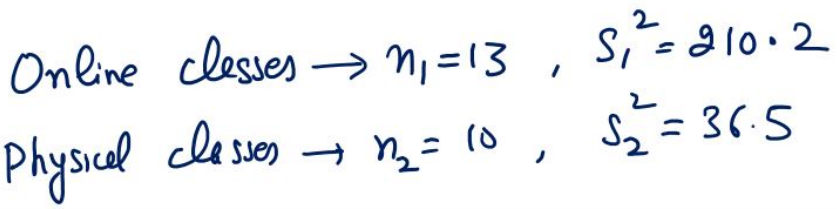
Since we rejected the null hypotheses, the population variances are not equal.

An examination of the sample variances reveals that the variance from machine 1 measurements is greater than that from machine 2 measurements. The operators and process managers might want to examine machine 1 further; an adjustment may be needed or some other reason may be causing the seemingly greater variations on that machine.

---

1. Variances of Two Populations

A coaching institute would like to determine whether there is more variability in the final exam scores of students taking the online classes in comparison to students taking the physical classes. Random samples of 13 online classes students and 10 physical classes students are selected. Following results are obtained from these sample.

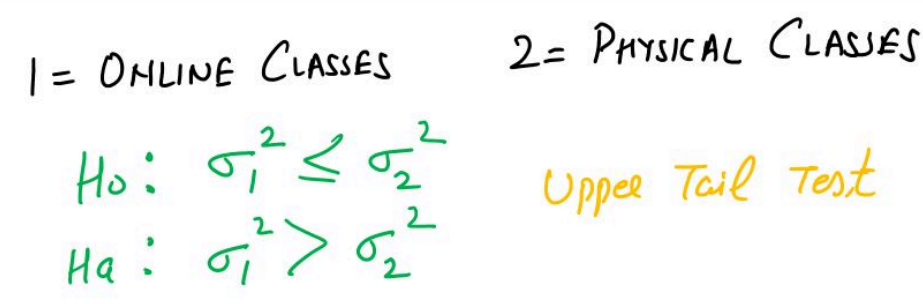


At the 0.05 level of significance, is there evidence that there is more variability in the final exam scores of students taking online classes than the students taking physical classes? Assume that the population final exam scores are normally distributed.

Solution:

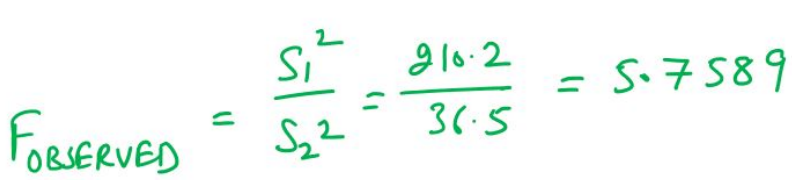
The test for the difference between the variances of two independent populations, often known as the F-test for variances, involves following steps.

1. Formulate Hypotheses

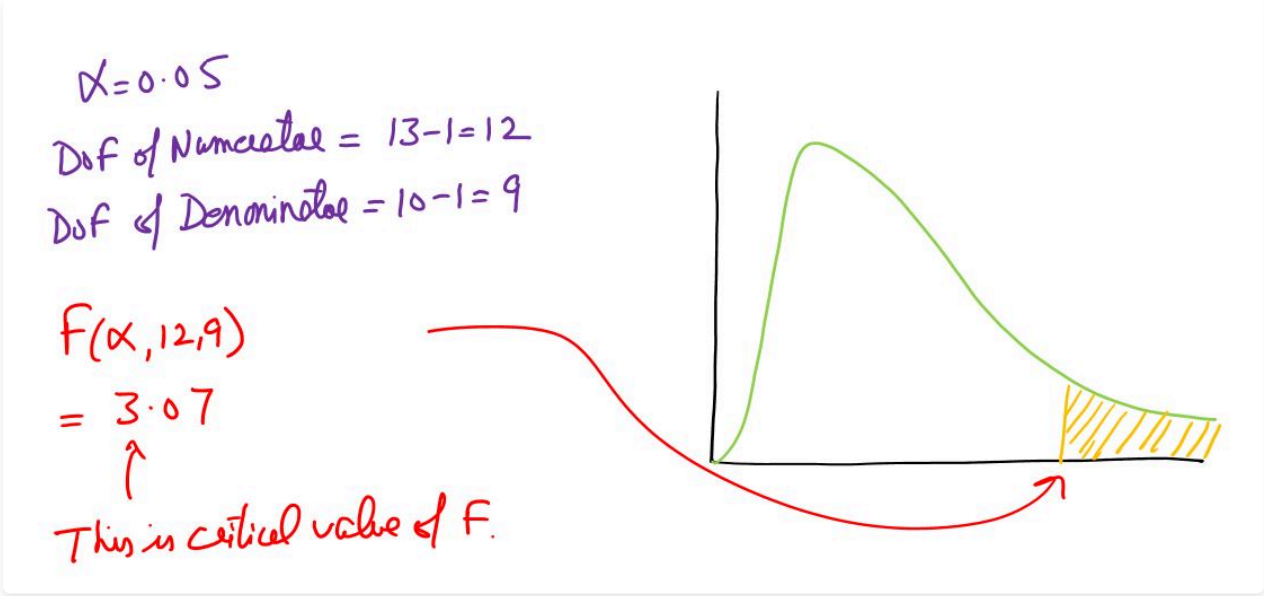


2. Choose Level of significance

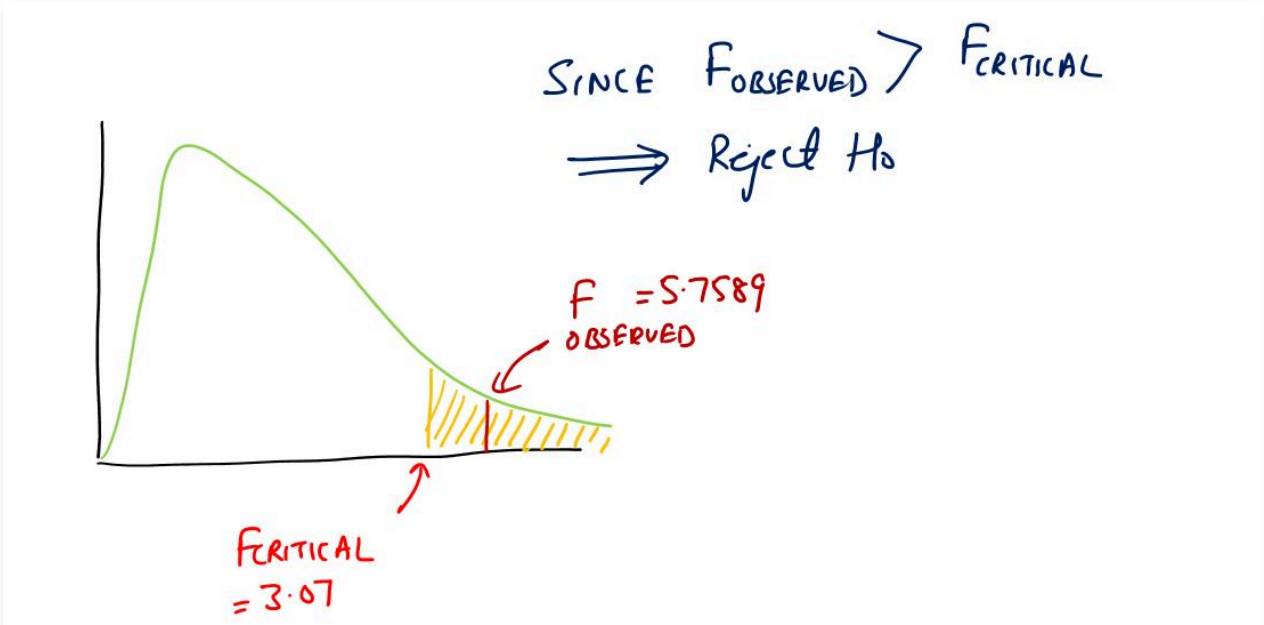
2. From sample data calculate 'Observed' value of F



4. Identify 'Critical' value of F from table



5. Make Decision



6. Interpret Decision

Rejecting the null hypothesis in this problem implies that there is statistical evidence at the 0.05 level of significance to suggest that there is more variability in the final exam scores of students taking online classes compared to students taking physical classes. In practical terms, it indicates that the spread or dispersion of final exam scores among online classes students is significantly different or greater than that among physical classes students.

---

1. Introduction

ANOVA, or Analysis of Variance, is a statistical technique used to compare means across multiple groups or populations. While the primary intention in ANOVA is to compare population means, the method achieves this comparison by analyzing the variances observed within and between these groups.

The fundamental concept behind ANOVA lies in partitioning the total variance observed in the data into different components. It separates the overall variability in the data into two main sources: variation between the group means and variation within each group.

By comparing these variances, ANOVA determines whether the observed differences between the group means are statistically significant or if they could likely occur due to random fluctuations within the groups. If the variation between group means is significantly larger than the variation within each group, it suggests that there are genuine differences among the population means.

Let us understand some key terms, before we understand process of ANOVA.

---

2. Key Terms

Let us understand the key concepts used in ANOVA—dependent variable, independent variable, and levels of factors—using an example to illustrate their meanings.

Suppose we are comparing mean intlligence levels score of calss X students at three schools run by Delhi Public School, DPS Panipat, DPS Delhi and DPS Mathura.

1. Dependent Variable

The dependent variable, also referred to as the response variable, signifies the outcome or measure being observed or assessed in a study. It is the aspect that is influenced or impacted by changes in the independent variable(s).
Example: Intelligence levels score—this is the metric or measure that's being evaluated or observed among the students.

2. Independent Variable

The independent variable, or the factor, is the variable under study that is manipulated or controlled. It's the variable that influences or affects the dependent variable.
Example: 'City' is the independent variable.

3. Levels of Factors

Levels of factors pertain to the specific categories or groups within the independent variable. They are also called classification levels, treatments, or groups.
Example: DPS Panipat, DPS Delhi, and DPS Mathura under the category of cities are levels of factors

Another Example

Similarly, consider the example of comparing the average revenue generated per day for various Uber services such as Uber X, Uber Go, and Uber Moto.

1. *Dependent Variable*: The dependent variable is the mean revenue generated per day. It's the measure or outcome being assessed and is expected to be influenced by the type of Uber service.

2. *Independent Variable*: The independent variable is the type of Uber service provided, representing the factor under investigation that may impact the mean revenue per day for each service.

3. *Levels of Factors*: Levels of factors refer to the distinct categories or conditions within the independent variable. In this case, the levels of factors are the different types of Uber services - Uber X, Uber Go, and Uber Moto. These levels represent the specific categories being compared to understand their influence on daily revenue.

One way and Two way ANOVA

When analyzing the impact of a single factor, it's termed as one-way ANOVA. However, if the analysis involves the comparison of two distinct factors, it's referred to as two-way ANOVA.

---

3. Assumptions of ANOVA

Three essential assumptions are necessary for the application of analysis of variance (ANOVA).

1. The response variable for each population adheres to a normal distribution. Considering DPS schools example, the Intelligence levels score must be normally distributed at all schools.

2. The variance (σ^2) of the response variable remains consistent across all populations. This condition is known as homogenity of variance. To check homogenity of variance, we use Levene Test.

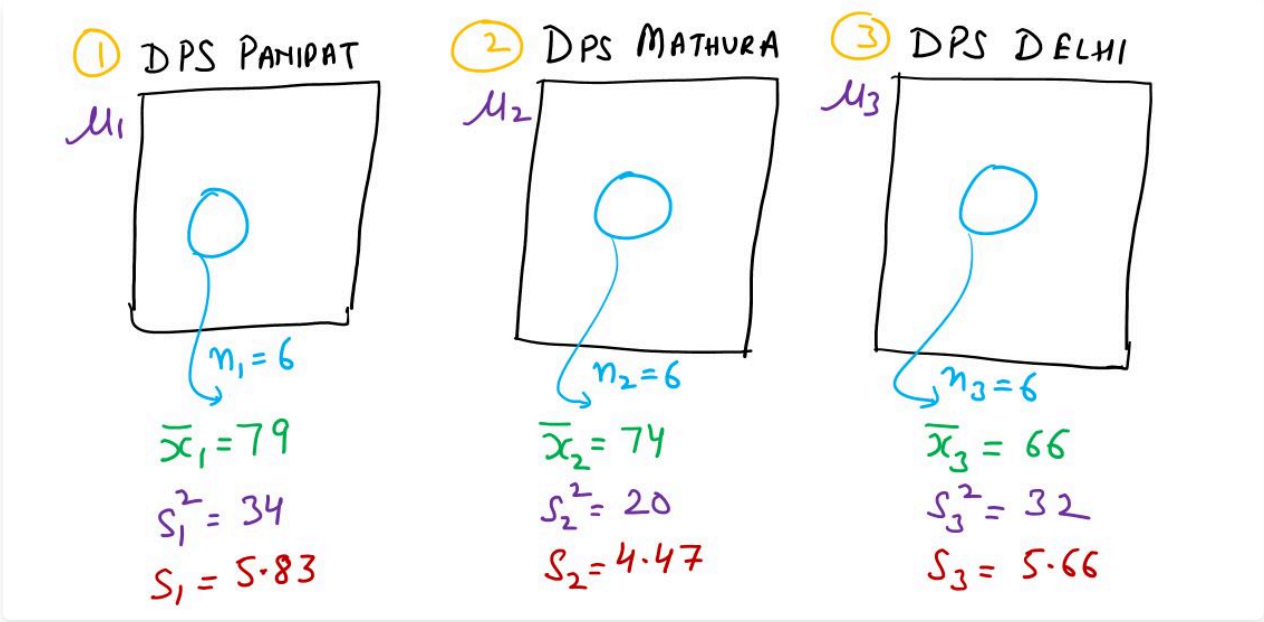
$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 \cdots = \sigma^2$$

3. The observations made must be independent of each other. Intelligence levels score of each student is independent of the Intelligence levels score of any other student. This can be achieved by random sampling.

4. Steps of ANOVA

Let us understand the steps of ANOVA with an example involving DPS schools.

We aim to assess the mean intelligence scores of Class X students across three DPS schools: DPS Panipat, DPS Mathura, and DPS Delhi. To conduct this analysis, samples from each school are gathered, and their sample mean, sample variance, and sample standard deviation are measured. The values obtained are given in the figure.



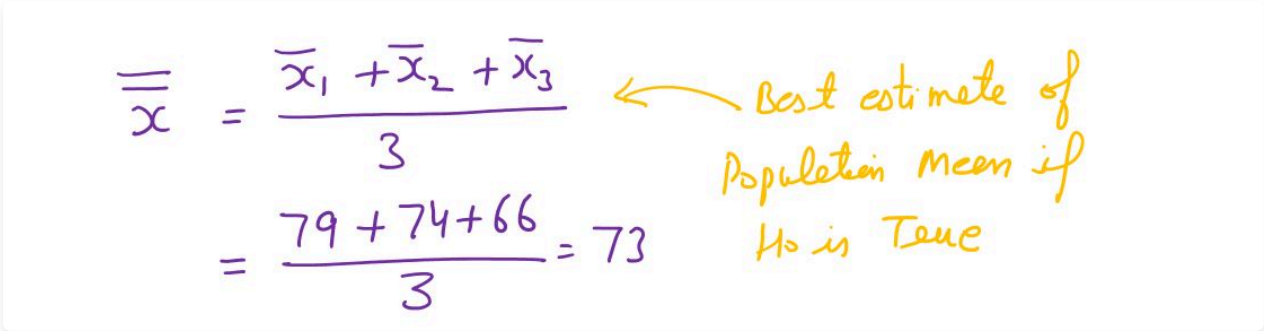
Step 1: Formulate Null and Alternate Hypothesis

The Null and Alternate Hypothesis are given below.

If the means of the three populations are identical, we anticipate observing reduced variability in the sample means, thereby corroborating our Null Hypothesis.

Step 2: Calculate mean of sample means

Following this, we determine the grand mean ($\bar{\bar{x}}$) by calculating the average of the sample means.



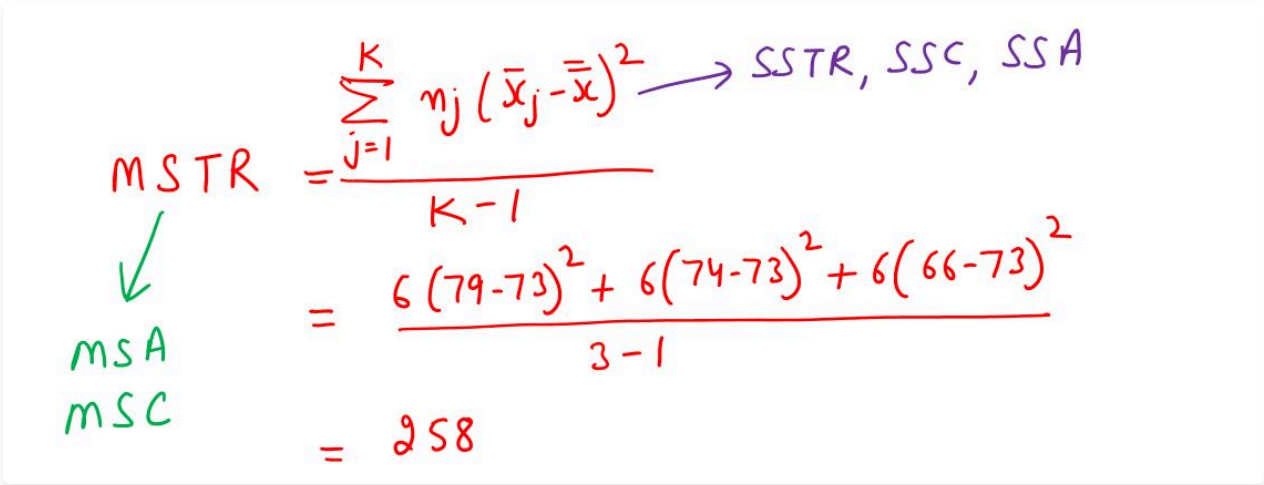
Step 3: Calculate Mean Square due to Treatments (MSTR)

Mean Square due to Treatments (MSTR) represents the variation among the means of different groups or treatments. MSTR is obtained by dividing the Sum of Squares due to Treatments (SSTR) by its degrees of freedom (DoF). It measures the variability between the means of various treatments or groups being compared.

It might have different names in different books, like MSA (Mean Square Among groups), MSC (Mean Square for Columns) or Between Treatment Estimate of σ^2.
MSTR = $\frac{SSTR}{k-1}$

Where, the numerator is called the Sum of Squares due to Treatments (SSTR) and is calculated as:
SSTR =$\sum_{j=1}^k n_j \left(\bar{x}_j - \bar {\bar x}\right)^2$
Just like MSTR, the SSTR also has different names in different books, like SSA (Sum of Squares Among groups), SSC (Sum of Squares for Columns).

The denominator, k-1, represents the degrees of freedom associated with SSTR.



If H_0 is true, then MSTR will provide unbiased estimate of σ^2. On the other hand if H_0 is not true, then MSTR will overestimate value of σ^2.

Step 4: Calculate Mean Square due to Error (MSE)

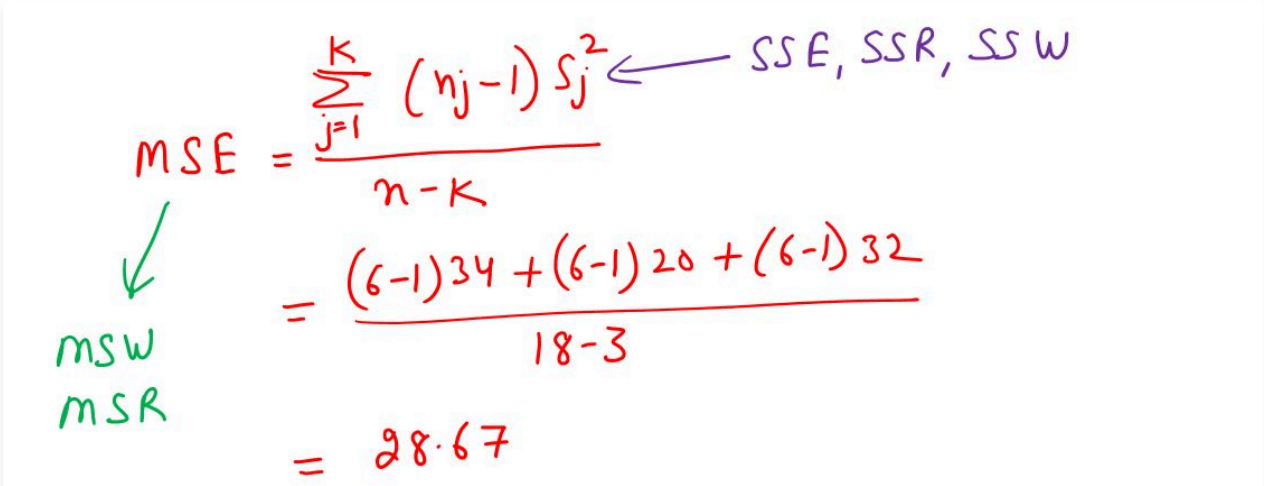
Mean Square due to Error (MSE) represents the variation within each treatment group, capturing the variability that is not explained by the differences between the treatment means.

It might have different names in different books, like MSW (Mean Square Within groups), MSR (Mean Square for Rows) or Within Treatment Estimate of σ^2.
MSE = $\frac{SSE}{n-k}$

Where, the numerator is called the Sum of Squares due to Error (SSE) and is calculated as:
SSE =$\sum_{j=1}^k (n_j-1) s_j^2$

Just like MSE, the SSE also has different names in different books, like SSW (Sum of Squares Within groups), SSR (Sum of Squares for Rows).

The denominator of MSE (n-k) is referred to as the degrees of freedom associated with SSE.



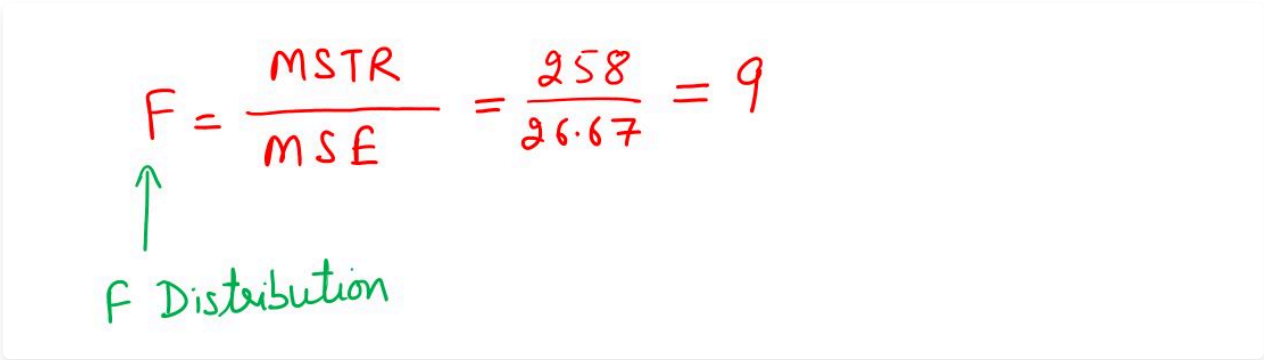
It may be noted that MSE will always give correct estimate of σ^2, irrespective of the fact that H_0 is True or not True.

Step 5: Calculate F value

For normal populations, the sampling distribution of the ratio of two independent estimates of σ^2 follows an F distribution. Hence, if the null hypothesis is true and the ANOVA assumptions are valid, the sampling distribution of MSTR/MSE is an F distribution with numerator degrees of freedom equal to k-1 and denominator degrees of freedom equal to n-k.

In other words, if the null hypothesis is true, the value of MSTR/MSE should appear to have been selected from this F distribution.

Thus, the F Statistic is calculated as
F = $\frac{MSTR}{MSE}$



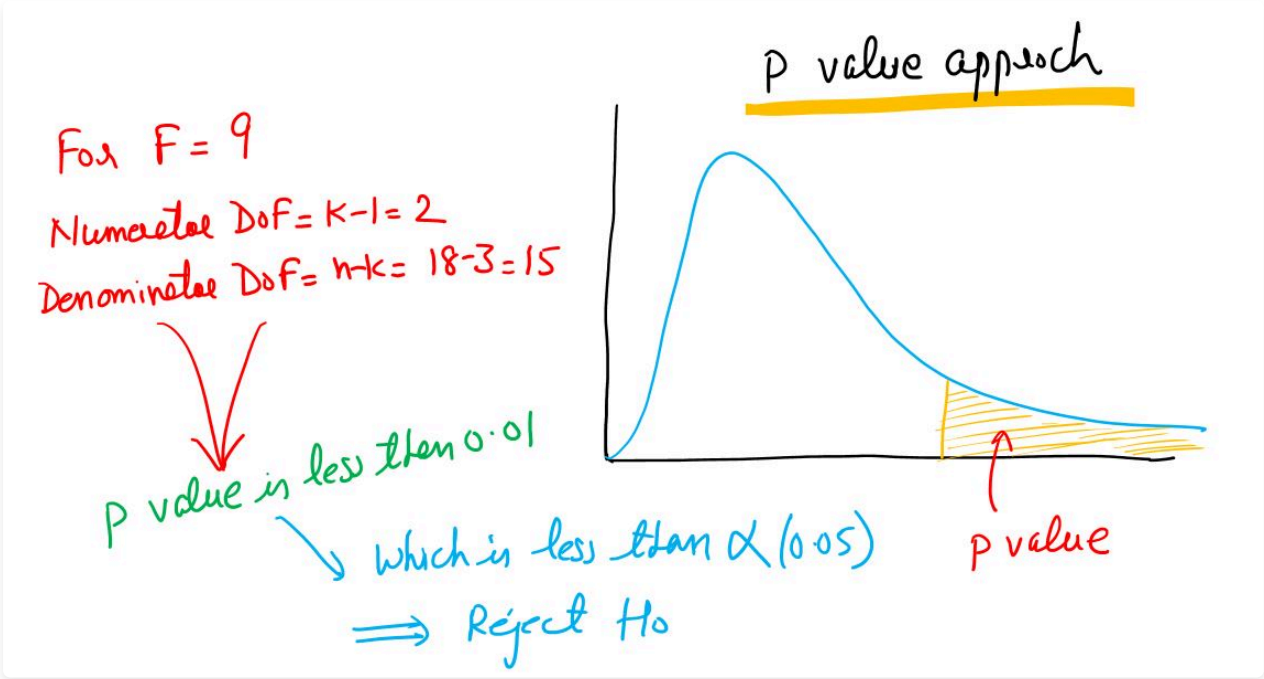
Step 6: Rejection Rule

With numerator degrees of freedom as k–1 and denominator degrees of freedom as n-k, along with a specified level of significance α, we calculate corresponding p value from F Table.

We Reject H_0 if, p-value ≤ α.

In Our example, the numerator degrees of freedom is k – 1 = 3 – 1 = 2 and the denominator degrees of freedom is n-k = 15 – 3 = 12. We use a level of significance α = 0.05.

The corresponding p value is 0.004 (from the table), which is less than α = 0.05.



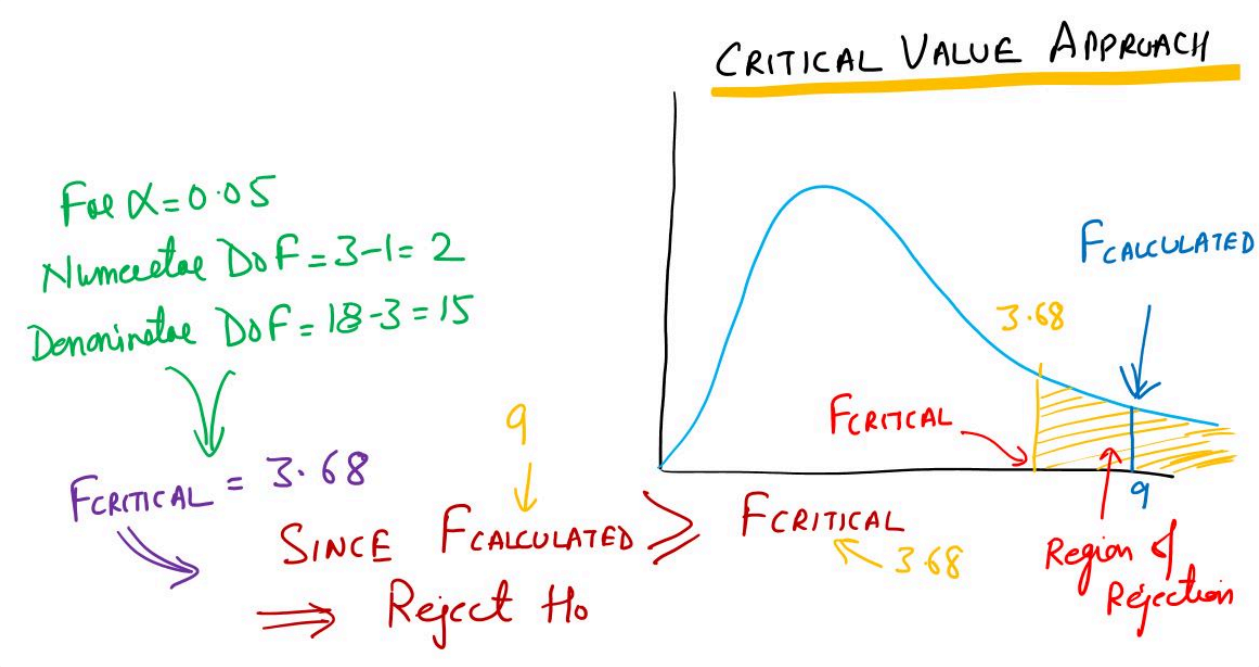
Hence, we reject Null Hypothesis and conclude that the means of the three populations are not equal (we conclude mean intelligence scores of students of 3 DPS schools are not same).

Please note that, if H_0 is rejected, we cannot conclude that all population means are different. Rejecting H_0 only means that at least two population means have different values.

Critical Value Approach

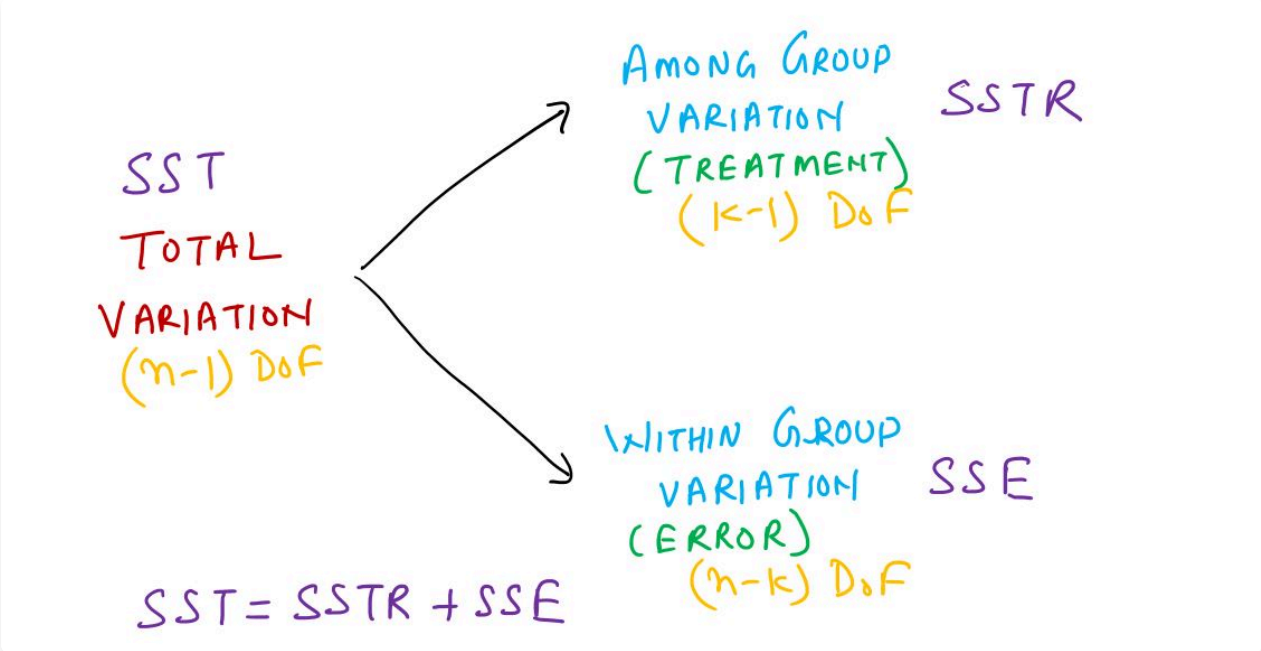
Critical value approach: Reject H_0 ,if $F \geq F_\alpha$

Where the value of F_α is based on an F distribution with k – 1 numerator degrees of freedom and \(n_T-k \) denominator degrees of freedom.



5. Error Partioning

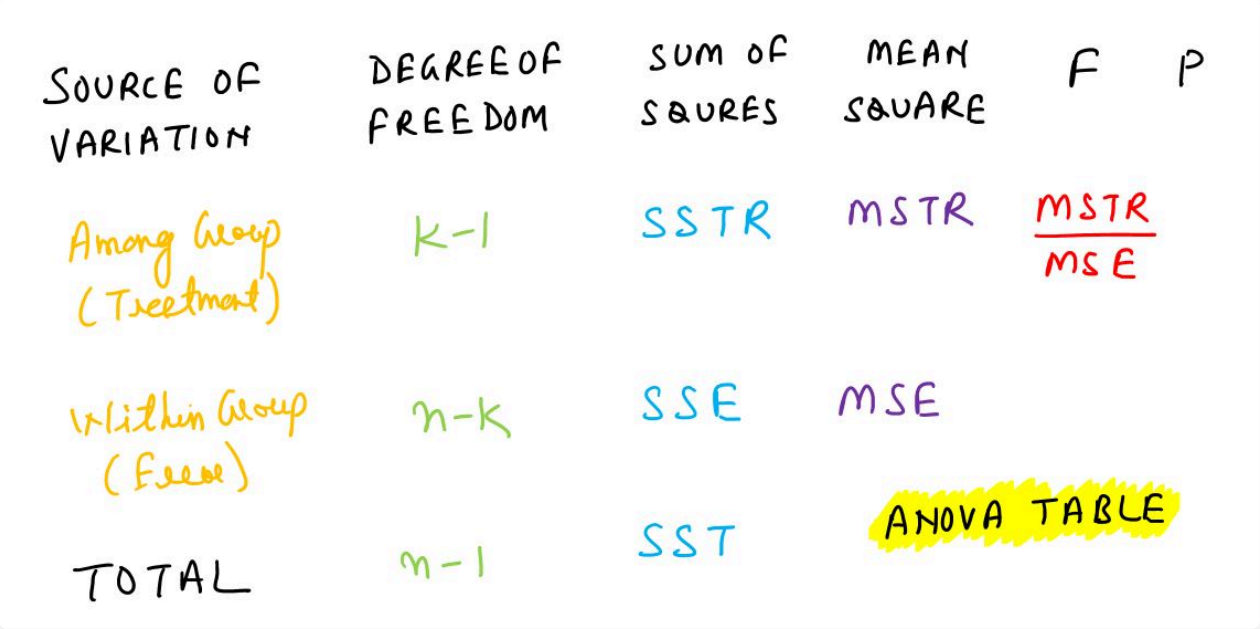
ANOVA, or Analysis of Variance, hinges on a fundamental principle. The total variability observed in a dataset can be divided into distinct components. It's segmented into variations among groups, showcasing differences from one group to another, and variations within groups, reflecting random discrepancies.



This division involves breaking down the Sum of Squares Total and its corresponding degrees of freedom into separate sources. The entire variation (with a degree of freedom of n-1) is broken into two parts:
(i) Treatment, which accounts for differences among groups (with a degree of freedom of k-1), and
(ii) Error, signifying variation within groups (with a degree of freedom of n-k).

---

6. ANOVA Table



The ANOVA table is a comprehensive summary of the results derived from the Analysis of Variance. It presents the sources of variation, the degrees of freedom associated with each source, the sum of squares for each source, the mean squares, and the F-statistic along with its associated p-value.

The table typically includes the following components:

(i) **Source of Variation**: This section lists the various sources of variability in the data. For instance, it may include the treatment or factor being studied and the residual or error term.

(ii) **Degrees of Freedom (DoF)**: Indicates the degrees of freedom for each source of variation, showing the number of independent observations used in calculating the variance.

(iii) **Sum of Squares (SSTR or SSE)**: Represents the sum of squared deviations from the mean, reflecting the variability within each source.

(iv) **Mean Squares (MSTR or MSE)**: Obtained by dividing the sum of squares by the degrees of freedom, providing an average variance estimate for each source.

(v) **F-statistic**: The ratio of mean squares for the treatment and error, used to test the hypothesis that there are no significant differences among group means.

(vi) **p-value**: Indicates the probability of obtaining the observed F-statistic (or a more extreme value) under the assumption that the null hypothesis is true. A small p-value suggests evidence against the null hypothesis.

The ANOVA Table for our example on DPS schools is given below.

SOURCE OF VARIATION	DEGREE OF FREEDOM	SUM OF SQURES	MEAN SQUARE	F	P
Among Group (Treatment)	2	516	258	9	less then 0.01
Within Group (Error)	15	430	28.67		
TOTAL	17	946			

ANOVA TABLE

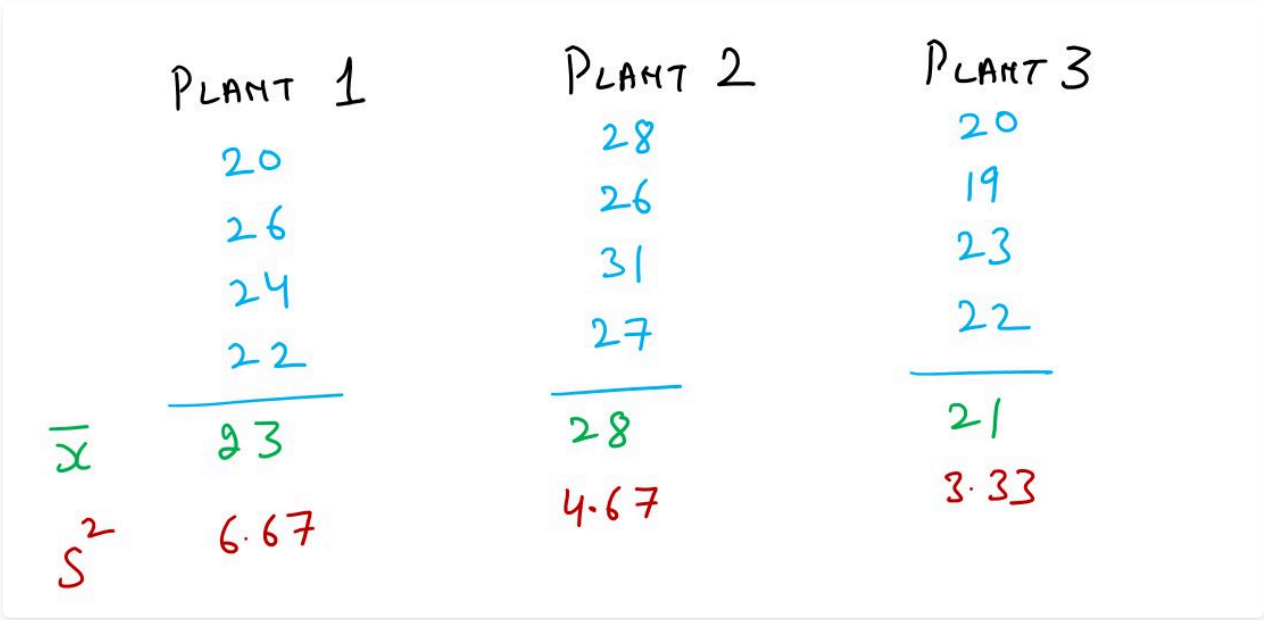
7. Another Illustration ANOVA

To test whether the mean time needed to stitch a shirt is the same for machines produced by three plants, the H&M Company obtained the following data on the time (in minutes) needed to stitch a shirt. Use these data to test whether the population mean times for stitching a shirt differ for the three plants. Use $\alpha = 0.05$. The data is given below.

Plant 1	Plant 2	Plant 3
20	28	20
26	26	19
24	31	23
22	27	22

Solution:

First we calculate mean and variance for data from each plant.



Step 1: Formulate Null and Alternate Hypothesis

The Null and Alternate Hypothesis are given below.

If the means of the three populations are identical, we anticipate observing reduced variability in the sample means, thereby corroborating our Null Hypothesis.

Step 2: Calculate mean of sample means

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1+n_2+n_3} = \frac{4\times23 + 4\times28 + 4\times21}{4+4+4}$$

$$= 24$$

Step 3: Calculate Mean Square due to Treatments (MSTR)

$$MSTR = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} = \frac{4(23-24)^2 + 4(28-24)^2 + 4(21-24)^2}{3-1}$$

$$= \frac{104}{2} = 52$$

Step 4: Calculate Mean Square due to Error (MSE)

$$MSE = \frac{\sum_{j=1}^{k}(n_j-1)s_j^2}{n-k} = \frac{3(6.67) + 3(4.67) + 3(3.33)}{12-3}$$

$$= \frac{44.01}{9} = 4.89$$

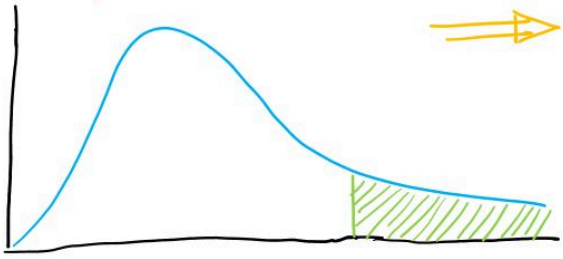
Step 5: Calculate F value

$$F = \frac{MSTR}{MSE} = \frac{52}{4.89} = 10.63$$

Step 6: Rejection Rule

Corresponding p is < 0.01, which is less $\leq \alpha$ ≤ 0.05

$\Rightarrow$ REJECT H_0



ANOVA Table

SOURCE OF VARIATION	DEGREE OF FREEDOM	SUM OF SQURES	MEAN SQUARE	F	P
Among Group (Treatment)	2	104	52	10.63	Less then 0.01
Within Group (Error)	9	44.01	4.89		
TOTAL	11	148.01			

Rejecting the null hypothesis indicates that there is a statistically significant difference among the mean stitching times of the three plants.

It does not specifically identify which particular means are different, only that there is evidence to support the idea that there is variability among them.

---

7. Another Illustration ANOVA

In previous question, when we rejected the null hypothesis, it meant that there's a statistically significant difference among the average stitching times of the three plants. However, it doesn't pinpoint which exact averages are distinct. We use Tukey's test to pinpont, which groups are different.

Tukey's test, or Tukey's honestly significant difference test, is a post hoc test commonly used after an analysis of variance (ANOVA) to identify which specific groups differ significantly from each other when there are three or more groups being compared. It determines the differences between group means.

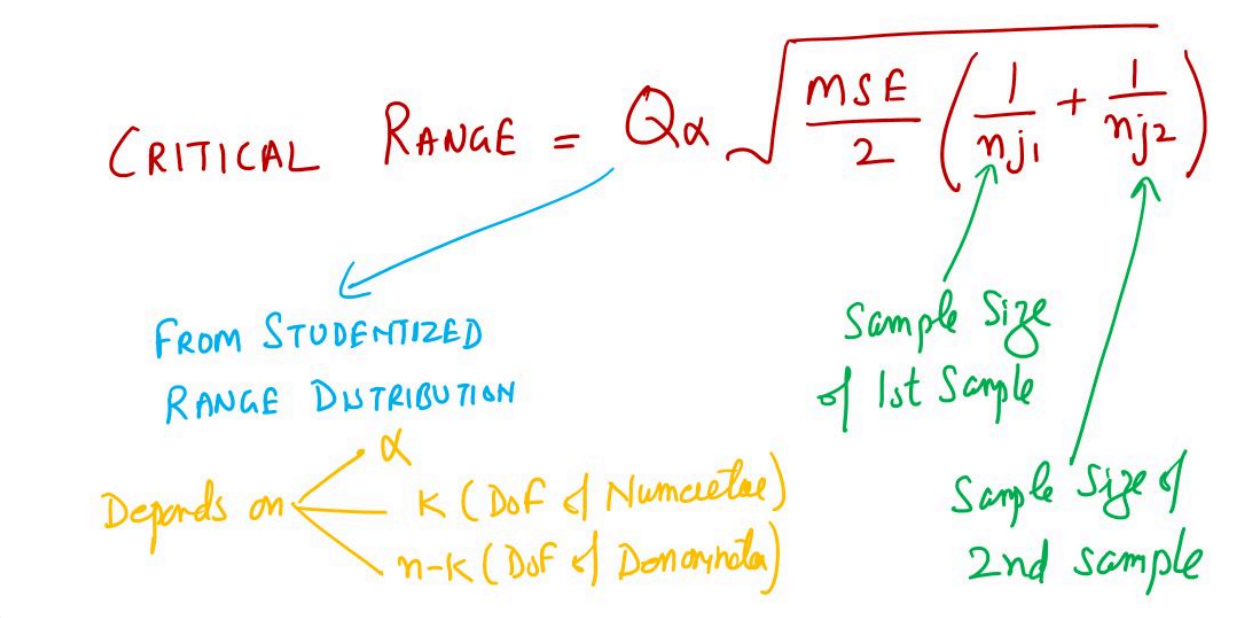
The Procedure Follows 3 steps.

Step 1: Determine the significance level

Set the desired significance level (usually $\alpha = 0.05$) to control the probability of incorrectly rejecting a true null hypothesis (Type I error rate).

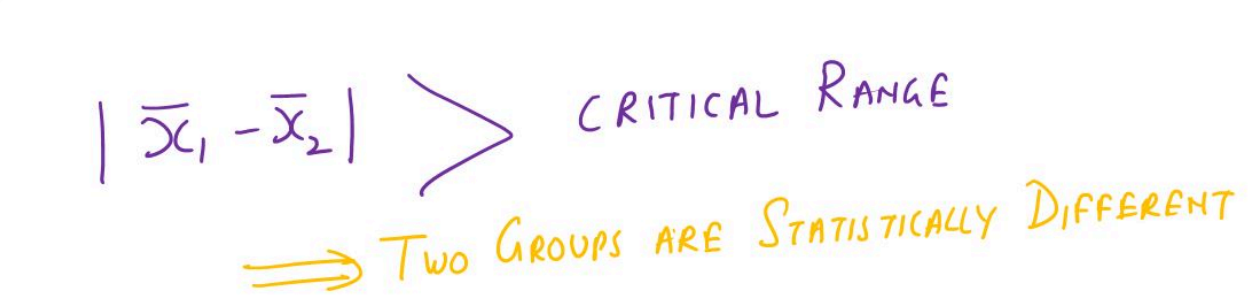
Step 2: Calculate Critical Range or Tukey's HSD

Compute Tukey's Honestly Significant Difference (HSD) statistic using the formula:



Step 3: Compare group means

Calculate the differences between the means of all possible pairs of groups.



If the difference between two group means exceeds the computed HSD value (Critical Range), it indicates a statistically significant difference between those groups.

Comparing Plant 1 and 2

$|\bar{x}_1 - \bar{x}_2| = |23-28| = 5$

$\alpha = 0.05$, $k = 3$, $n-k = 12-3 = 9$ (Using Table)

$\text{CRITICAL RANGE} = 3.95 \times \sqrt{\frac{4.89}{2}\left(\frac{1}{4}+\frac{1}{4}\right)} = 4.37$

SINCE $|\bar{x}_1 - \bar{x}_2| >$ CRITICAL RANGE

$\Rightarrow$ PLANT 1 and PLANT 2 are steliscelly Synificant Diffeent

Comparing Plant 2 and 3

$|\bar{x}_2 - \bar{x}_3| = |28-21| = 7$

$\alpha = 0.05$, $k = 3$, $n-k = 12-3 = 9$ (Using Table)

$\text{CRITICAL RANGE} = 3.95 \times \sqrt{\frac{4.89}{2}\left(\frac{1}{4}+\frac{1}{4}\right)} = 4.37$

SINCE $|\bar{x}_2 - \bar{x}_3| >$ CRITICAL RANGE

$\Rightarrow$ PLANT 2 and PLANT 3 are significantly different.

Comparing Plant 1 and 3

$|\bar{x}_1 - \bar{x}_3| = |23-21| = 2$

$\alpha = 0.05$, $k = 3$, $n-k = 12-3 = 9$ (Using Table)

$\text{CRITICAL RANGE} = 3.95 \times \sqrt{\frac{4.89}{2}\left(\frac{1}{4}+\frac{1}{4}\right)} = 4.37$

SINCE $|\bar{x}_1 - \bar{x}_3| <$ CRITICAL RANGE

$\Rightarrow$ PLANT1 and PLANT 3 are not significatly diffeent

7. Another Illustration ANOVA

When we use ANOVA to test whether the means of k populations are equal, rejection of the null hypothesis allows us to conclude only that the population means are not all equal.

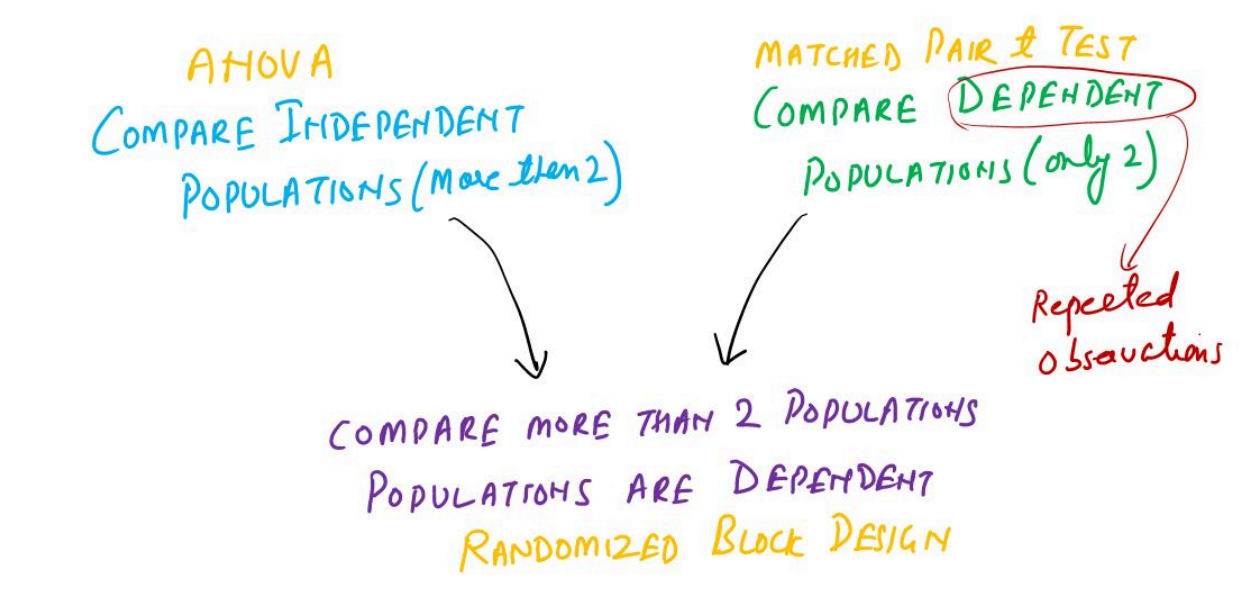
In some cases, we want to go a step further and determine where the differences among means occur. In this case, **Fisher's least significant difference (LSD)** procedure can be used to determine where the differences occur.

---

8. Randomized Block Design

Previously, we have understood **ANOVA** to compare means across more than 2 independent populations.

We have also learnt, how we use **matched pair t Test** for comparing two population means with repeated observations on the same subjects (Dependent population).



What if, we need to compare means across more than 2 populations and these observations are repeated within the same population (dependent populations).

In this case, we utilize the **randomized block design**.

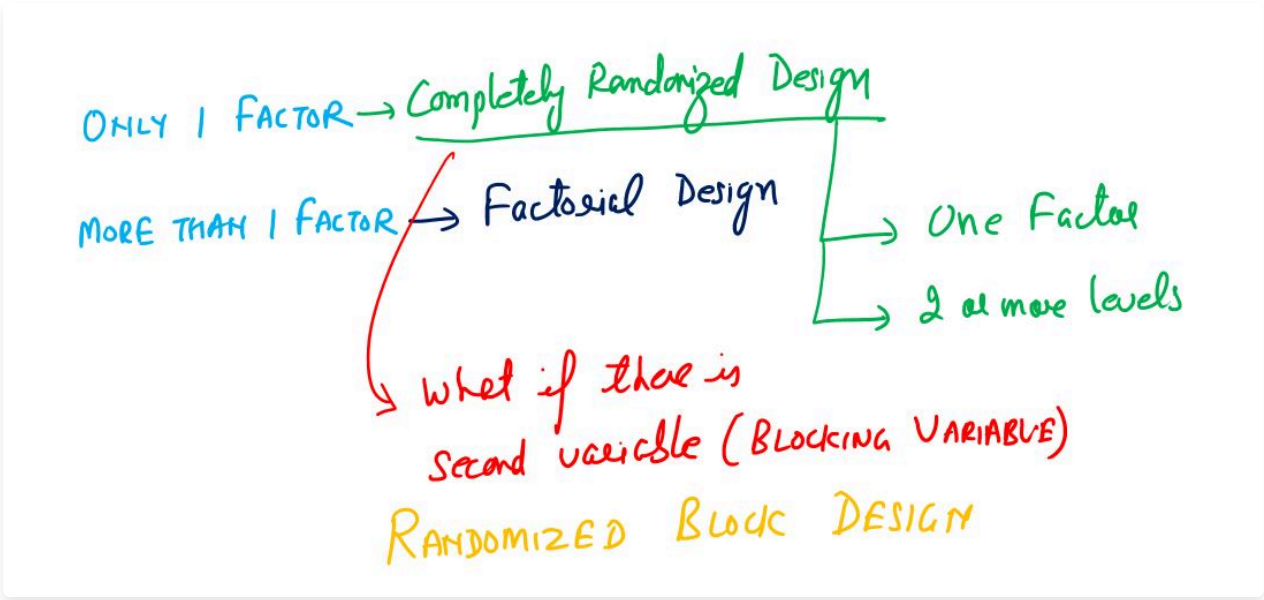
Blocks refer to the individuals or units on whom these repeated observations are taken.

For instance, in a medical study comparing the effectiveness of various treatments for different patients, where each patient receives multiple treatments, the patients themselves act as blocks. The treatments are applied within each patient, and the responses to these treatments are measured. By considering patients as blocks, the analysis can account for the variability among individuals, allowing for a more accurate comparison of the treatment effects.

---

8. Randomized Block Design

In experimental designs, when one factor is considered, it's referred to as a **Completely Randomized Design**. Conversely, when more than one factor is involved, it's known as a **Factorial Design**.



In a Completely Randomized Design, subjects are randomly allocated to different treatments. This design typically involves a single independent variable (or factor) with two or more levels.

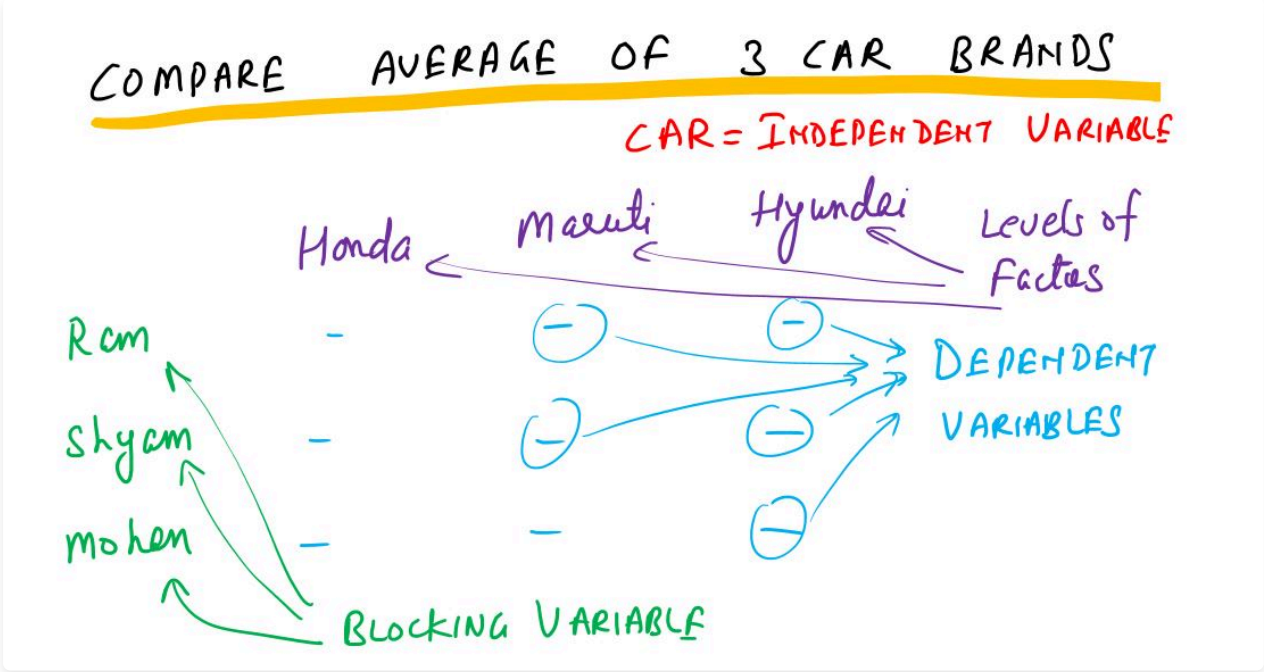
However, there exists an alternative version of the Completely Randomized Design that includes a second variable called a **blocking variable**.

A blocking variable is not the primary treatment variable but is instead used by the researcher to control or manage other variables that may affect the experiment's outcomes. These variables, often known as confounding or concomitant variables, can have an impact on the observed results.

For instance, let's consider a scenario where we aim to compare the average mileage of three cars—Honda, Maruti, and Hyundai.

In this case, the average mileage serves as the dependent variable. The type of car (Honda, Maruti, Hyundai) represents the independent variable or factor, consisting of three levels.

Additionally, the three drivers—Ram, Shyam, and Mohan—act as the blocking variables. While these drivers are not the main focus of the study, their driving styles may significantly impact the average mileage of each car. Hence, by considering these drivers as blocking variables, we control for any potential influence their driving styles might have on the results.



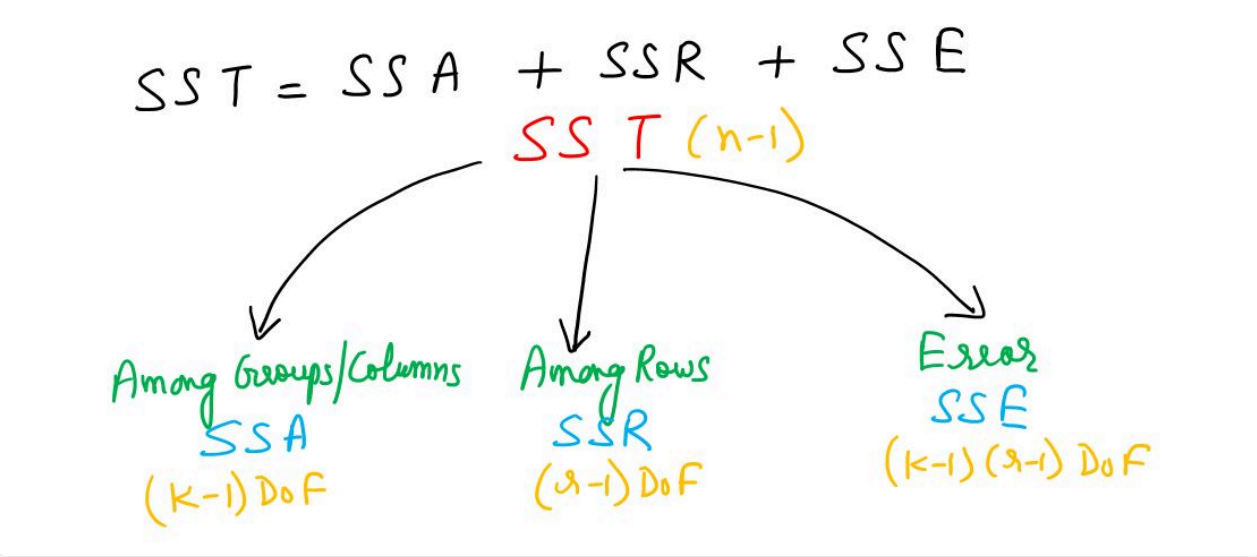
This approach is crucial in ensuring that any observed differences in mileage among the cars are not solely influenced by individual driving habits, allowing for a more accurate assessment of the impact of the car's make or model on fuel efficiency.

Ronald A. Fisher introduced the concept of blocking variables while investigating seed variety in plants. He sought to compare the characteristics of different seed types, but he encountered an additional factor that influenced plant growth: soil moisture. In response, Fisher devised a method to mitigate this variable's impact by dividing the land into distinct sections or blocks. These blocks were used to control the influence of soil moisture, ensuring that it didn't disproportionately affect the comparison of seed varieties. Consequently, these partitioned sections were referred to as blocking variables, a term coined by Fisher due to their function in blocking or neutralizing the effect of certain confounding factors within the experiment.

---

8. Randomized Block Design

In ANOVA, blocking, or the inclusion of blocking variables, is a technique employed to control for extraneous factors that might influence the outcome of an experiment. By introducing blocking variables, researchers aim to reduce variability caused by these factors and focus more precisely on the effects of the variables they are specifically investigating.



When blocking is introduced into ANOVA, it refines the analysis by segregating the sources of variation more comprehensively.

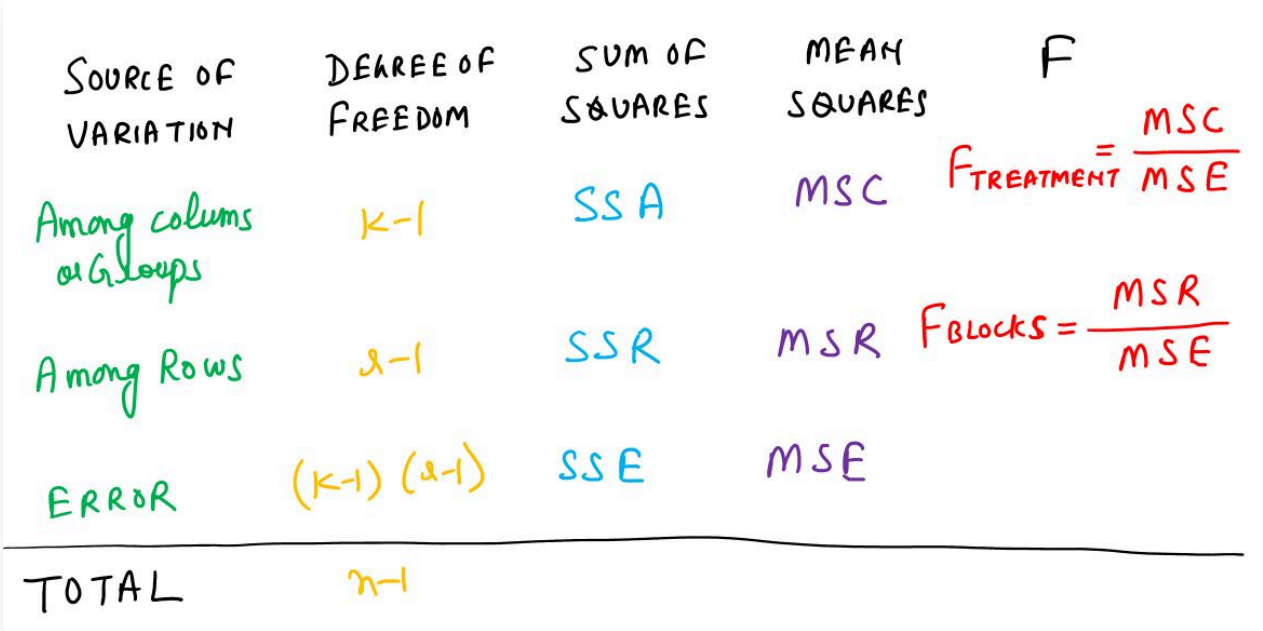
Conventionally, ANOVA partitions the total variation (SST) into components representing variation among treatment groups (columns) and the error term. However, with the inclusion of blocking, this **partitioning expands to include a third factor**: variation among blocks (or rows) in addition to variation among treatment groups and the residual error term.

This partitioning helps researchers differentiate and understand the contribution of different factors to the observed variability. It separates out the variability attributable to the various levels or groups under study (columns), the variability stemming from the blocks or subsets used to control extraneous factors (rows), and the residual variability that remains unaccounted for by the specified factors or blocks.

---

8. Randomized Block Design

The ANOVA table in a randomized block design would include the sources of variation: among columns (treatments), among rows (blocks), and the residual error.



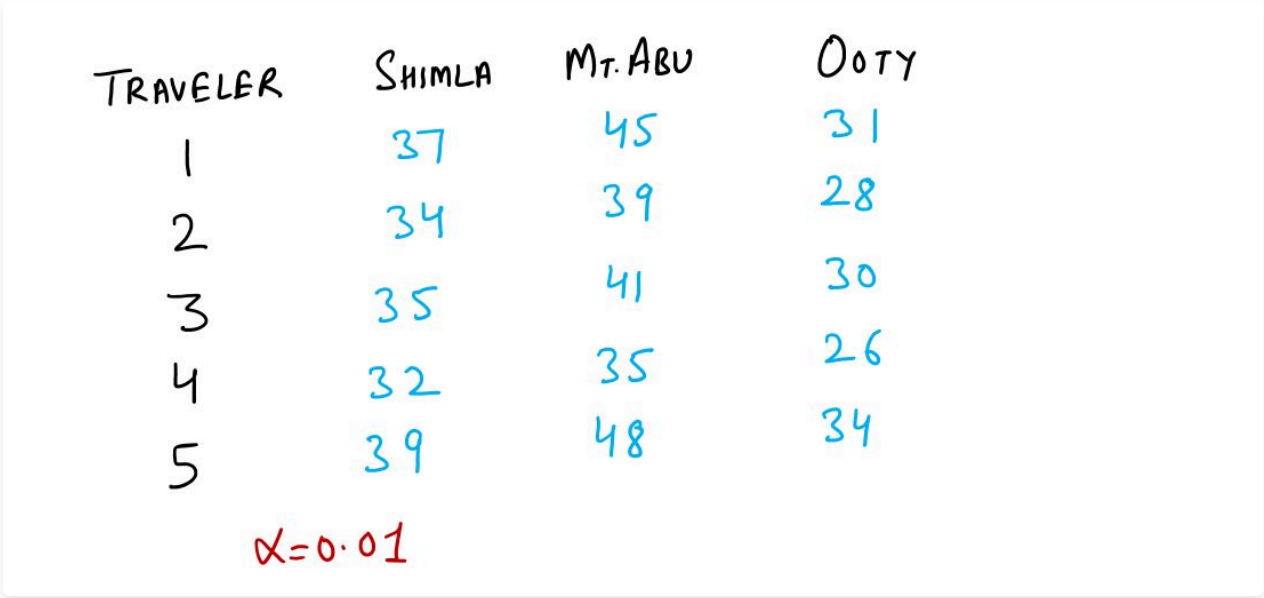
This ANOVA table showcases the partitioning of the total variability (SST) into three distinct sources:
(i) among columns (representing treatment effects), SSA
(ii) among rows (reflecting block effects), SSR and
(iii) residual variability that remains unexplained, SSE.

The table includes the respective sum of squares, degrees of freedom, mean squares, and F-ratios associated with each source of variation.

---

9. Illustration on Randomised Block Design

A travel agency wanted to compare holiday experiences across three destinations: Shimla, Mount Abu, and Ooty. They sent five travelers to each location to rate their experience on a scale of 1 to 100. Using a significance level of 0.01, conduct an analysis to determine if there are any significant differences in the holiday experiences across these destinations. The feedback is tabulated below.



Solution:

1. First let us calculate treatment means, block means and grand mean.

2. Calculate SST, SSA (SSC) and SSR.

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij} - \bar{\bar{x}})^2$$

$$= (37-35.6)^2 + (34-35.6)^2 + \cdots (34-35.6)^2 = 517.6$$

$$SSC/SSA = n\cdot\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2$$

$$= 5\left[(35.4-35.6)^2 + (41.6-35.6)^2 + (29.8-35.6)^2\right] = 348.4$$

$$SSR = k\cdot\sum_{i=1}^{n}(\bar{x}_i - \bar{\bar{x}})^2$$

$$= 3\left[(37.7-35.6)^2 + (33.7-35.6)^2 + \cdots (40.3-35.6)^2\right] = 154.9$$

3. Since SST = SSA(SSC) + SSR + SSE, we can calculate value of SSE.

$$SST = SSA + SSR + SSE$$

$$517.6 \quad 348.4 \quad 154.9 \quad ?$$

$$\Rightarrow SSE = 517.6 - (348.4 + 154.9) = 14.3$$

4. Calculate MSC(MSA), MSR and MSE.

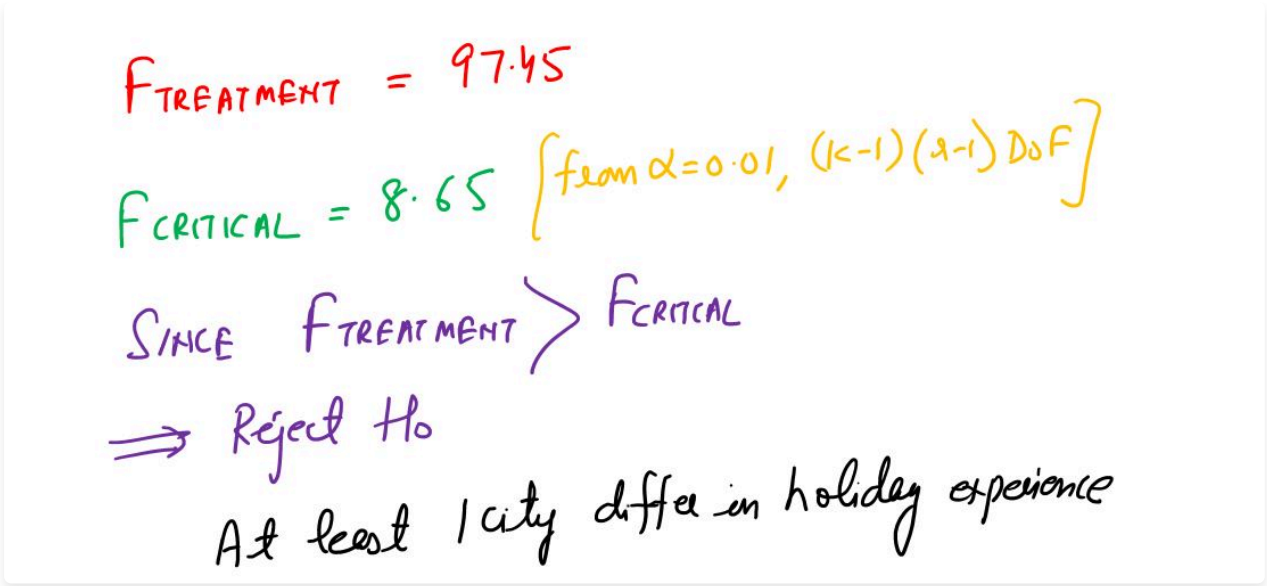
$$MSC/MSA = \frac{SSC/SSA}{k-1} = \frac{348.4}{3-1} = 174.2$$

$$MSR = \frac{SSR}{n-1} = \frac{154.9}{5-1} = 38.72$$

$$MSE = \frac{SSE}{(k-1)(n-1)} = \frac{14.3}{2\times 4} = 1.79$$

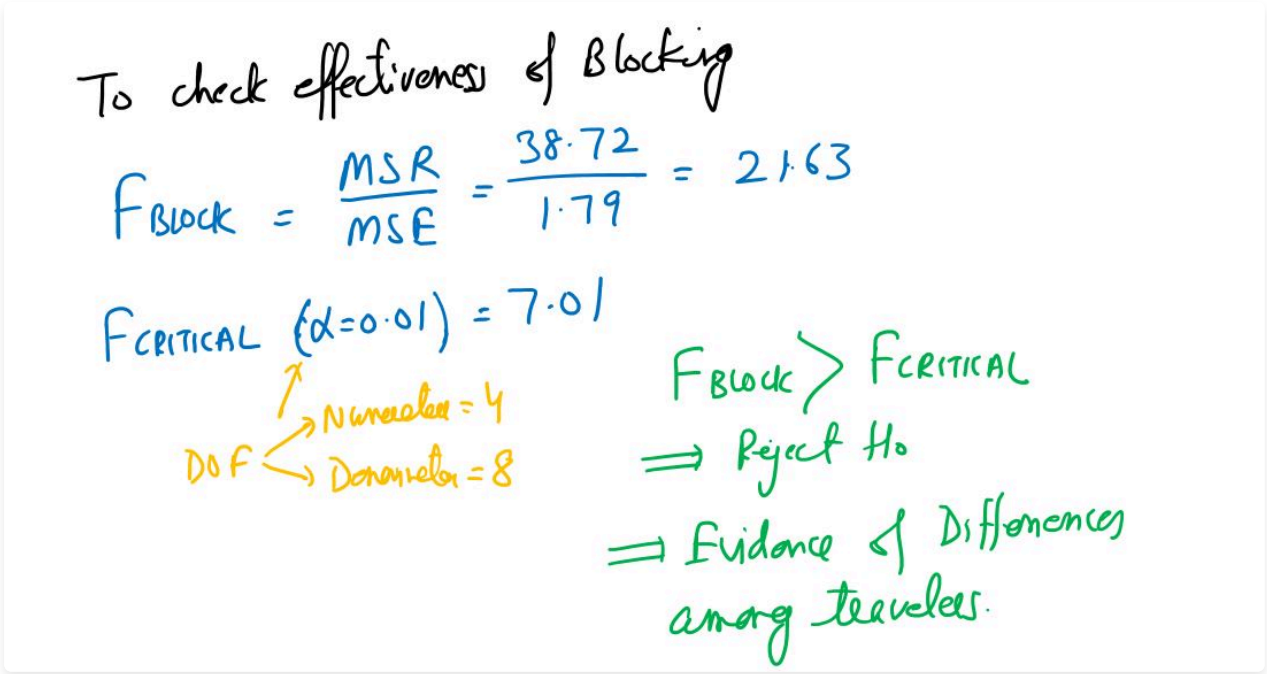
$$F_{TREATMENT} = \frac{MSC}{MSE} = \frac{174.2}{1.79} = 97.45$$

5. Reject Null Hypothesis if calculated value of F is greater than critical value of F.



Rejecting the null hypothesis for the treatment value of F would indicate that there are statistically significant differences in the holiday experiences among the destinations (Shimla, Mount Abu, and Ooty), based on the ratings given by the travelers.

6. Check effectiveness of Blocking by calculating Block Value of F and compare it with critical value of F.



Rejecting the null hypothesis for the block value of F would indicate that there are statistically significant differences among the travelers' experiences or ratings across the blocks (i.e., travelers' opinions might significantly differ from one another or from one specific group to another based on the destinations they visited).

It means that the Blocking has been advantageous to us.

But how much has blocking help? to find this, we will calculate Estimated Relative Frequency.

7. Calculate Estimated Relative Frequency.

ESTIMATED RELATIVE FREQUENCY

$$= \frac{(a-1)MSR + a(k-1)MSE}{(n-1)(MSE)} = \frac{4 \times 38.72 + 5 \times 2 \times 1.79}{14 \times 1.79}$$

$$= 6.89$$

It means that, it would take 6.89 times observations in completeley randomized design as compared to randomized block design to have same precision in comparing 3 holiday destinations.

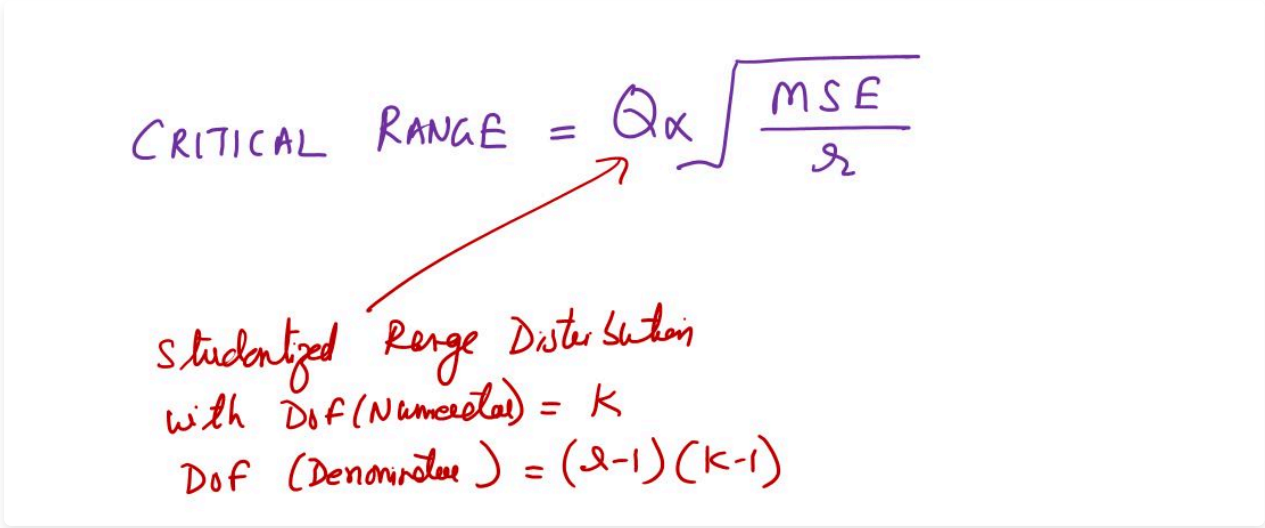
8. ANOVA Table is given below:

SOURCE OF VARIATION	DEGREE OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F
Among colums or Groups	2	348.4	174.2	$F_{TREATMENT} = \frac{174.2}{1.79} = 97.45$
Among Rows	4	154.9	38.72	$F_{BLOCK} = \frac{38.72}{1.79} = 21.63$
ERROR	8	14.3	1.79	
TOTAL	14			

10. Multiple Comparisons

As in the case of the completely randomized design, once you reject the null hypothesis of no differences between the groups, you need to determine which groups are significantly different from the others. For the randomized block design, you can use a procedure developed by Tukey.

The following equation gives the critical range for the Tukey multiple comparisons procedure for randomized block designs.



Let us understand the steps, using our example of Shimla, Ooty and Mt. Abu.

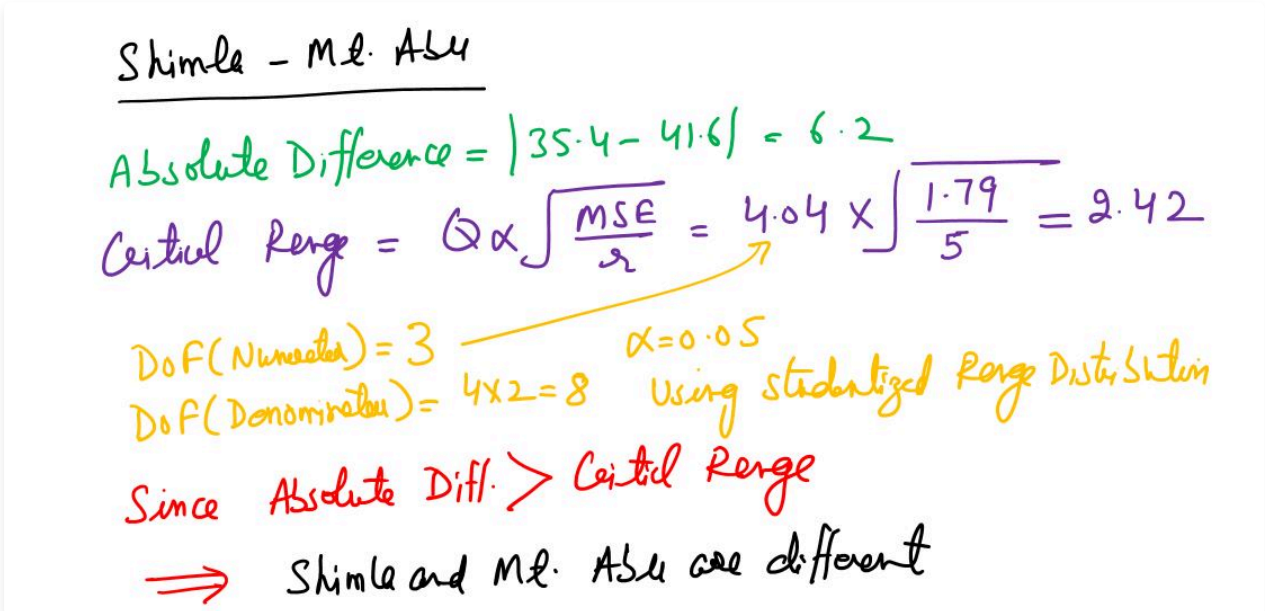
The Procedure Follows 3 steps.

Step 1: Determine the significance level
Set the desired significance level (usually α = 0.05) to control the probability of incorrectly rejecting a true null hypothesis (Type I error rate).

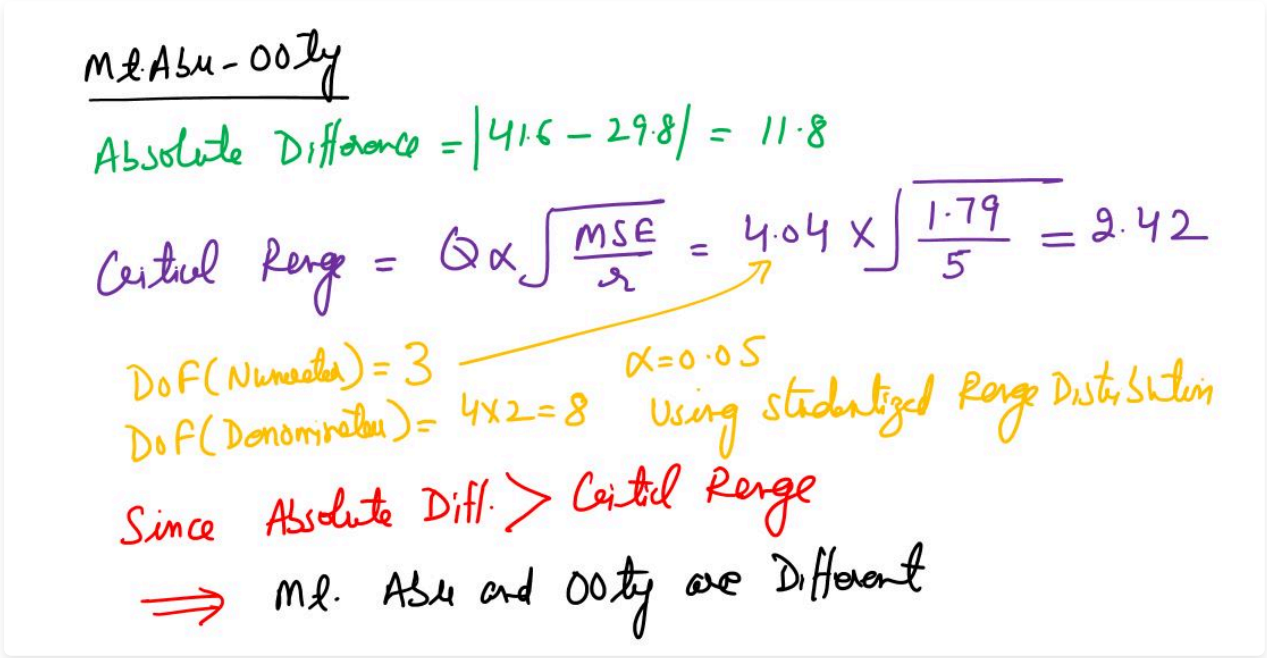
Step 2: Calculate Critical Range
Compute Critical Range using the formula.

Step 3: Compare and Infer
Calculate the absolute differences between the means of all possible pairs of groups. If the absolute difference between two group means exceeds the computed Critical Range, it indicates a statistically significant difference between those two groups.

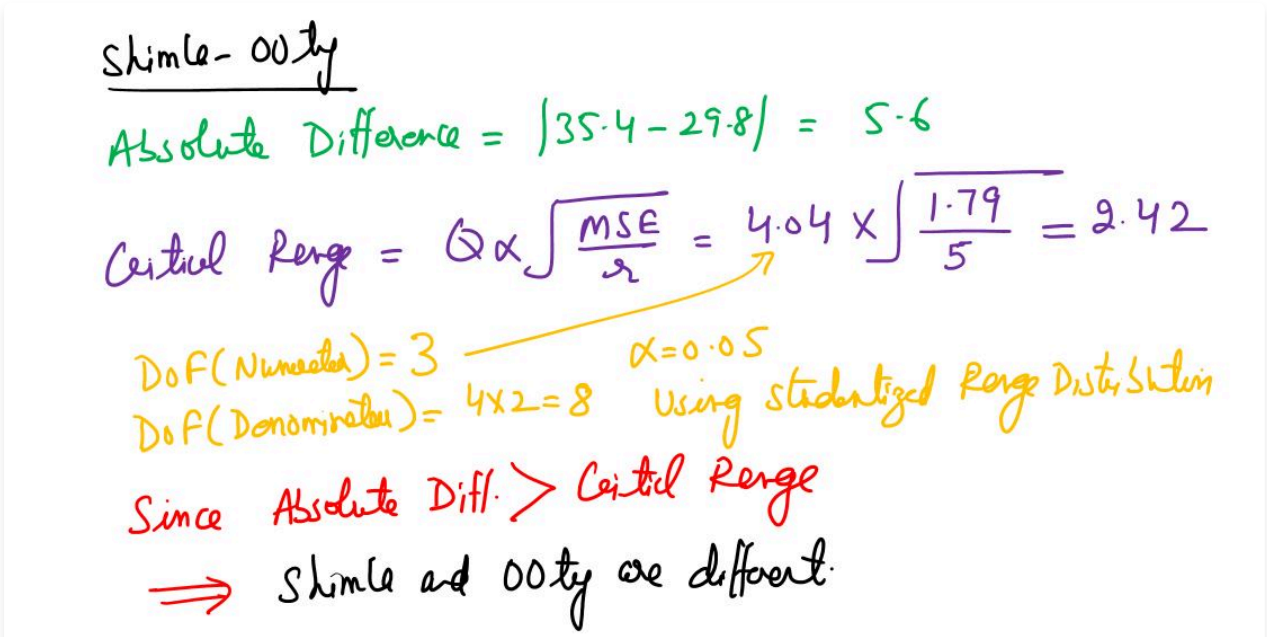
Comparing Shimla and Mt. Abu



Comparing Plant Mt. Abu and Ooty



Comparing Shimla and Ooty

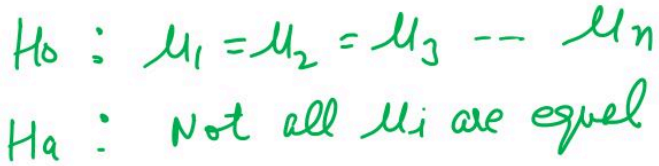


11. Two-way ANOVA

Two-way ANOVA, also known as factorial design, is used to analyze the influence of two independent categorical variables (factors) on a dependent variable. The primary objective is to determine whether there is a significant interaction between the two factors and whether each factor individually has a significant impact on the dependent variable.

In two-way ANOVA, there are two factors. Each factor has two or more levels.

The Interaction examines whether the combined effect of the factors on the dependent variable is significant or if they interact in influencing the outcome.



Consider a study investigating the effect of both diet and exercise on weight loss. The two factors are "Diet" (with levels A: Low-carb, B: High-carb) and "Exercise" (with levels X: Low intensity, Y: High intensity). The dependent variable is "Weight Loss."

Null Hypothesis (H_0): There is no significant difference in weight loss due to diet, exercise, or their interaction.
Alternative Hypothesis (H_1): There is a significant difference in weight loss, and it may be attributed to diet, exercise, or their interaction.

The main distinctions between One-Way and Two-Way ANOVA lie in the number of factors analyzed and whether interaction effects between these factors are considered. One-Way ANOVA focuses on a single factor, making it simpler and more straightforward. In contrast, Two-Way ANOVA involves two factors, introducing complexity as it investigates potential interactions between these factors, providing a more comprehensive understanding of their combined effects.

As an illustration, consider a One-Way ANOVA examining the impact of different training programs (factor) on employee performance. On the other hand, a Two-Way ANOVA might explore how both training programs (factor 1) and work experience (factor 2) collectively influence employee performance.

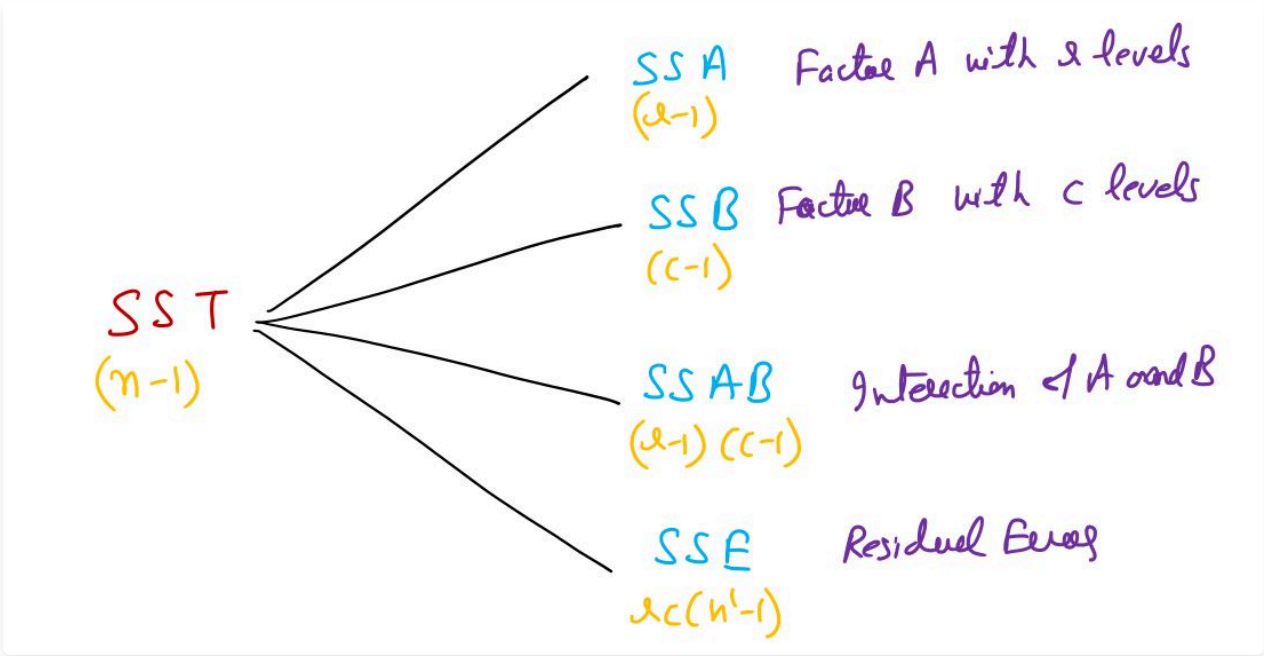
---

11. Two-way ANOVA

In Two-Way ANOVA analysis, consider two factors denoted as A and B.

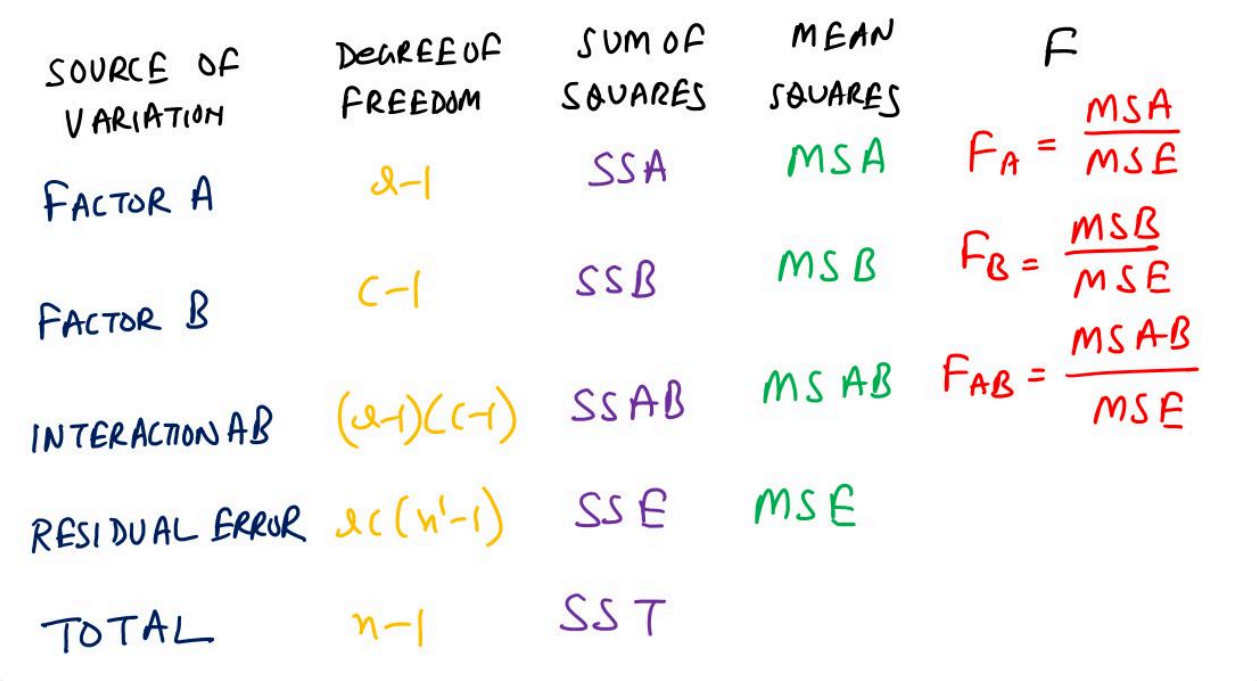
If factor A has r levels and factor B has c levels, and within each cell formed by a combination of a level from A and a level from B, there are n' values, then the total number of values, n, in the entire experiment is given by the product of the levels of A, B, and n', i.e., n = r \(\times \) c \(\times \) n'.

The partitioning of error in this analysis is distributed as follows:



11. Two-way ANOVA

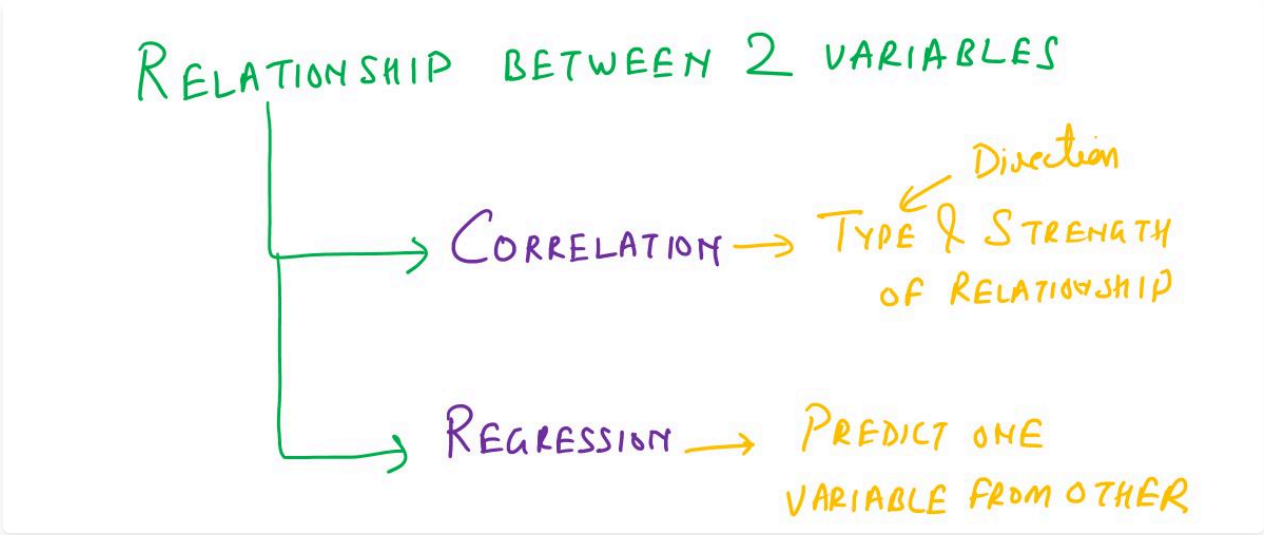
The ANOVA table for two-way ANOVA is given below:



1. Introduction

Our world is full of relationships.

Whether it is how ads impact sales, government actions affect economic growth, or even how your attitude can influence how productive you are, everything is related. This unit is all about understanding relationships, focusing on studying two things at the same time (bivariate data).



Think about it—the way a company plans its strategy affects how much of the market it gets, or how the temperature outside can make people buy more ice cream. Even something like where a college ranks might be connected to how much someone earns at their first job. And of course, hard work plays a role in success too. Relationsips are everywhere.

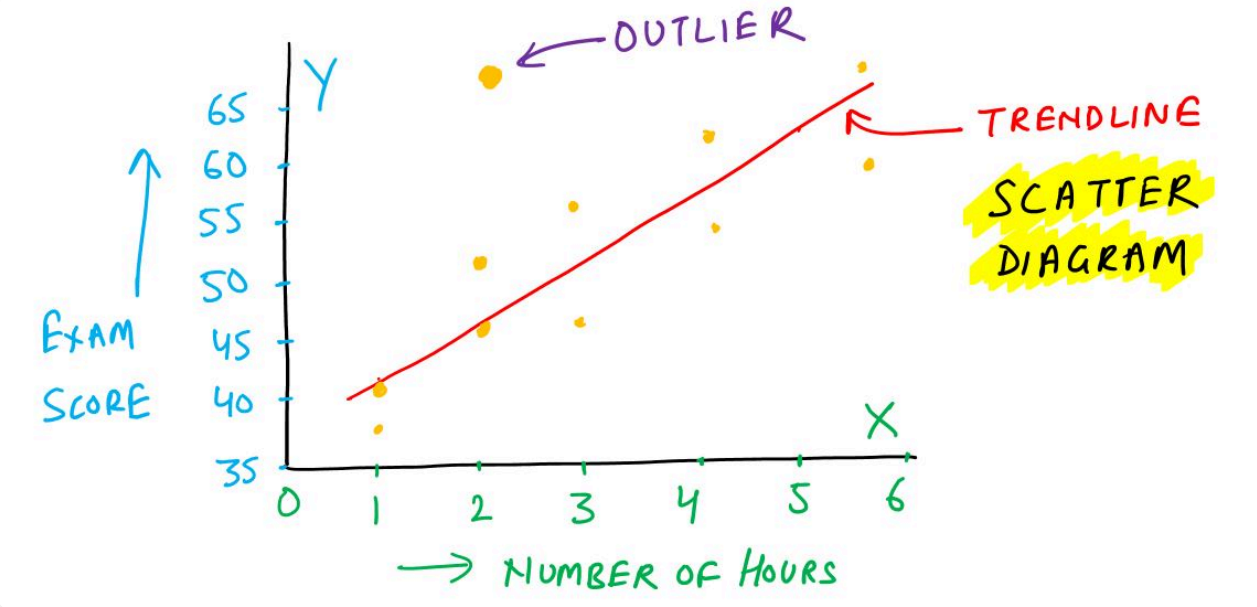
When we talk about **correlation**, we're trying to understand not just if things are related, but also type and strength of relationship.

Later We will explore **regression analysis**, a method that establishes relationships between two variables and enables the prediction of one from the other.

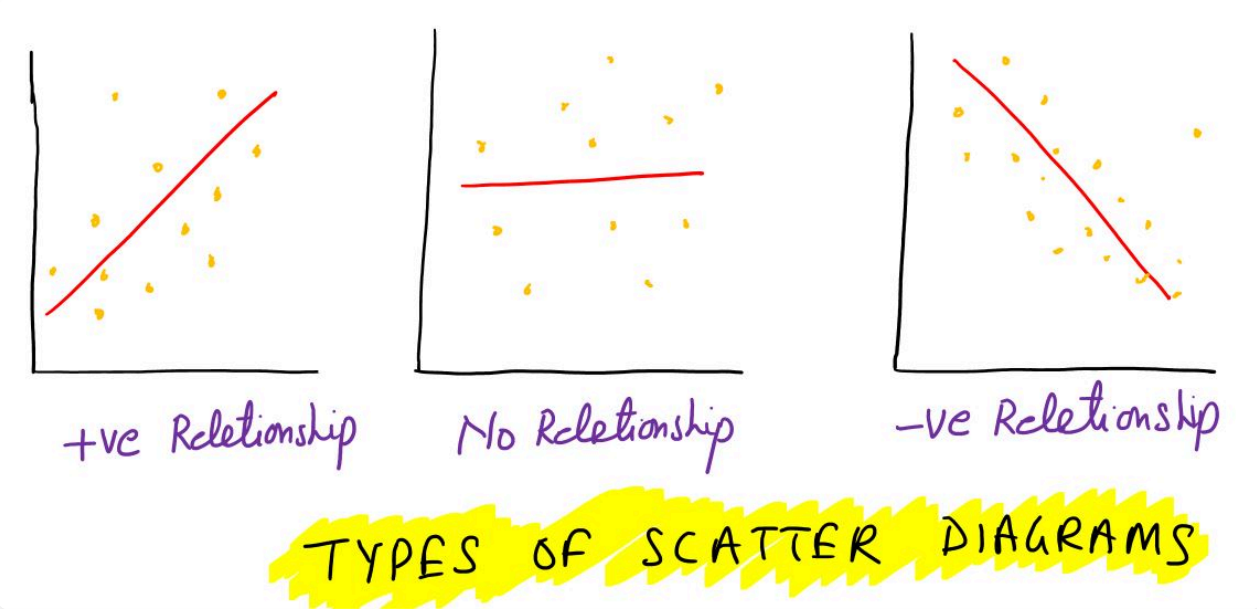
---

2. Scatter Diagram

A scatter diagram, also known as a scatter plot, is a visual representation of the relationship between two variables. It consists of points plotted on a two-dimensional graph, where each point represents the values of the two variables. The horizontal axis (x-axis) typically represents one variable, while the vertical axis (y-axis) represents the other.



Suppose we want to examine the relationship between the number of hours students spend studying (x-axis) and their exam scores (y-axis). We collect data from a group of students, recording the number of hours each student studied and their corresponding exam scores.



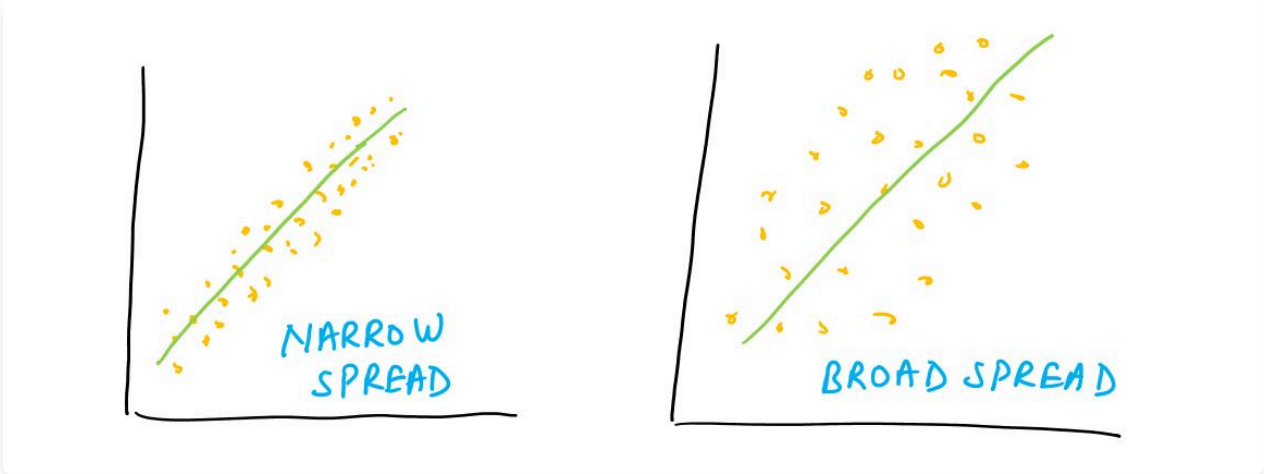
The overall pattern of the points in the scatter diagram provides insights into the relationship between the variables.

1. **Positive Relationship**: Points generally move upward from left to right, indicating that as one variable increases, the other tends to increase.(higher study hours associated with higher scores)

2. **Negative Relationship**: Points generally move downward from left to right, suggesting that as one variable increases, the other tends to decrease. (higher study hours associated with lower scores)

3. **No Relationship**: Points are scattered without a clear pattern, indicating a lack of a strong relationship between the variables. (no clear relation between study hours and scores)

Outliers are points that significantly deviate from the overall pattern. They may indicate unusual or extreme observations and can influence the interpretation of the relationship.



The **spread or concentration of points** can provide a sense of the strength of the relationship. A narrow spread suggests a strong relationship. A broader scatter indicates a weaker relationship between two variables.

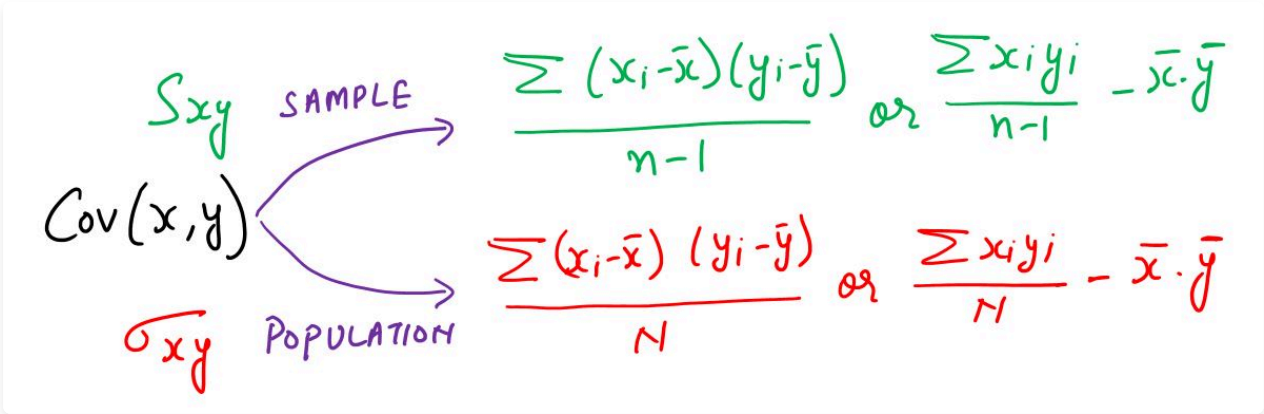
From a scatter diagram, we can derive a **trendline**, which is drawn with the assistance of regression analysis. This line is a representation of the overall trend or pattern observed in the scatter plot. Trendlines are instrumental in providing insights into the relationship between two variables. They offer a visual summary of the general direction and slope of the data points, aiding in the identification of trends, patterns, or correlations. Essentially, trendlines help us understand the underlying behavior or tendency in the data and make predictions about future observations based on this observed trend.

---

3. Covariance

Covariance mesaures the degree to which two variables change together. In other words, it assesses the joint variability of two random variables. Specifically, covariance measures whether an increase in one variable is associated with an increase or decrease in another variable.

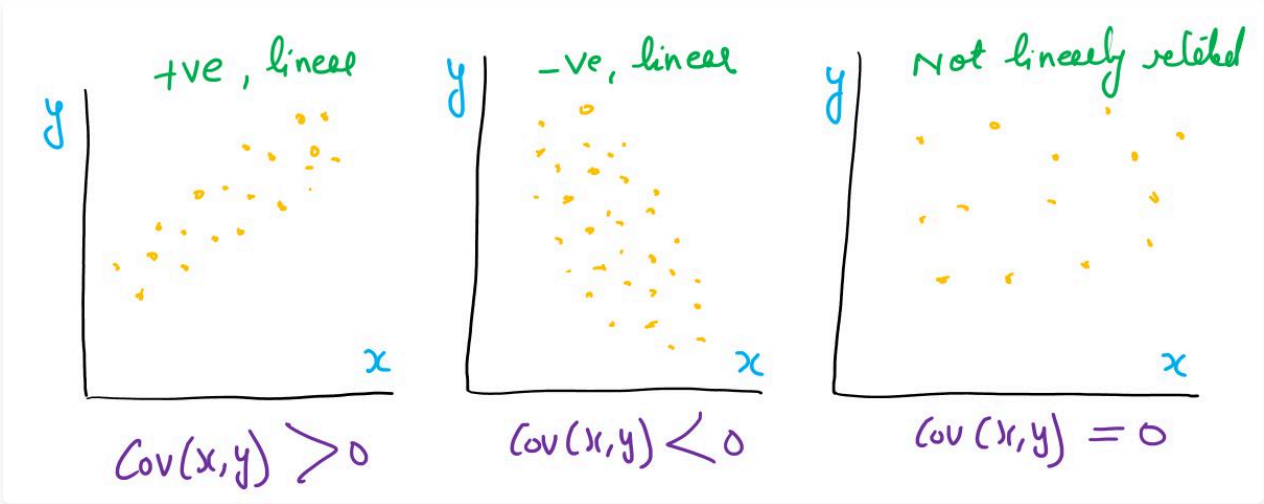
The formula for the Covariance, denoted by Cov(x,y), between two variables, x and y, based on a set of n paired observations, can be expressed as:



A **positive covariance** indicates that as one variable increases, the other tends to increase as well.

A **negative covariance** suggests that as one variable increases, the other tends to decrease.

A **Zero covariance** suggests that there is no relationship between two variables.



Covariance is expressed in the product of the units of the two variables. The magnitude of covariance is not standardized and depends on the units of the variables. Therefore, it can be challenging to interpret the strength of the relationship.

For example, if we have two variables, the weight of an object (x) measured in kilograms and the distance it is thrown (y) measured in meters, and we calculate the covariance, the result would be in units of kilogram-meters. Now, if we decide to change the weight values from kilograms to grams (keeping the relationship the same), the numerical value of the covariance will change.

To address this, correlation coefficient is often used, as it scales the covariance by the standard deviations of the variables.

---

3. Covariance

Suppose we have two variables, X and Y, representing the number of hours spent studying (X) and the corresponding exam scores (Y) for a group of students. The table below shows the data for a sample of five students. Calculate the covariance between the number of hours spent studying and the corresponding exam scores for this sample.

STUDENT	HOURS OF STUDY	EXAM SCORE
A	3	75
B	5	88
C	4	80
D	2	60
E	6	92

Solution:

STUDENT	HOURS OF STUDY (x_i)	EXAM SCORE (y_i)	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$
A	3	75	-1	-4	4
B	5	88	1	9	9
C	4	80	0	1	0
D	2	60	-2	-19	38
E	6	92	2	13	26
	20	395			77

$\bar{x} = \frac{20}{5} = 4$

$\bar{y} = \frac{395}{5} = 79$

$$Cov(x,y) = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

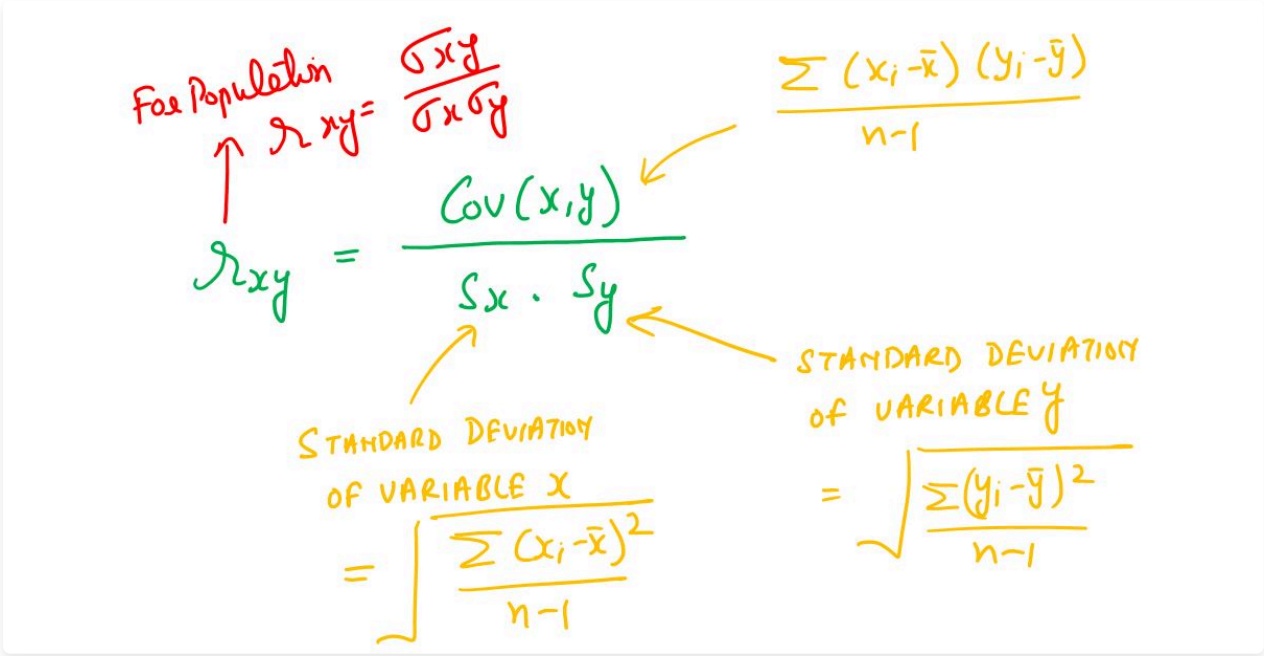
$$= \frac{77}{5-1} = 19.25$$

The covariance between the number of hours spent studying and the corresponding exam scores for this sample is 19.25.

4. Correlation Coefficient

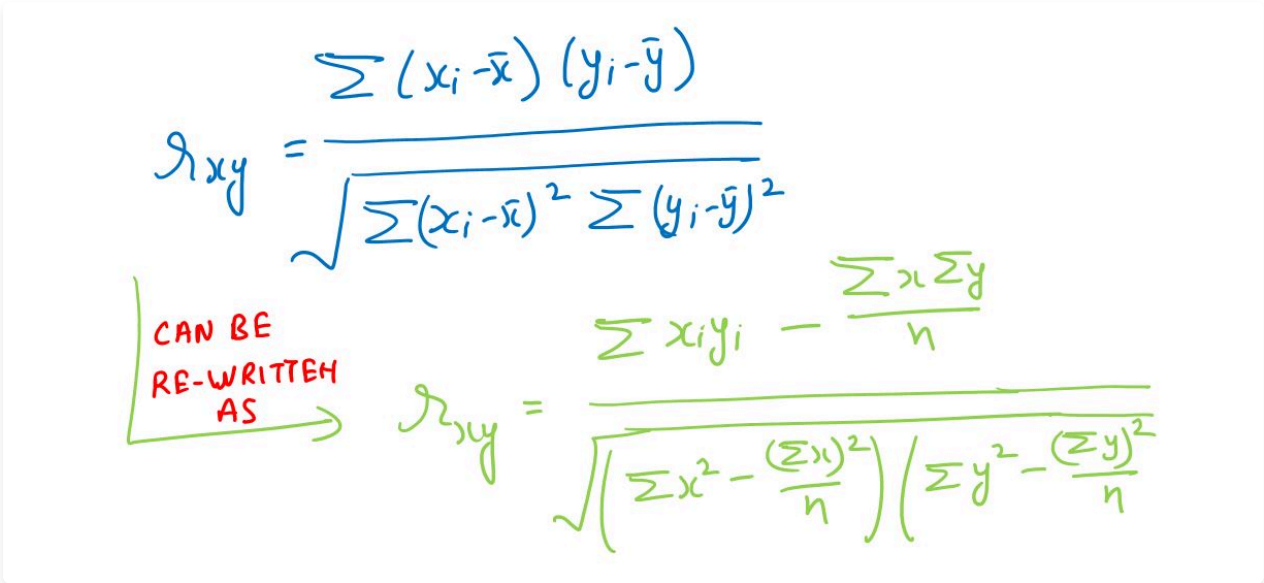
To overcome shortcomings of Covarianvce (depends upon unit of measurement), we use Correlation Coefficient.

The Karl Pearson Product-Moment Correlation Coefficient, commonly referred to as Pearson correlation coefficient or simply Pearson's r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is denoted by the symbol r_{xy}.

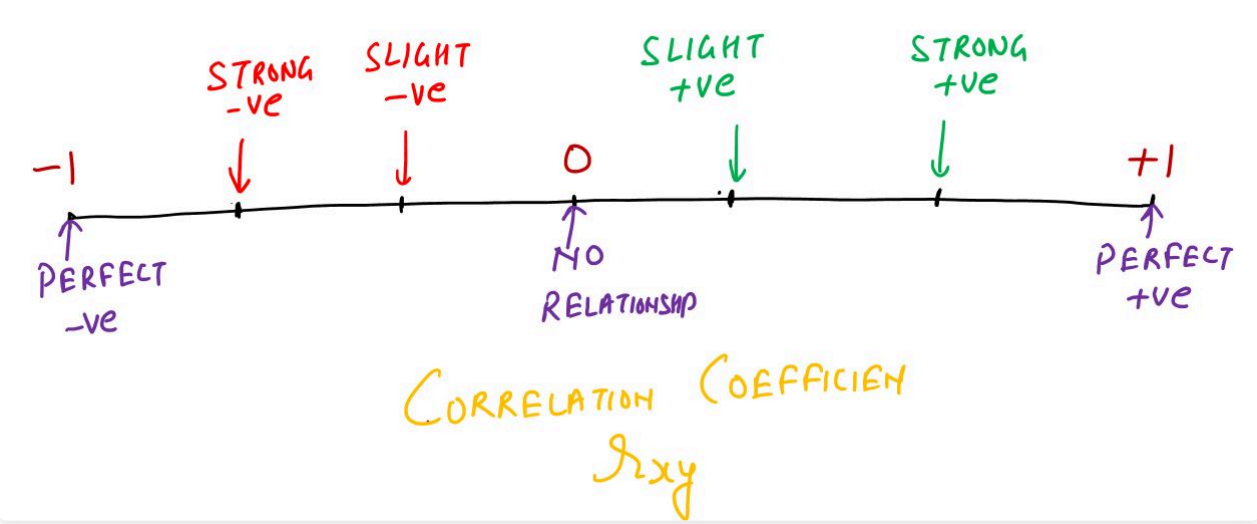


This formula represents the standardized form of covariance, allowing for a more easily interpretable measure of the linear relationship between two variables. By dividing the covariance by the product of the standard deviations, Pearson's correlation coefficient becomes independent of the specific units of measurement used for the variables.

The formula can be rewriiten as given below:



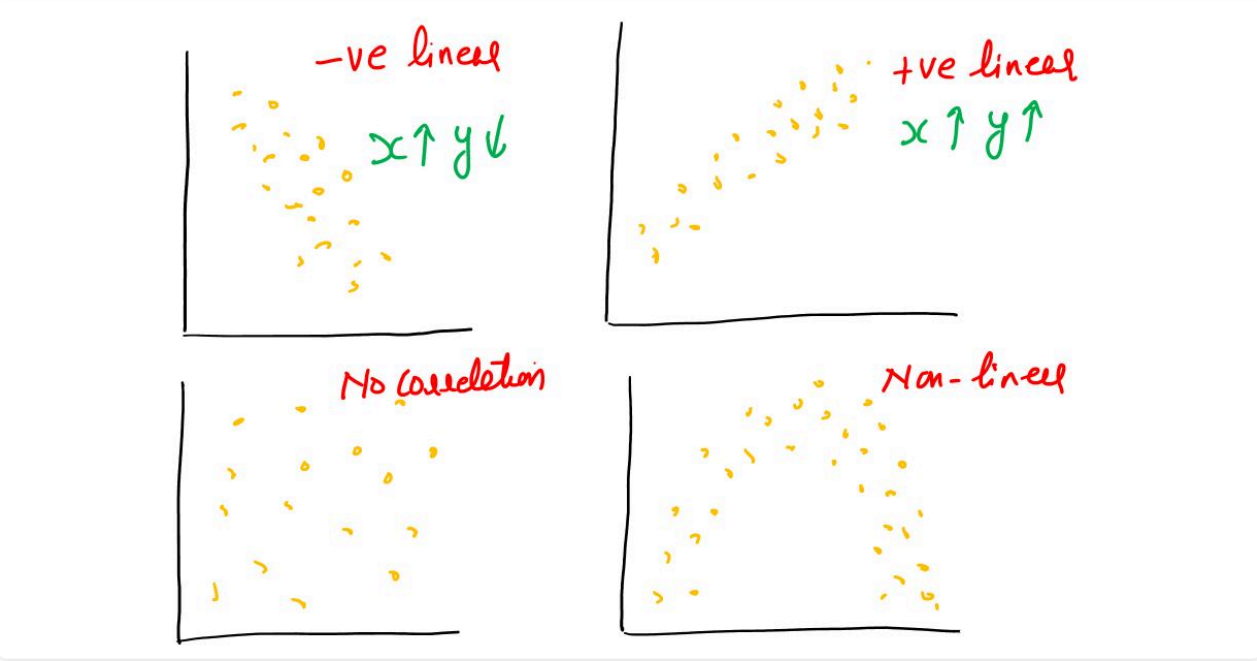
Pearson's r ranges from -1 to 1. The closer the absolute value of r is to 1, the stronger the linear relationship between the variables.



Let us discuss examples for each type of correlation:

1. Perfect Positive Linear Relationship (r = 1)
Suppose we are studying the relationship between the number of hours students spend studying and their exam scores. If every student in a class who studies more hours tends to have a higher exam score, and the relationship is perfectly aligned, we might observe a Pearson correlation coefficient (r) close to 1.

2. Perfect Negative Linear Relationship (r = -1)
Imagine we are examining the relationship between the amount of time spent commuting to work and job satisfaction. If every individual who spends more time commuting reports lower job satisfaction, and the relationship is perfectly consistent, we could observe a Pearson correlation coefficient (r) close to -1.



3. No Linear Relationship (r = 0)
Consider a scenario where we analyze the relationship between the number of hours spent watching TV and shoe size in a group of students. If, upon plotting the data, we find no clear trend or pattern suggesting that one variable influences the other, the Pearson correlation coefficient (r) may be close to 0, indicating no linear relationship.

4. Positive Correlation (0 < r < 1)
In a study exploring the relationship between regular exercise and heart health, if individuals who engage in more regular exercise tend to have better heart health, we might observe a positive correlation. The Pearson correlation coefficient (r) would be positive, indicating a positive linear relationship.

5. Negative Correlation (0 > r > -1)
Suppose we investigate the relationship between the number of hours spent on social media and self-reported productivity. If

individuals who spend more time on social media tend to report lower productivity levels, the Pearson correlation coefficient (r) would be negative, indicating a negative linear relationship.

Pearson's correlation coefficient assumes that the relationship between the variables is linear.

Pearson's r is a standardized measure, making it suitable for comparing the strength of relationships across different pairs of variables.

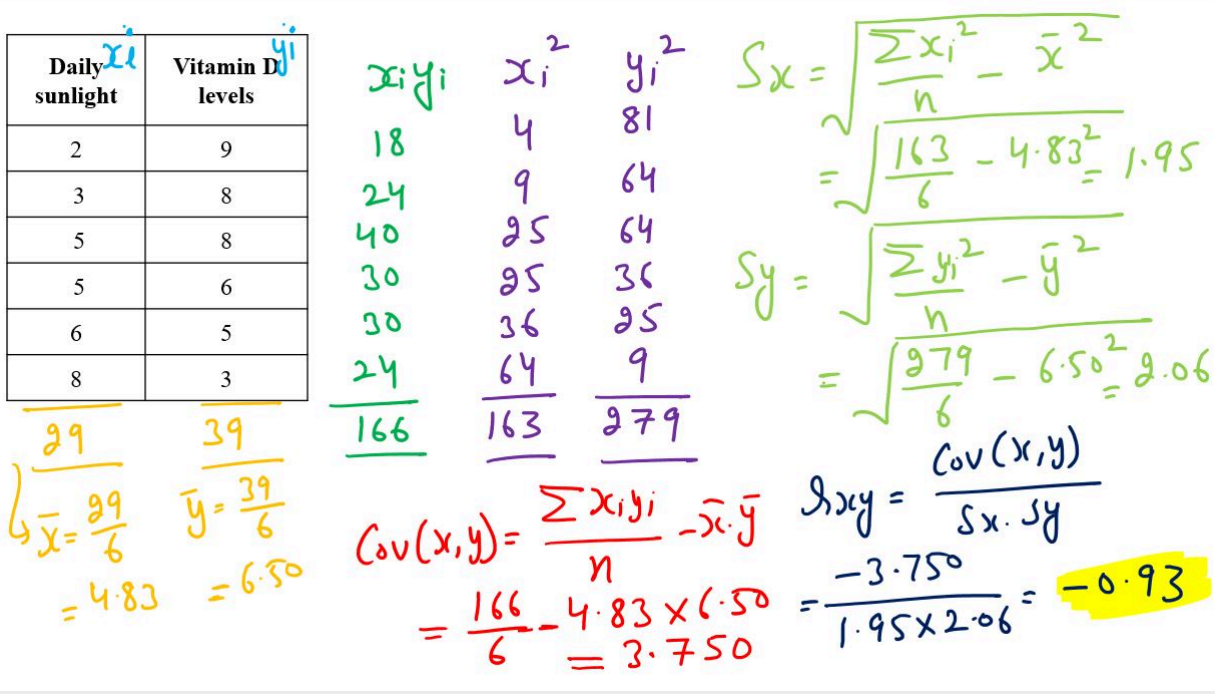
---

4. Correlation Coefficient

In a study examining the relationship between daily sunlight exposure and vitamin D levels in individuals, a researcher collected data from a sample of 6 participants. The number of hours of daily sunlight exposure and the corresponding vitamin D levels were recorded. The goal is to investigate the correlation between sunlight exposure and vitamin D levels.

Number of hours of daily sunlight	Vitamin D levels
2	9
3	8
5	8
5	6
6	5
8	3

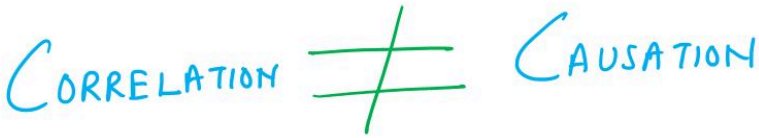
Solution:



The correlation coefficient of -0.93 suggests a very strong negative linear relationship between daily sunlight exposure and vitamin D levels. As sunlight exposure increases, vitamin D levels tend to decrease significantly in the studied sample of individuals.

4. Correlation Coefficient

Correlation coefficient measures the strength and direction of a linear relationship between two variables but it does not imply causation.



For instance, for a restaurant, there might be a positive correlation between review ratings and mean prices. But it does not necessarily mean that higher prices are causing better ratings. This type of association is known as a spurious relationship.

Thus, causation implies a direct influence of one variable on another, leading to a correlation. However, the reverse is not necessarily true—just because two variables are correlated does not imply a causal relationship. Other factors, known as confounding variables, may contribute to the observed correlation, making it crucial to carefully analyze causation beyond the presence of correlation.

Consider few more examples to understand this better.

Consider the correlation between the number of firefighters at a fire scene and the extent of damage caused by the fire. While there might be a positive correlation, indicating that larger fires attract more firefighters, it would be a mistake to infer causation. The true causation is that larger fires naturally require more firefighting resources, leading to both a correlation and a causal relationship.

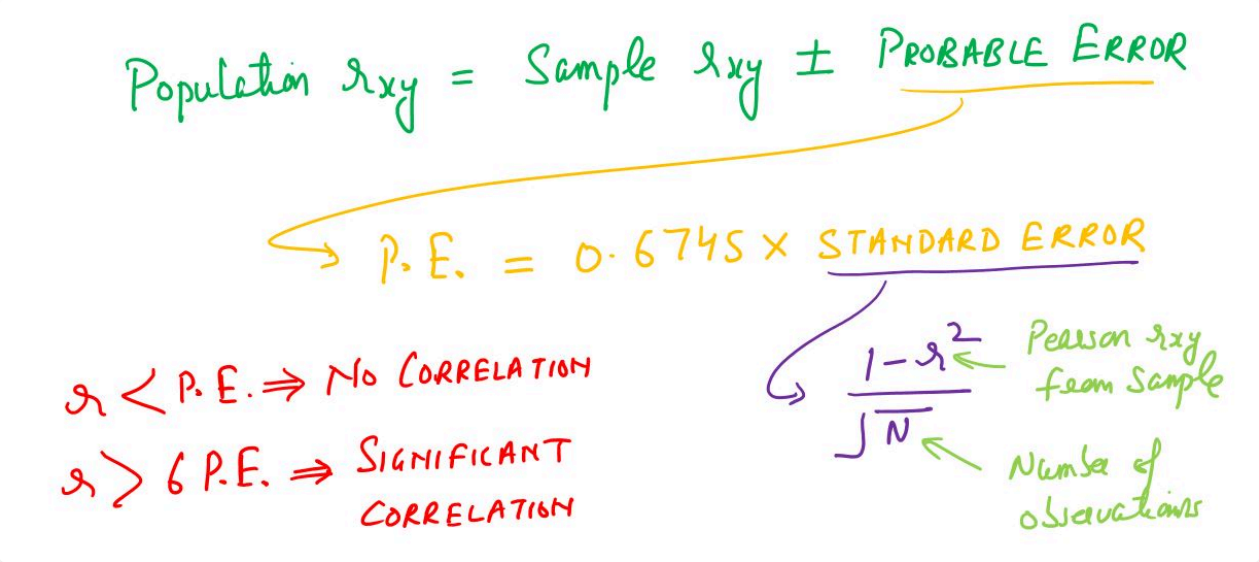
In a study on office productivity, there might be a negative correlation between the number of breaks employees take and their overall productivity. However, assuming that taking fewer breaks would automatically boost productivity oversimplifies the relationship. The true causation might involve a third factor, such as workload or job satisfaction, influencing both break frequency and productivity.

---

5. Probable Error

The probable error is a measure of the variability or precision associated with the correlation estimate of population from sample.

It is given by following formula:



If 0.6745 is omitted from the formula of **Probable Error**, we get the Standard Error of coefficient of correlation.

However, it is important to note that the concept of probable error is not commonly used these days. In the past, probable error was employed as a way to estimate the likely range within which the true correlation might fall. It represented a margin of error around the correlation coefficient. A smaller probable error indicated a more precise estimate.

In modern statistical approaches, the confidence interval provides a more intuitive and widely accepted measure of precision. A 95% confidence interval, for example, gives a range within which we can be 95% confident that the true correlation lies.

---

6. Spearman rank-order correlation

Spearman Rank Correlation (r_s) is a non-parametric measure of statistical dependence between two variables. It was given by Edward Spearman.

Unlike Pearson correlation, Spearman correlation does not assume that the variables have a linear relationship. Instead, it assesses the strength and direction of monotonic relationships, whether they are increasing or decreasing.

A monotonic relationship is a type of association between two variables where the direction of the relationship (increasing or decreasing) is consistent across the entire range of values of the variables.

Spearman Rank Correlation is particularly useful when dealing with ordinal or ranked data. This method is also robust to outliers.

Steps to Calculate Spearman Rank Correlation are given below.

1. Ranking

For each variable, rank the data from lowest to highest. In case of ties, we use either Bracket Rank method or Average rank method to rank. Following example shows how these methods are used to rank, in case of ties.

2. Calculate the Differences

Find the differences (d_i) between the ranks of corresponding pairs of observations.

3. Square the Differences

Square each difference (d_i^2).

4. Sum the Squared Differences

Calculate the sum of the squared differences.

5. Apply the Formula to calculate Spearman rank correlation coefficient (r_s)

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

IF THERE ARE TIES $\Rightarrow$

$$r_s = 1 - \frac{6\left[\sum d_i^2 + \frac{\sum (t_j^3 - t_j)}{12}\right]}{n(n^2-1)}$$

t_j = Tied Rank

The Spearman rank correlation coefficient (r_s) ranges from -1 to 1.

A positive value indicates a monotonic increasing relationship, while a negative value indicates a monotonic decreasing relationship. A value of 0 suggests no monotonic relationship.

---

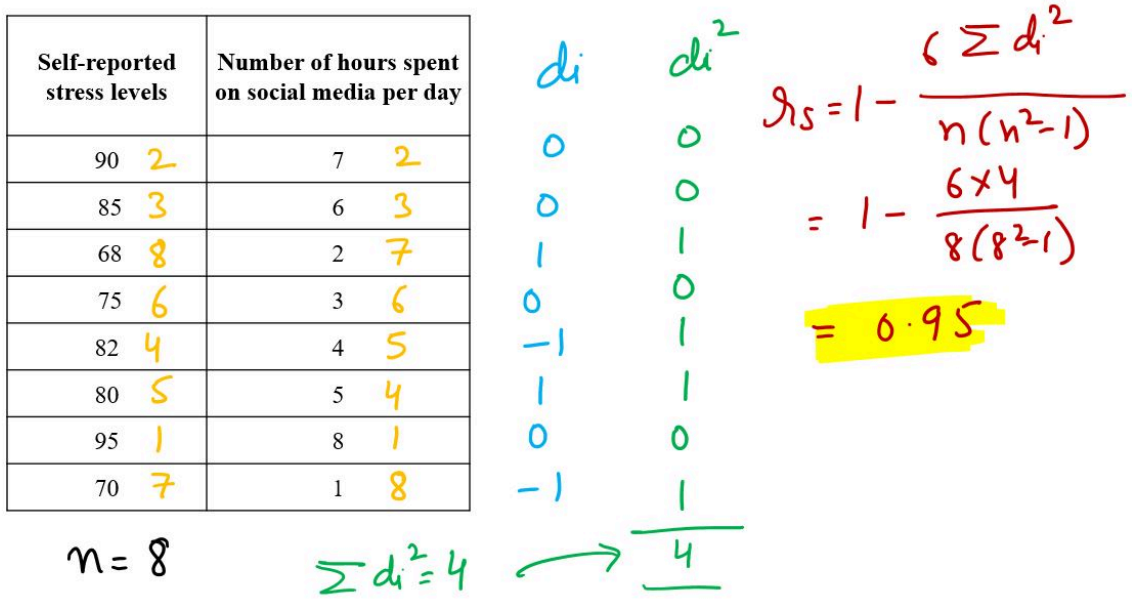
6. Spearman rank-order correlation

A researcher is investigating the relationship between the number of hours spent on social media per day and self-reported stress levels in a group of 8 individuals. The data is given below.

Self-reported stress levels	Number of hours spent on social media per day
90	7
85	6
68	2
75	3
82	4
80	5
95	8
70	1

Use the Spearman Rank Correlation to compute the correlation coefficient. Interpret the result regarding the strength and direction of the relationship between social media usage and self-reported stress levels.

Solution:



A Spearman Rank Correlation coefficient of 0.95 indicates a very strong positive monotonic relationship between the number of hours spent on social media per day and self-reported stress levels in the studied group of 8 individuals. This suggests that as the time spent on social media increases, there is a consistent and substantial tendency for higher self-reported stress levels.

---

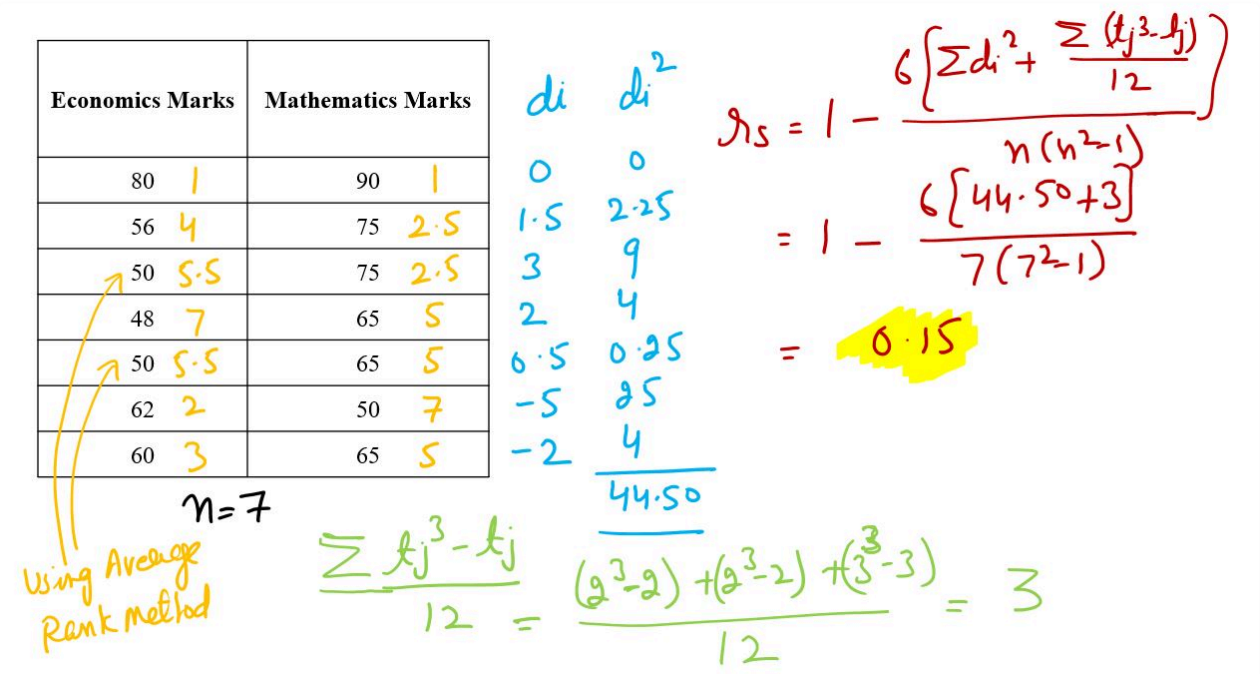
6. Spearman rank-order correlation

A researcher is exploring the relationship between the marks obtained in Economics and Mathematics among a group of 7 students. Each student's marks in Economics and Mathematics have been recorded.

Economics Marks	Mathematics Marks
80	90
56	75
50	75
48	65
50	65
62	50
60	65

Use the Spearman Rank Correlation to compute the correlation coefficient.

Solution:



A Spearman Rank Correlation coefficient of 0.15 suggests a very weak positive monotonic relationship between the marks obtained in Economics and Mathematics among the group of 7 students. This indicates that there is a slight tendency for higher marks in Economics to be associated with slightly higher marks in Mathematics. However, the relationship is weak, and the correlation is not substantial.

---

7. Coefficient of Concurrent Deviation

The Coefficient of Concurrent Deviation is another approach to assessing the correlation between two variables.

However, it is considered a poor method, and its results may lack reliability.

Let us delve into this method step by step to better comprehend its procedure. The data is given below.

X values	Y values
25	35
28	34
30	35
23	30
35	29
38	28
39	26
42	23

First, we consider values of x variable. We mark a postive sign if value of x in the any row is greater than the value in previous row. Otherwise we mark it with negative sign. The first row is left blank.

X values	Y values
25	35
28 +	34
30 +	35
23 −	30
35 +	29
38 +	28
39 +	26
42 +	23

Similary, we do this for y variable.

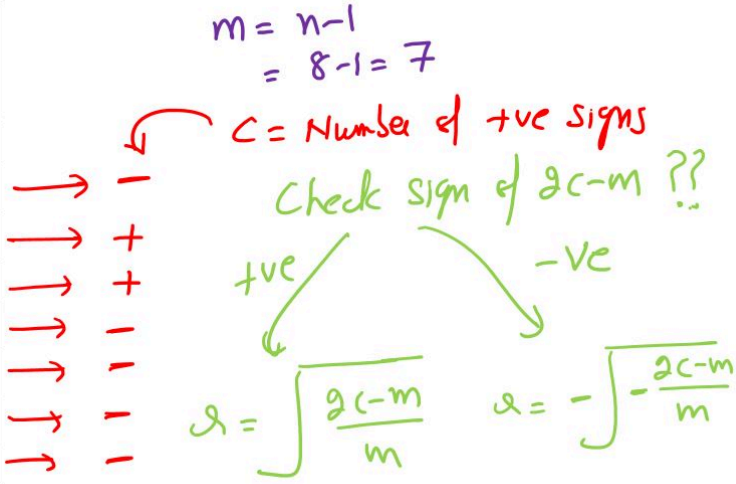
X values	Y values
25	35
28 +	34 −
30 +	35 +
23 −	30 −
35 +	29 −
38 +	28 −
39 +	26 −
42 +	23 −

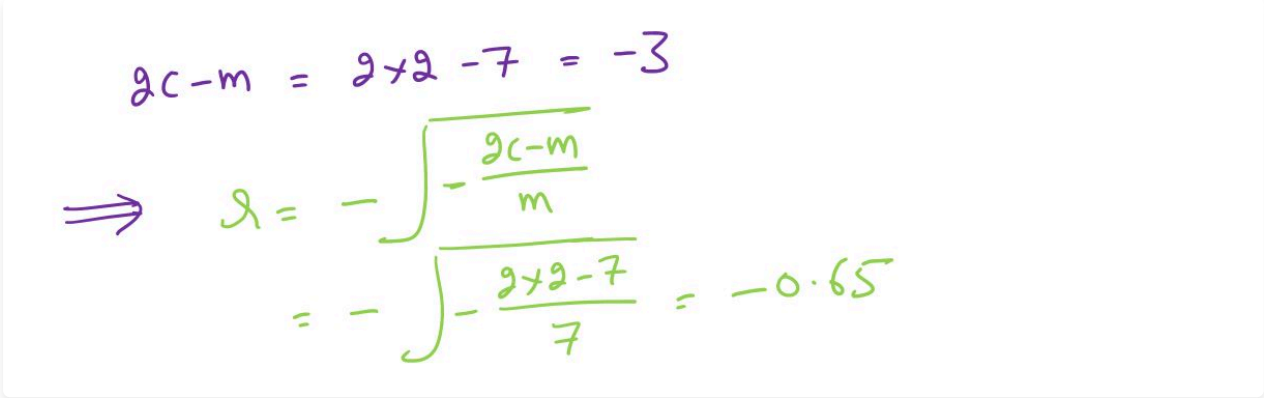
Then we multiply signs of x and y variables, for each correponding pair.

X values	Y values	
25	35	
28 +	34 −	→ −
30 +	35 +	→ +
23 −	30 −	→ +
35 +	29 −	→ −
38 +	28 −	→ −
39 +	26 −	→ −
42 +	23 −	→ −

The formula of correlation coefficient depends upon the sign of (2c-m).

X values	Y values	
25	35	
28 +	34 −	→ −
30 +	35 +	→ +
23 −	30 −	→ +
35 +	29 −	→ −
38 +	28 −	→ −
39 +	26 −	→ −
42 +	23 −	→ −





The value of -0.65 imply that there is negative correlation between x and y.

---

8. Comparison of Pearson and Spearman coefficients

The Pearson and Spearman correlation coefficients can range in value from −1 to +1. For the Pearson correlation coefficient to be +1, when one variable increases then the other variable increases by a consistent amount. This relationship forms a perfect line. The Spearman correlation coefficient is also +1 in this case.

If the relationship is that one variable increases when the other increases, but the amount is not consistent, the Pearson correlation coefficient is positive but less than +1. The Spearman coefficient still equals +1 in this case.

When a relationship is random or non-existent, then both correlation coefficients are nearly zero.

Pearson correlation coefficients measure only linear relationships. Spearman correlation coefficients measure only monotonic relationships. So, a meaningful relationship can exist even if the correlation coefficients are 0. One should examine a scatterplot to determine the form of the relationship.

Similar to Spearman's rank-order correlation coefficient, **Kendall's tau correlation coefficient** is designed to capture the association between two ordinal variables.

---

9. Comparison between Covariance and Correlation

Let us understand few important differences between Covariance and Correlation.

1. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.

2. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.

3. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between $-\infty$ and $+\infty$.

4. Covariance is affected by the change in scale, i.e. if all the value of one variable is multiplied by a constant and all the value of another variable are multiplied, by a similar or different constant, then the covariance is changed. As against this, correlation is not influenced by the change in scale.

5. The coefficient of correlation is independent of change of scale and origin of the variable x and y. Change of origin means some value has been added or subtracted in the observation. Change of scale means some value is multiplied or divided to observations.

6. Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.

---

1. Introduction

Regression analysis involves constructing a mathematical model, often an equation, to predict a specific variable based on the influence of one or more other variables.

The variable being predicted is termed the **dependent variable**, and the variables used to make predictions are known as **independent variables**.

The objective is to establish a relationship that enables the estimation or forecasting of the dependent variable by considering the effects of the independent variables.

For example, regression analysis can be applied to predict the future value of a stock (dependent variable) based on various factors like market trends, interest rates, and company performance (independent variables).

Similarly, regression analysis can be utilized to predict students' academic performance (dependent variable) based on factors such as study hours, attendance, and socioeconomic background (independent variables).

If there is only a single independent variable involved, the method is referred to as **simple linear regression** (also called Bivariate Regression). Conversely, when there are multiple independent variables in the analysis, it is termed as **multiple regression**.

Sometimes, in academic world, the dependent variable is also called Predicted variable, Explained variable, Response variable or Endogenous variable.

Similarly the independent variable is also called Predictor variable, Explainatory variable, Stimulus variable or Exogenous variable.

---

2. Simple Linear Regression

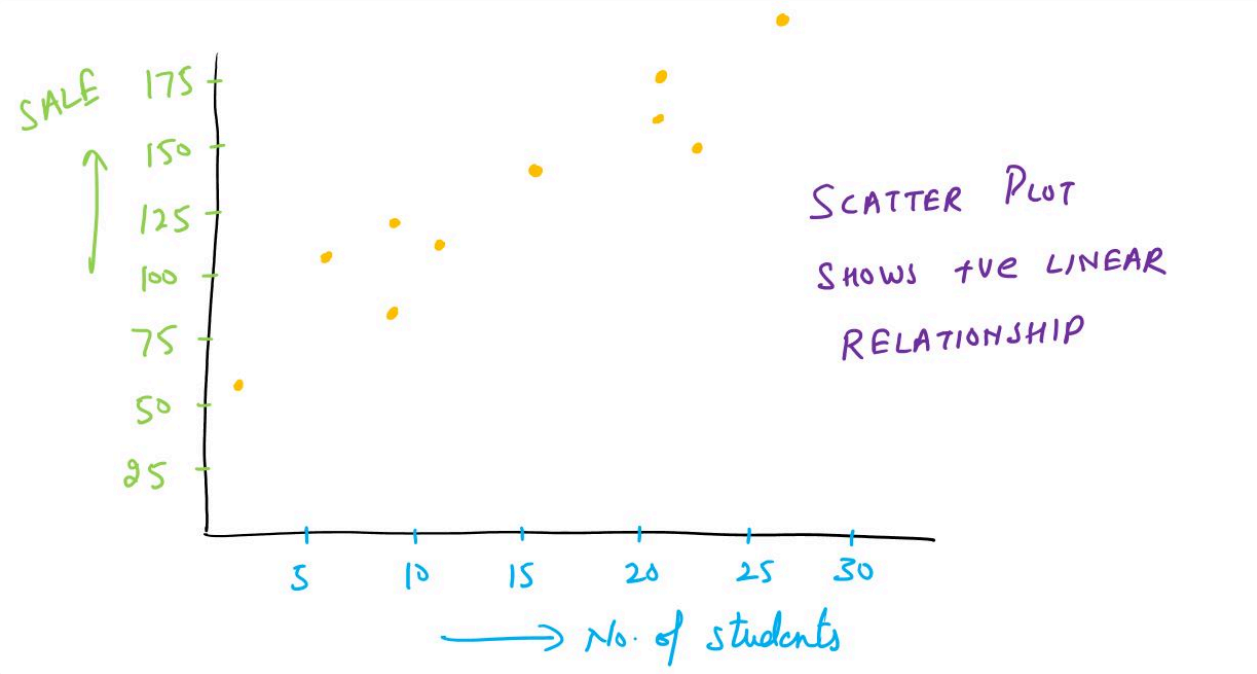
Let us walk through the process of building a simple linear regression model using the Least Squares method step by step.

For illustration, let us consider hypothetical data on the number of students in a college and the corresponding canteen sales.

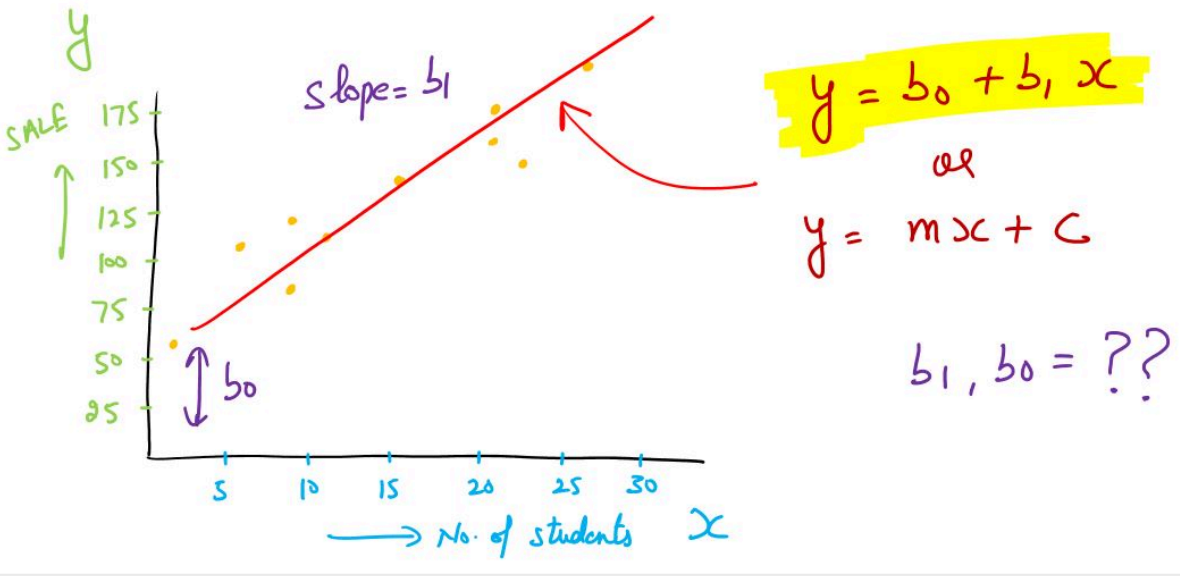
Step 1: Data Collection
Collect data on the number of students (x) and canteen sales (y). The data is given below.

Population of college (in '000)	Sale in canteen (in lakhs)
2	58
6	105
8	88
8	118
12	117
16	137
20	157
20	169
22	149
26	202

Step 2: Plotting the Data
Create a scatter plot with the number of students (x) on the x-axis and canteen sales (y) on the y-axis. Visualize the scatter plot to determine if a linear relationship exists between the number of students and canteen sales.



Step 3: Formulate Simple Linear Equation
Then let us formulate the equation of a straight line. It is given by following equation.

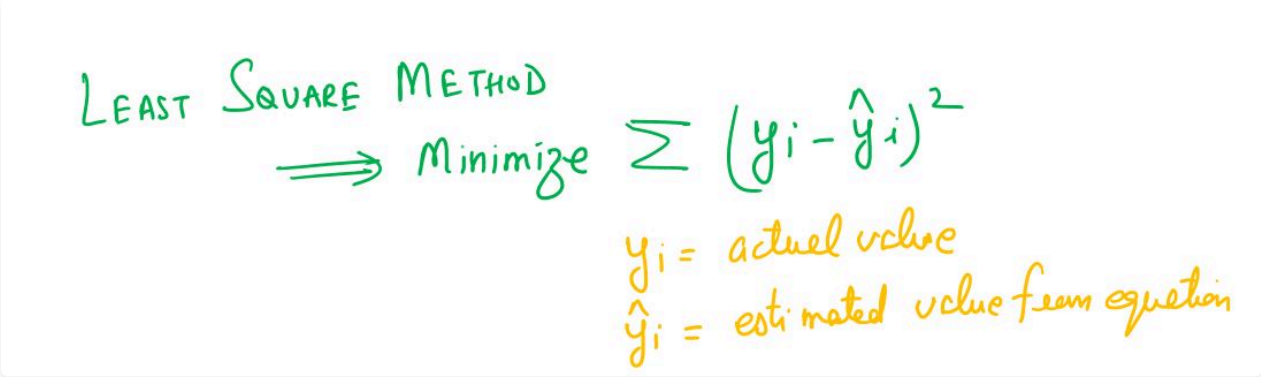


Here y is canteen sales, x is the number of students, b_1 is the slope, and b_0 is the intercept.

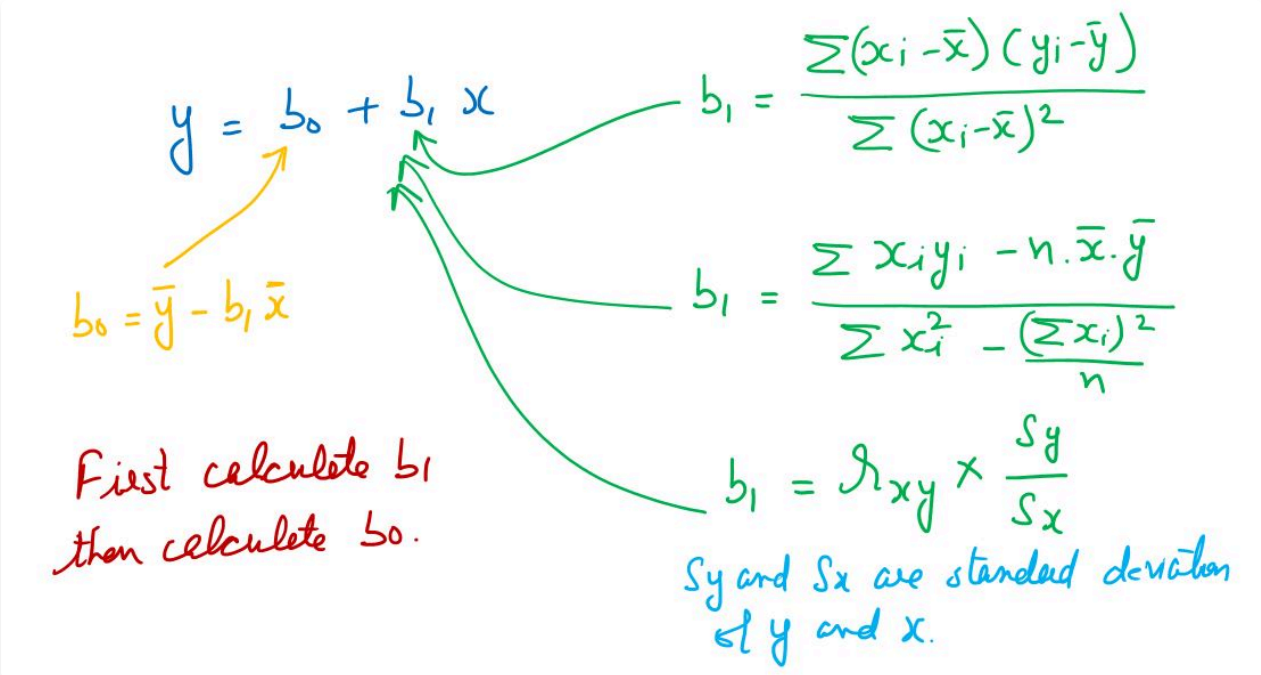
Sometimes, this equation is written as y = mx+c, where m is the slope, and c is the intercept.

Step 4: Least Squares Method

Use the Least Squares method to find the values of b_1 and b_0 that minimize the sum of squared differences between the observed and predicted values.



The formulas for b_1 (slope, m) and b_0 (intercept, c) are:



This equation represents the relationship between the number of students and canteen sales based on the provided data. You can use this equation to make predictions and interpret the impact of the number of students on canteen sales.

The calculations of b_0 and b_1 are given below:

Population of college (in '000)	Sale in canteen (in lakhs)	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
2	58	−12	−72	864	144
6	105	−8	−25	200	64
8	88	−6	−42	252	36
8	118	−6	−12	72	4
12	117	−2	−13	26	4
16	137	2	7	14	36
20	157	6	27	162	36
20	169	6	39	234	36
22	149	8	19	152	64
26	202	12	72	864	144
140				2840	568

$\bar{x} = \frac{140}{10} = 14$

1300 $\bar{y} = \frac{1300}{10} = 130$

$y = b_0 + b_1 x$

$b_1 = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2} = \frac{2840}{568} = 5$

$b_0 = \bar{y} - b_1 \bar{x} = 130 - 5 \times 14 = 60$

EQUATION

$y = 60 + 5x$

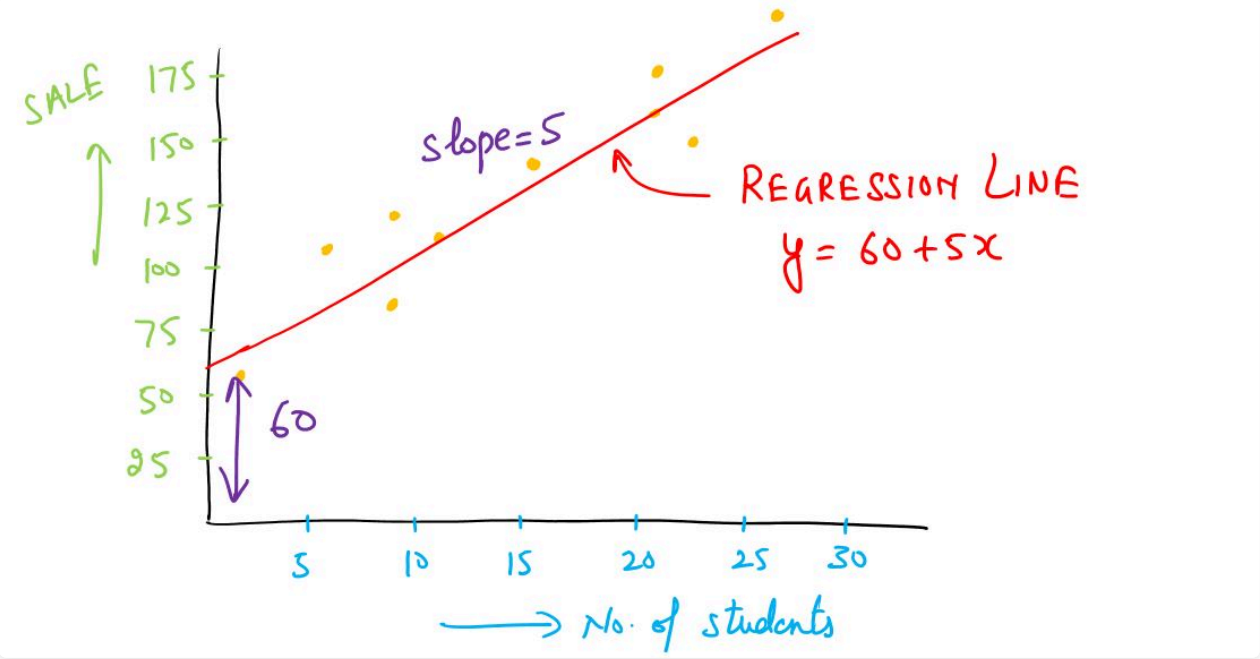
b_0 b_1

Step 5: Model Evaluation

Assess the goodness of fit using metrics like R-squared, which quantifies the proportion of variance explained by the model. We will do this step leter.

Step 6: Prediction and Interpretation

The final equation y=5x+60 implies the relationship between the number of students (x) and canteen sales (y) based on the provided data.



Here is how you can interpret the equation and draw conclusions:

1. For each additional student, canteen sales increase by 5 units. The positive slope indicates a positive correlation between the number of students and canteen sales.

2. Even with zero students, the expected canteen sales would be 60 units. This represents the initial level of sales when there are no students.

3. The positive slope suggests that as the number of students increases, there is a positive impact on canteen sales. This aligns with the intuitive expectation (from scatter plot) that a higher student population would likely lead to increased sales.

4. We can use the equation to predict canteen sales for different numbers of students. For example, if there are 70 students ($x=70$), you can predict y (canteen sales) as $5\times70+60=410$.

5. It is important to note that this model is based on the specific data provided, and its accuracy depends on the representativeness of the dataset. It assumes a linear relationship between the number of students and canteen sales, and any non-linearity or additional influential factors may affect the model's accuracy.

6. We should validate the model's performance with new data to ensure its generalizability beyond the given dataset.

---

2. Simple Linear Regression

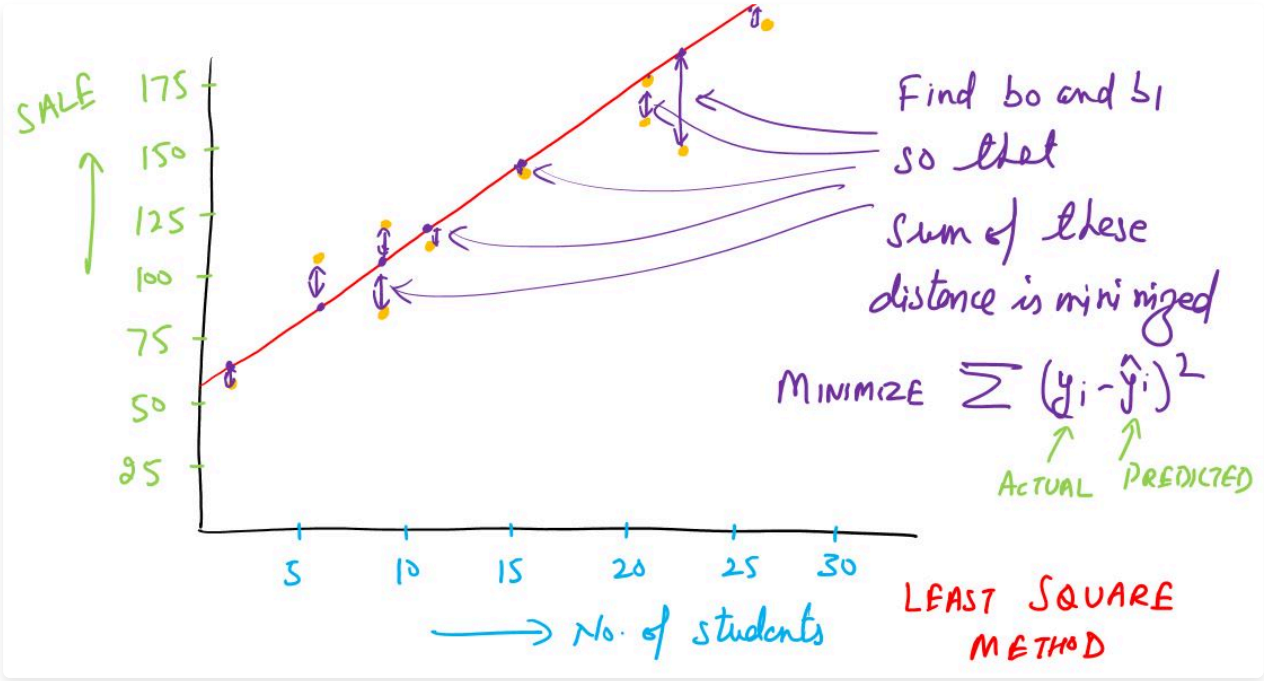
Let us understand, how Least Squares Method works, in our regression analysis.

The Least Squares method in regression analysis to find the best-fitting line (or curve) through a set of data points. The objective is to minimize the sum of the squared differences between the observed values and the values predicted by the model.

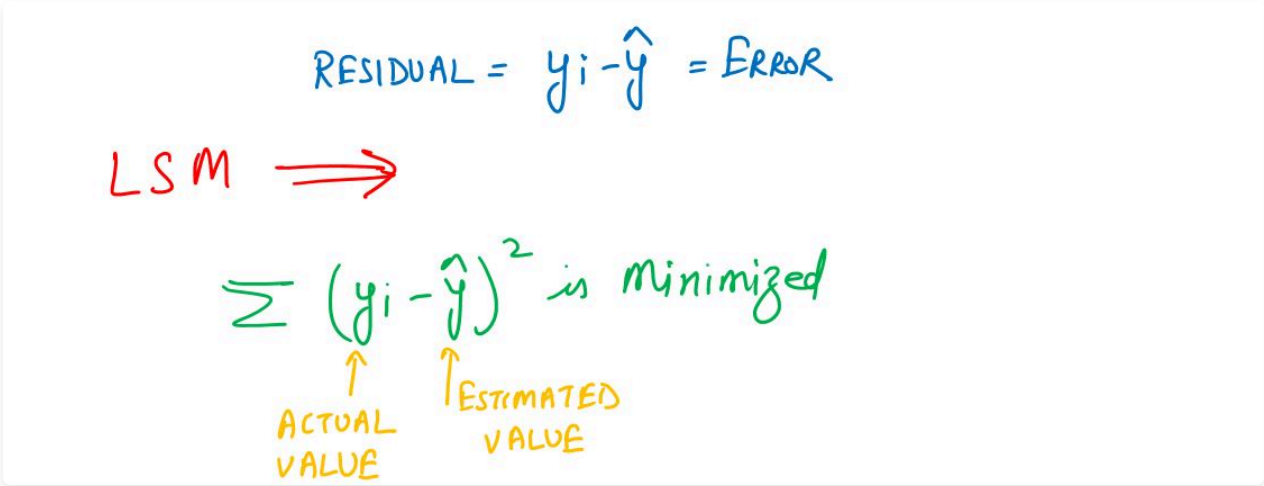
From the data on the number of students and canteen sales, we know that at x=6, the corresponding canteen sales is 105. This is the actual value of y, represented by y_i.

By putting the value of x=6 into our regression equation (y=60+5x), we obtain the estimated (predicted) value of canteen sales to be 90. This is represented by $\hat{y}$.

The vertical distance between y_i and $\hat{y}$ is called the residual.



The Least Squares method works in such a way that it finds those b_0 and b_1 (sometimes written as m and c if the equation is y=mx+c) so that the sum of all "Residuals" is minimized.



This is how the Least Squares method works.

3. Regression Coeffifcients

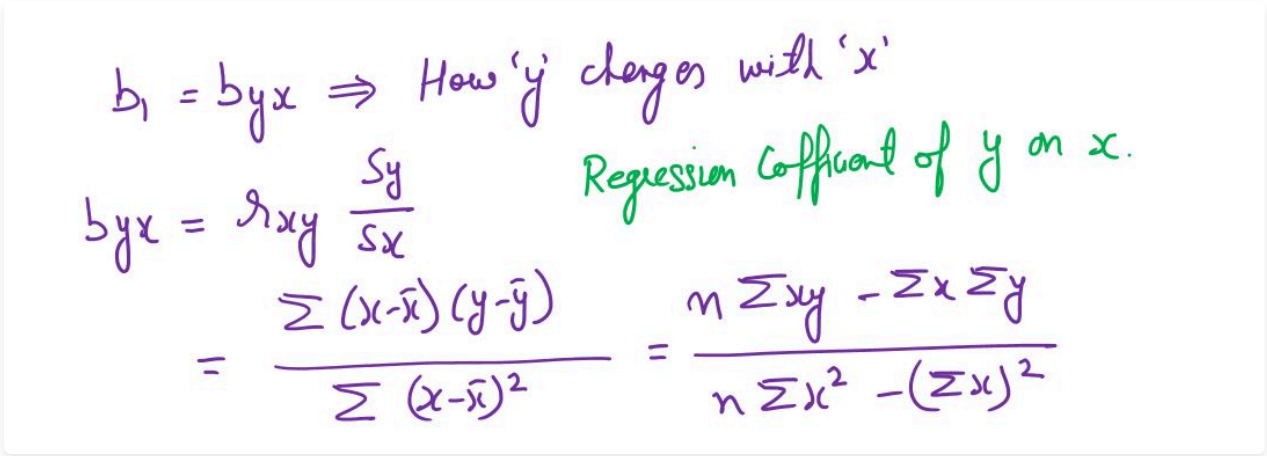
From the data on the number of students and canteen sales, we determined the regression equation to be y=60+5x. The values of the coefficients b_0 and b_1 are 60 and 5, respectively.

Regression of y on x

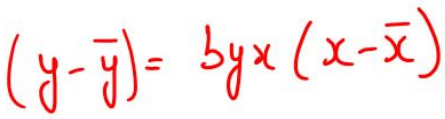
Through this equation, we can understand how y changes with variations in x. This is termed the regression of 'y on x'. The coefficient b_1 is referred to as the regression coefficient of 'y on x' and is also denoted by b_{yx}.

In ourexample, it means how canteen sales (y) change as the number of students (x) changes.

As discussed earlier, the value of b_1 or b_{yx} is given by:



The regression equation can also be expressed as:



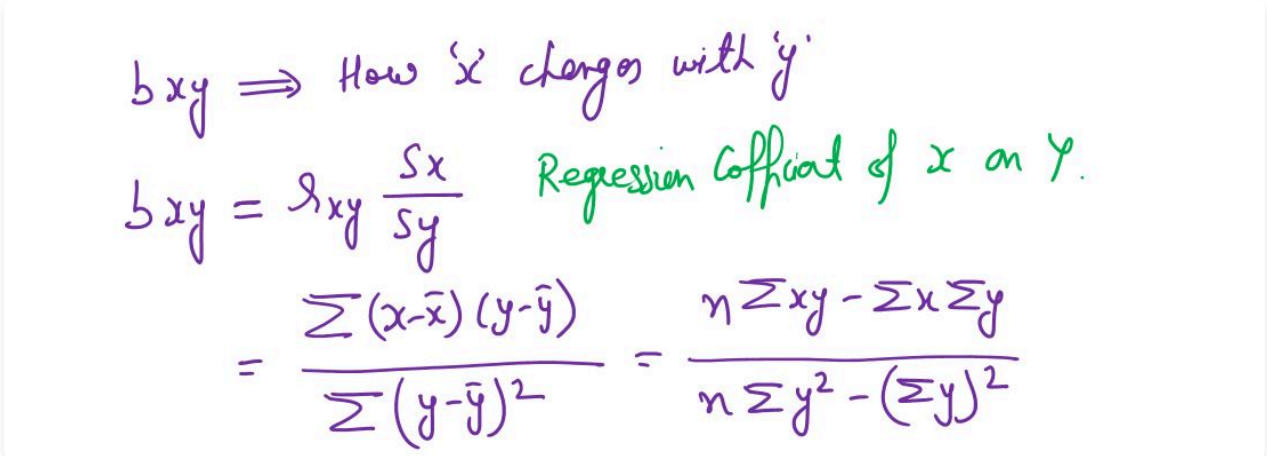
Regression of x on y

Similary, we can also identify, how x changes with variations in Y.

This is termed the regression of 'x on y'. The coefficient is referred to as the regression coefficient of 'x on y' and is denoted by b_{xy}.

In our example, it means how the number of students (x) changes with variations in canteen sales (y).

The value of b_{xy} is given by:



The regression equation can also be expressed as:

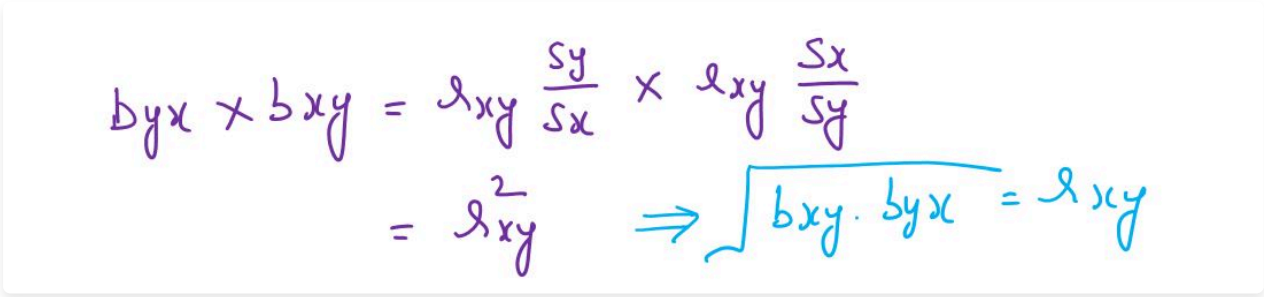
$$(x-\bar{x}) = b_{xy}(y-\bar{y})$$

3. Regression Coeffifcients

Let us discuss few important properties of Regression Coefficients (b^{yx} and b_{xy}).

Property 1

Multiplying \(b_{xy} \) and \(b_{yx} \) we get:



Thus product of two regression coefficients in equal to square of correlation coefficient. We can also say that the geometric mean between two regression coefficients is equal to the coefficient of correlation.

Property 2

The Sign of \(b_{xy} \) and \(b_{yx} \) are always same. When both are positive, r_{xy} is positive, when both are negative, then r_{xy} is negative.

Property 3

The product of two regression coefficients will always be between 0 to 1.

If one regression coefficient is greater than unity, then the other regression coefficient must be lesser than unity.

Property 4

Arithmetic mean of both regression coefficients is equal to or greater than coefficient of correlation.

Property 5

Regression coefficients are independent of the change of origin but not independent of change of scale.

Property 6

The point where two regression lines intersect is determined by the means of the x variable and the y variable (Point is \(\bar x \) , \(\bar y \)) .

$(y-\bar{y}) = b_{yx}(x-\bar{x})$
$(x-\bar{x}) = b_{xy}(y-\bar{y})$
} meet at $(\bar{x}, \bar{y})$

Property 7

Both regression lines will coincide if r_{xy} is 1 or -1.

If $r = \pm 1$

$$\frac{x-\bar{x}}{S_x} = \pm \frac{y-\bar{y}}{S_y}$$ (only one regression equation)

Property 8

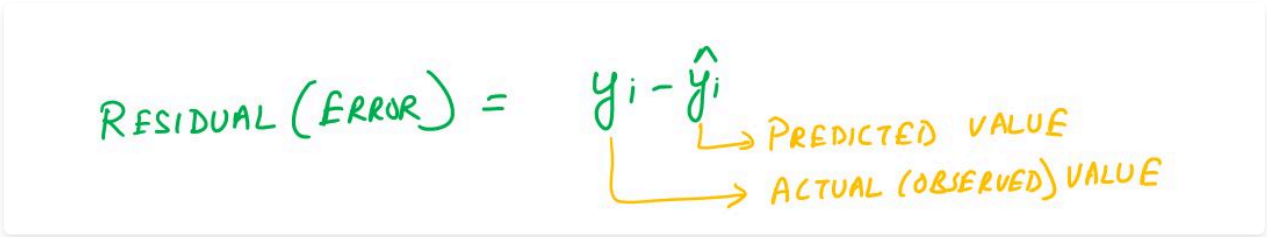
If r_{xy} is 0, the the regression equations will be y = \(\bar y \) and x = \(\bar x \). These are parallel to x and y axis.

If $r = 0$

Regression equations will be $y = \bar{y}$ and $x = \bar{x}$

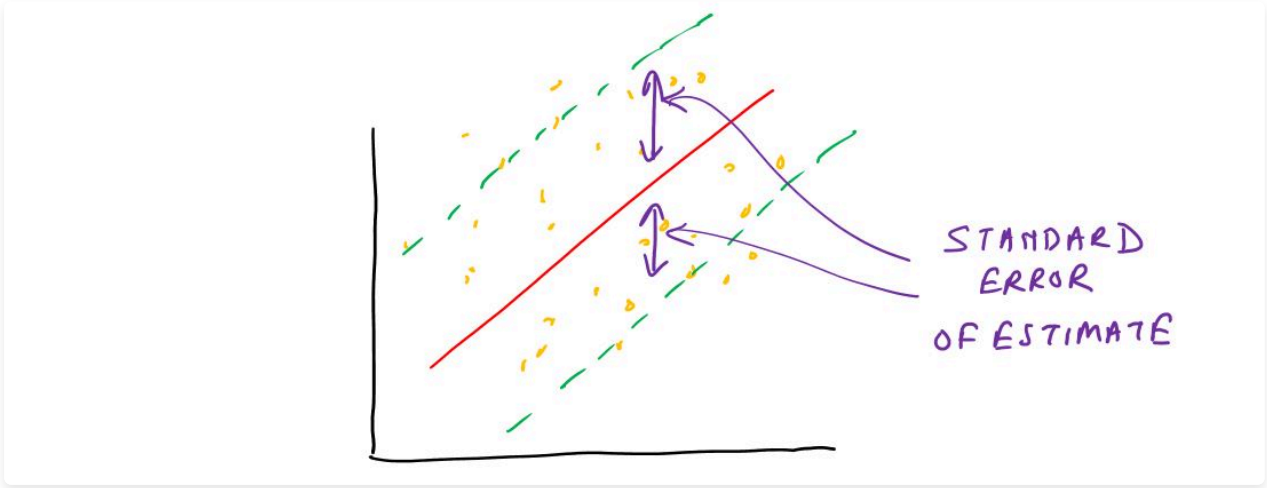
4. Standard Error of Estimate

We have learnt that the Least Squares Method (LSM) yields a line that best fits the data, but it is important to note that this line doesn't necessarily pass through all actual data points. Consequently, there are errors known as residuals.

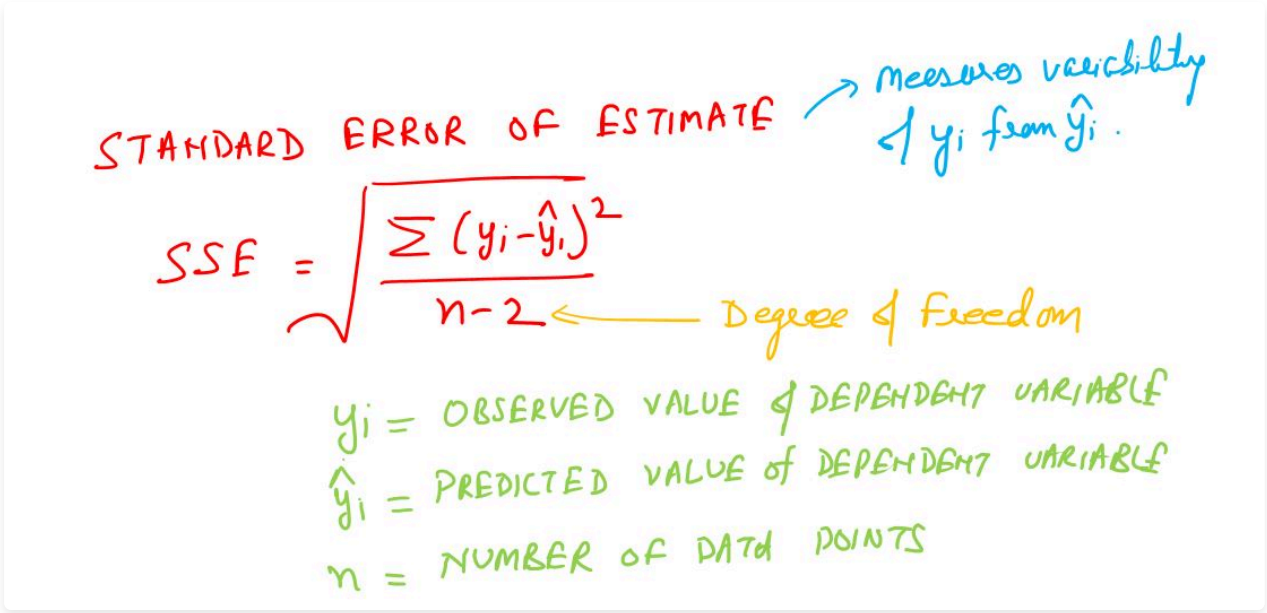


The effectiveness of our regression model hinges on our ability to minimize these residual errors.

The standard error of the estimate (SEE), also known as the standard deviation of the residuals, is a measure of the variability or dispersion of the observed values around the regression line in a regression analysis. It provides information about how well the regression equation predicts the dependent variable.



The standard error of the estimate is calculated using the following formula:



Thus, the standard error of the estimate is essentially a measure of the spread or dispersion of the residuals (the differences between observed and predicted values). Lower values suggest that the residuals are closer to the regression line.

A smaller standard error of the estimate indicates a better fit of the regression model to the data.

The formula includes an adjustment for the degrees of freedom (n-2), where n is the number of data points. This adjustment accounts for the fact that the estimated slope (b_1 or m) and intercept (b_0 or c) are based on sample data.

The standard error of the estimate is expressed in the same units as the dependent variable.

SSE allows for the comparison of different regression models. Models with lower standard errors of the estimate are considered better at explaining the variability in the dependent variable.

---

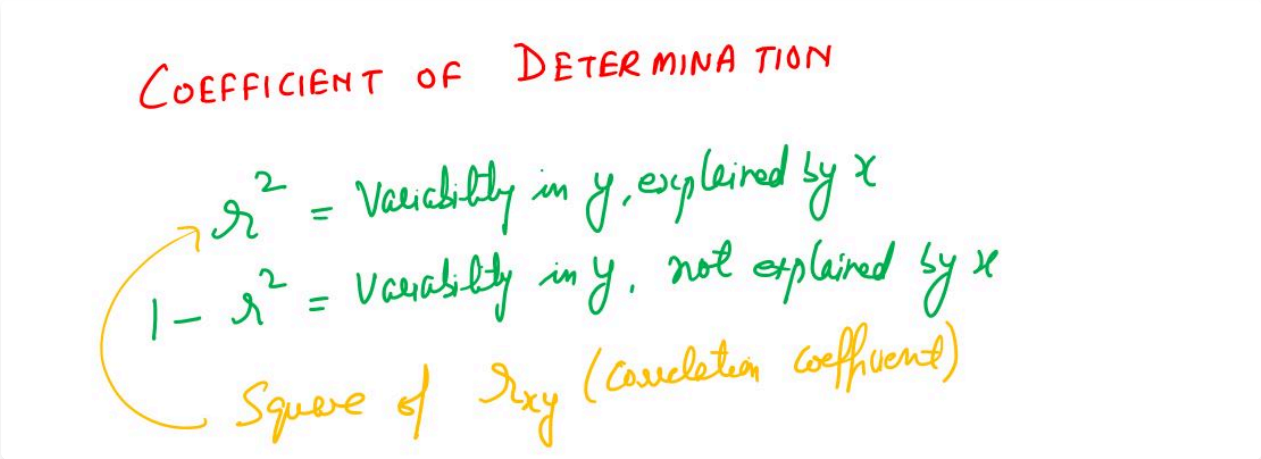
5. Coefficient of Determination

From the dataset on the number of students and canteen sales, we established the regression equation as y = 60+5x.

To evaluate how well this regression aligns with the actual data, we can employ the method of analyzing the standard error of the estimate, as previously discussed.

Now, let's delve into another method to assess the representation of the regression line in the actual data, **Coefficient of Determination**.

The Coefficient of Determination, represented by \(r^2 \) (R-square), indicates the proportion of variability in y that is explained by x. In mathematical terms:



A large \(r^2 \) suggests that a significant portion of the variability in y is explained by x, indicating a robust fit of the regression line. On the other hand, a small \(r^2 \) implies that y is not predominantly determined by x, suggesting the presence of other variables.

Imagine you are investigating the relationship between the number of hours spent exercising (x) and weight loss (y) over a month. After performing regression analysis, you find the equation y = −2+1.5x.

Coefficient of Determination, \(r^2 \), Suppose is 0.80. This means that 80% of the variability in weight loss (y) is explained by the number of hours spent exercising (x). In simpler terms, 80% of the changes in weight loss can be attributed to the variations in exercise time.

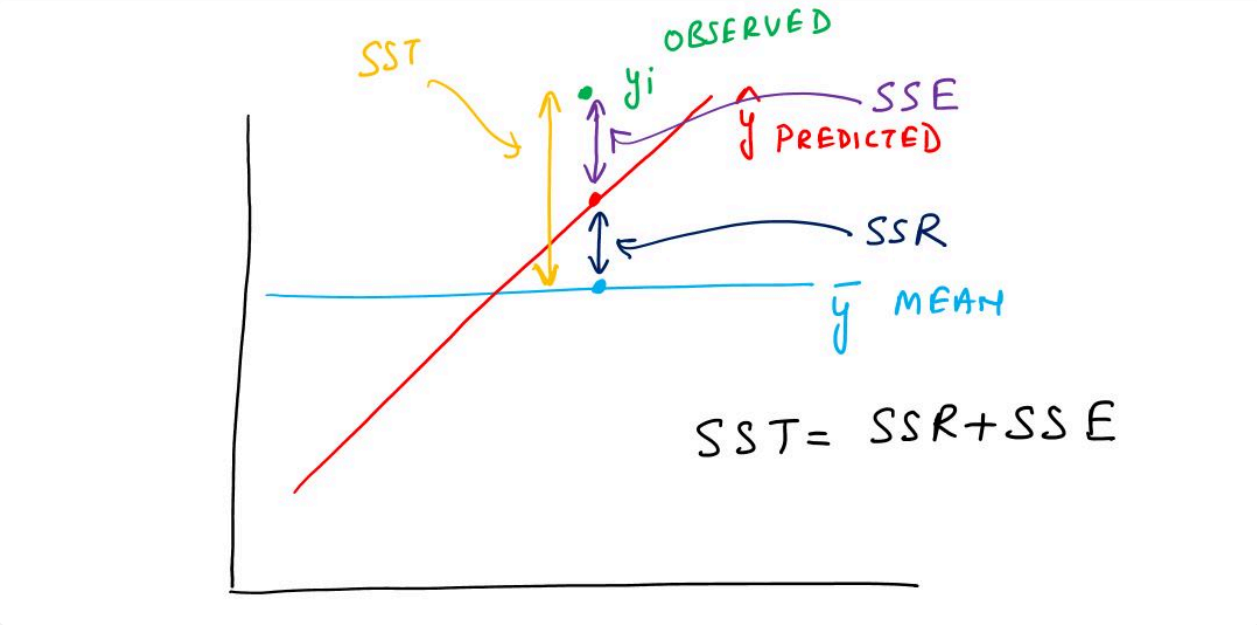
It is noteworthy that the Coefficient of Determination is nothing but the **square of the correlation coefficient**. Therefore, a high correlation between x and y will result in a large \(r^2 \), reinforcing the relationship between the variables in the regression model.

---

5. Coefficient of Determination

Let us consider example of the number of students and canteen sales.

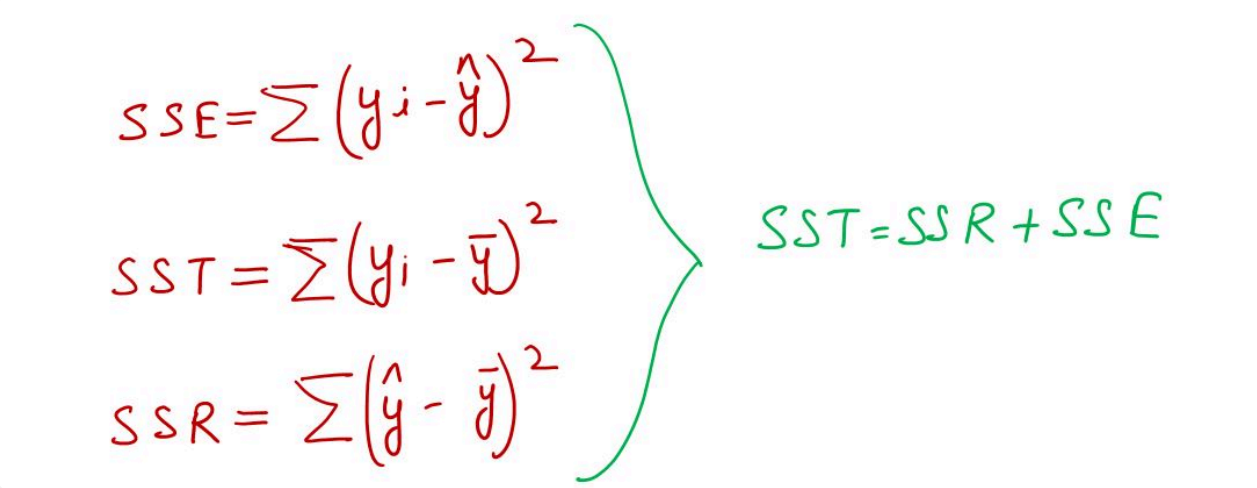
y_i : Actual (observed) values of the dependent variable.
$\hat{y}$: Predicted (estimated) values of the dependent variable.
$\bar{y}$ = Mean value of the dependent variable.



Sum of Squares due to Error (SSE)
SSE is the sum of squares of the difference between actual values (y_i) and predicted values ($\hat{y}$) of the dependent variable. This represents the residual error, and our goal is to minimize it.

Total Sum of Squares (SST)
SST is the sum of squares of the difference between actual values (y_i) and the mean value ($\bar{y}$) of the dependent variable. It indicates how well the data is clustered around the $\bar{y}$ line.

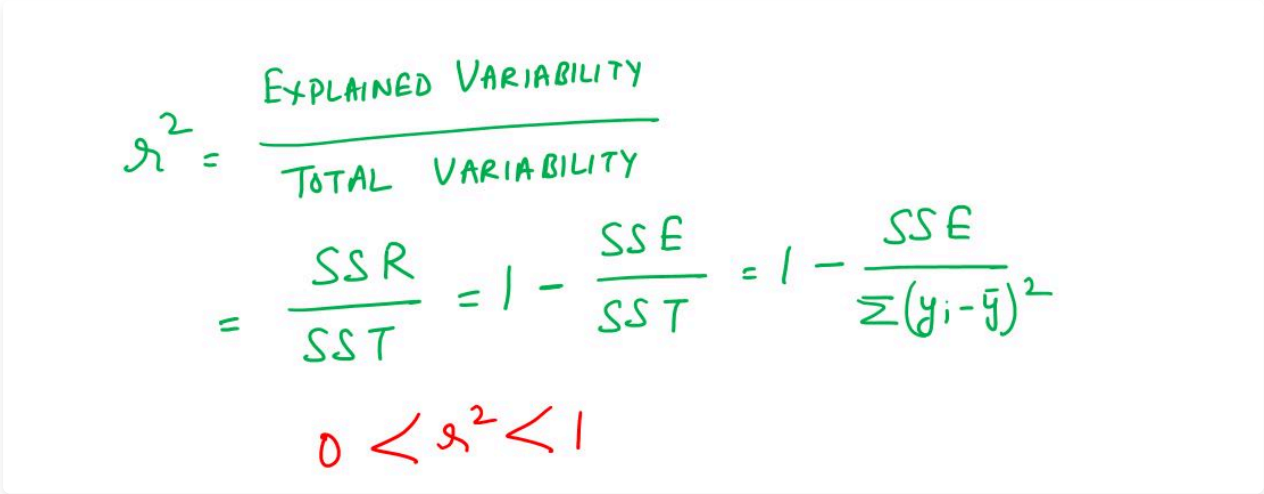
Sum of Squares due to Regression (SSR)
SSR shows how estimated values ($\hat{y}$) of the dependent variable deviate from the $\bar{y}$ line.



From the diagram, it can be see that:
SST = SSR + SSE

The relationship SST = SSR + SSE indicates the decomposition of variability into the explained variability (SSR) and the unexplained variability (SSE).

Since the Coefficient of Determination (\(r^2 \)) is defined as the ratio of the explained variability (SSR) to the total variability (SST). Therefore:



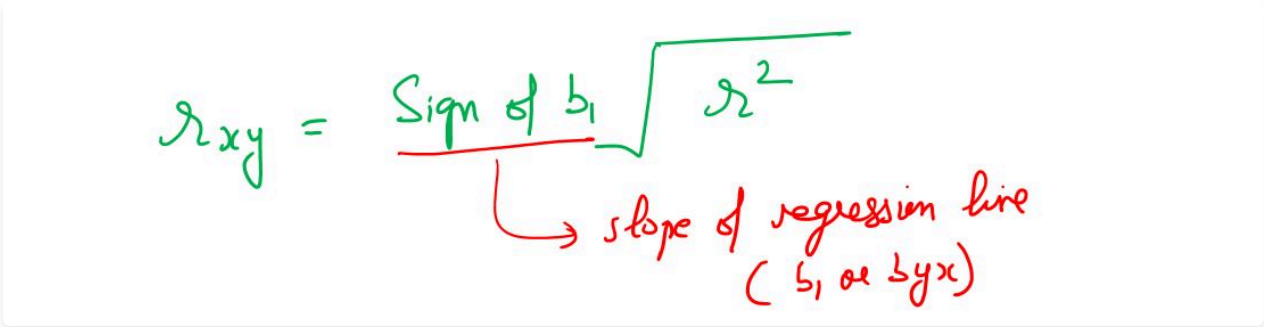
In simpler terms, \(r^2 \) expresses the proportion of total variability in the dependent variable that is explained by the regression model. It ranges from 0 to 1, where a higher \(r^2 \) value indicates a greater proportion of variability explained by the model.

---

6. Correlation Coefficient and Coefficient of Determination

Correlation Coefficient (r_{xy}) and Coefficient of Determination (r^2) are two measures that convey the strength of a relationship between two variables.

If we want to determine the correlation coefficient (r_{xy}) from the coefficient of determination (r^2), we simply take the square root of the coefficient of determination (r^2). Additionally, the sign of the correlation coefficient (r_{xy}) aligns with the sign of the slope of the regression line (b_1 or b_{yx}).



Both Correlation Coefficient (r_{xy}) and Coefficient of Determination (r^2) provide insights into the strength of the relationship between two variables. They quantify the degree to which changes in one variable are associated with changes in another.

While both measures assess the relationship strength, they differ in their interpretation and presentation.

The correlation coefficient (r_{xy}) ranges from -1 to 1, indicating the direction and strength of the linear relationship. On the other hand, the coefficient of determination (r^2) expresses the proportion of variability in the dependent variable explained by the independent variable, ranging from 0 to 1.

Notably, while the Correlation Coefficient (r_{xy}) is specifically for linear relationships, the Coefficient of Determination (r^2) can be applied to both linear and non-linear relationships.

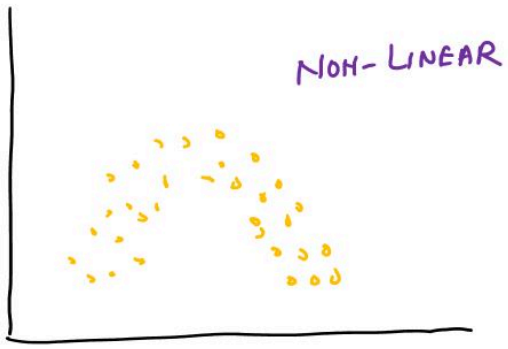
Additionally, the Correlation Coefficient (r_{xy}) is designed for one independent variable, whereas the Coefficient of Determination (r^2) can accommodate more than one independent variable.

---

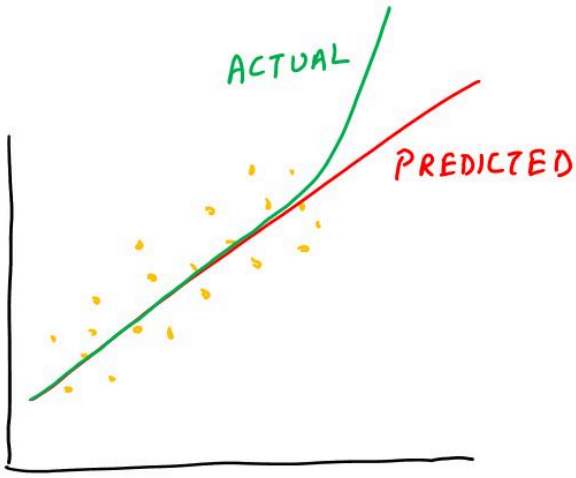
7. Limitations of Regression Analysis

Regression analysis has its limitations, and it may fail to provide accurate insights in certain situations:

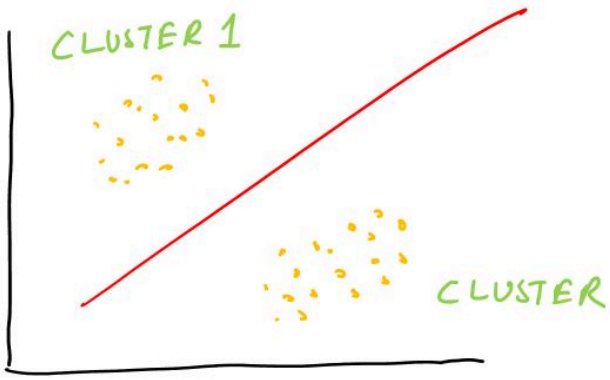
1. Nonlinear Relationships
Regression analysis assumes a linear relationship between variables. When the actual relationship is nonlinear, the model may not accurately capture the dynamics.



2. Extrapolation Beyond Data Range
Extrapolating predictions beyond the observed range of data can lead to inaccurate results. Behavior outside the range may differ, and predictions may not be reliable.

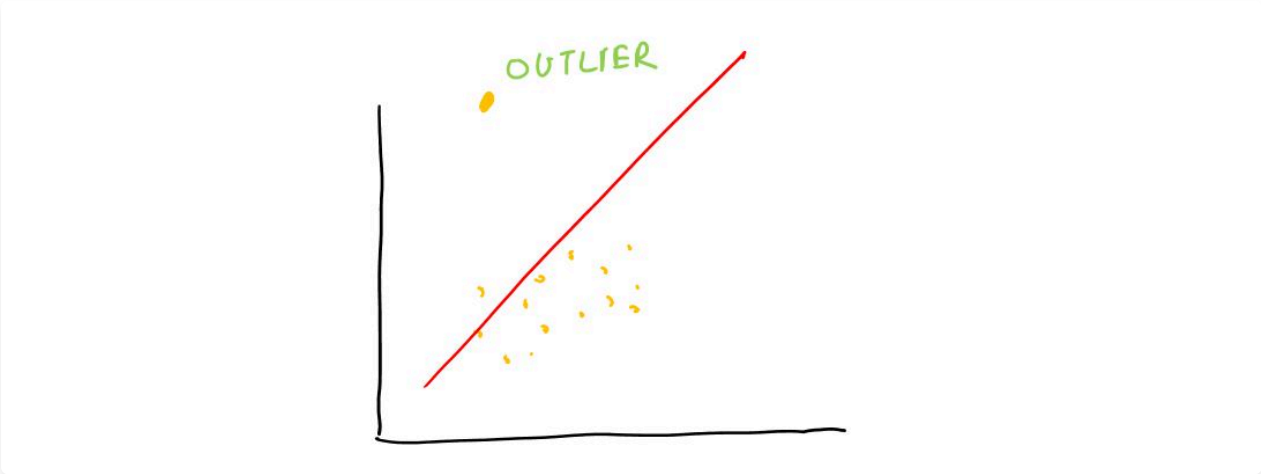


3. Clustering
If the data exhibits separate clusters, applying a single regression line using the Least Squares Method (LSM) may present a misleading picture. In such cases, conducting separate linear analyses for each cluster is advisable.



4. Outliers

Outliers, extreme values in the data, can distort the regression analysis by disproportionately influencing the model. Addressing outliers or considering robust regression techniques may be necessary.



5. Meaningless Intercept

The intercept in regression analysis may be meaningless in certain contexts. For instance, if salary is on the x-axis and mean age on the y-axis, a negative intercept is nonsensical, as age cannot be negative. Such situations require careful interpretation.

6. Multiple Variables

When dealing with more than one variable, a single-variable regression may not suffice. Multiple regression, which considers the impact of multiple independent variables on a dependent variable, is more appropriate in these cases.

---

1. Introduction

The statistical methods for inference, discussed till now, are known as **parametric methods**. These methods begin with an assumption about the probability distribution of the population (often that population has a normal distribution). Based upon this assumption, statisticians are able to derive the sampling distribution that can be used to make inferences about one or more parameters of the population, such as the population mean (μ) or the population standard deviation (σ) .

Let us now understand, **non-parametric methods** which can be used to make inferences about a population without requiring an assumption about the specific form of the population's probability distribution. For this reason, these nonparametric methods are also called **distribution-free methods.** Most of the statistical methods referred to as parametric methods require quantitative data, while nonparametric methods allow inferences based on either categorical or quantitative data.

---

2. Mann-Whitney-Wilcoxon Test

Mann-Whitney-Wilcoxon (MWW) test is a non-parametric test for the difference between two populations based on an independent sample from each population. The null hypothesis is that the two populations are identical. If the assumption can be made that the populations have the same shape, this test provides an inference about the difference between the medians of the two populations.

The combined ranks for the data from the two samples are obtained and the test statistic for the MWW test is the sum of ranks for the first sample.

In most applications, the samples sizes are large enough to use a normal approximation with the continuity correction factor in conducting the hypothesis test. If no assumption is made about the populations, the MWW procedure tests whether the two populations are identical. If the assumption can be made that the two populations have the same shape, the test provides an inference about the difference between the medians of the two populations.

Let us understand with an example.

Let us suppose that there are 4 boys and 5 girls in a class. They participate in a 100 metres race and are ranked from 1 to 9. We want to find out, if there is evidence of a significant difference in performance for boys compared to girls.

BOYS	RANK	GIRLS	RANK
B_1	4	G_1	5
B_2	1	G_2	6
B_3	7	G_3	9
B_4	2	G_4	3
		G_5	8

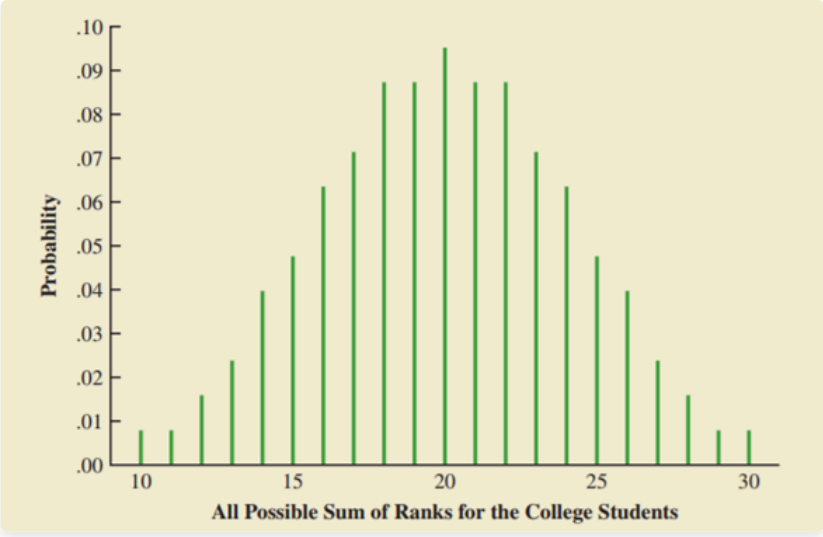
The Null and Alternate Hypothesis will be:

H_0 : Boys and Girls are identical in terms of performance in 100 m race.

H_a : Boys and Girls are not identical in terms of performance in 100 m race.

Next we sum the ranks for each sample. The sum of ranks for the first sample (Boys) will be the **test statistic W** for the MWW test. This sum, is W = 4 + 1 + 7 + 2 = 14.

All Possible Sum of Ranks for Boys are plotted, provides the probability distribution showing the exact sampling distribution of W, as shown in the figure next.



The sampling distribution of W is used to compute the p-value for the test just as we have done using other sampling distributions. If p value is less than level of significance (α), then Null Hypothesis is rejected.

The mean and standard deviation of W, are given by the following formulas n_1 and n_2 are sample sizes):

Mean: $\mu_w = \frac{n_1(n_1+n_2+1)}{2}$

Standard Deviation: $\sigma_w = \sqrt{ \frac{n_1n_2(n_1+n_2+1)}{12} }$

Some applications of the MWW test make it appropriate to assume that the two populations have identical shapes and if the populations differ, it is only by a shift in the location of the distributions. If the two populations have the same shape, the hypothesis test may be stated in terms of the difference between the two population medians.

Any difference between the medians can be interpreted as the shift in location of one population compared to the other. In this case, the three forms of the MWW test about the medians of the two populations are as follows:

Two-Tailed Test	**Lower Tail Test**	**Upper Tail Test**
$H_0: \text{Median}_1 - \text{Median}_2 = 0$	$H_0: \text{Median}_1 - \text{Median}_2 \geq 0$	$H_a: \text{Median}_1 - \text{Median}_2 \leq 0$
$H_a: \text{Median}_1 - \text{Median}_2 \neq 0$	$H_0: \text{Median}_1 - \text{Median}_2 < 0$	$H_a: \text{Median}_1 - \text{Median}_2 > 0$

3. Kruskal-Wallis Test

The Kruskal-Wallis test (a nonparametric test) extends the MWW test to 3 or more populations. It is an alternative to the parametric analysis of variance test for the differences among the means of three or more normally distributed populations.

The Kruskal-Wallis test does not require any assumption about the distribution of the populations and uses the null hypothesis that the k populations are identical. If the assumption can be made that the populations have the same shape, the test provides an inference about differences among the medians of the k populations.

Let us understand with an example.

Patanjali hires management graduates from 3 colleges (A, B and C). The Hiring Manager is interested in finding out, whether the three populations of managers are identical in terms of performance ratings.

Performance rating data are available for independent samples of 7 managers who graduated from college A, 6 managers who graduated from college B, and 7 managers who graduated from college C. We will use a 0.05 level of significance for the test.

The form of the null and alternative hypotheses is as follows:
H_0 : Populations of all 3 colleges are identical in performance
H_a : Populations of all 3 colleges are not identical in performance

The performance rating data and their assigned ranks are shown in Figure below. Note that we assigned the average ranks to tied performance ratings of 60, 70, 80, and 90. The Figure also shows the sum of ranks for each of the three samples.

Colleges	College A	Rank	College B	Rank	College C	Rank
	25	3	60	9	50	7
	70	12	20	2	70	12
	60	9	30	4	60	9
	85	17	15	1	80	15.5
	95	20	40	6	90	18.5
	90	18.5	35	5	70	12
	80	15.5	-	-	75	14
Sum of Ranks		95		27		88

The Kruskal-Wallis test statistic uses the sum of the ranks for the three samples and is computed as follows.

H = \(\frac{12}{n_T(n_T+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n_T+1) \)

Where

k = number of populations

n_i = the number of observations in sample i

n_T =\(\sum_{i=1}^k n_i \)

R_i = The sum of the ranks for sample i

Kruskal and Wallis were able to show that, under the null hypothesis assumption of identical populations, the sampling distribution of H can be approximated by a chi-square distribution with (k-1) degrees of freedom.

This approximation is generally acceptable if the sample sizes for each of the k populations are all greater than or equal to 5. The null hypothesis of identical populations will be rejected if the test statistic H is large. As a result, the Kruskal-Wallis test is always expressed as an **upper tail test.**

In our example, n_1 = 7, n_2 = 6, n_3 = 7 and n_T = 7 + 6 + 7 = 20

H Statistic is calculated as below:

H =$$ \frac{12}{n_T(n_T+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n_T+1) $$ = $$ \frac{12}{20(20+1)} [\frac{95^2}{7}+ \frac{27^2}{6}+ \frac{88^2}{7}] - 3(20+1) $$ = 8.92

We can now use the chi-square distribution table to determine the p-value for the test, which comes out to be p = 0.0116 (corresponding to $$ \chi^2 $$ =8.92). Since p value is less than Level of significance, we reject Null Hypothesis and Conclude that Performance Rating of 3 colleges are not identical.

Further, Because, the sum of the ranks is relatively low for the sample of managers who graduated from college B, it would be reasonable for the company to either reduce its recruiting from college B, or at least evaluate the college B graduates more thoroughly before making a hiring decision.

---

4. Rank Correlation Test

We know that, the Pearson product moment correlation coefficient is a measure of the linear association between two variables using quantitative data. In this section, we provide a correlation measure of association between two variables when ordinal or rank-ordered data are available. The Spearman rank-correlation coefficient has been developed for this purpose.

Rank Correlation Coefficient measures the degree of similarity between two rankings and can be used to assess the significance of relation between them.

Lizzat Papad hired 10 sales managers last year. At the time of hiring, each manager was allocated a score based on his performance in the selection process and thus was given a rank from 1 to 10.

Now after 1 year, each of 10 manager was given a rank based on performance in Sales. The Hiring Manager wants to determine whether individuals who had a greater potential at the time of interview turn out to have higher sales records. Based on sales performance, each employee is ranked 1 to 10.

First we compute the Spearman rank-correlation coefficient using following equation.

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where:
n = the number of observations in the sample
$d_i = x_i - y_i$
x_i = the rank of observation i with respect to the first variable
y_i = the rank of observation i with respect to the second variable

Using the equation for r_s,we calculate Spearman rank-correlation coefficient, as shown in the Figure below.
$r_s = 0.733$

Salesperson	x_i (Ranking of Potential)	y_i (Ranking of Sales Performance)	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	-5	25
E	6	7	-1	1
R	3	4	-1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	-1	1
J	5	2	3	9
Total				44

Thus, $\sum d_i^2 = 44$

$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6(44)}{10(100-1)} = 0.733$

Please note that, the Spearman rank-correlation coefficient ranges from -1.0 to +1.0 and its interpretation is similar to the Pearson product moment correlation coefficient for quantitative data. The Spearman rank-correlation coefficient provides the same value that is obtained by using the Pearson product moment correlation coefficient procedure with the rank-ordered data.

In both cases, a rank-correlation coefficient near +1.0 indicates a strong positive association between the ranks for the two variables, while a rank-correlation coefficient near -1.0 indicates a strong negative association between the ranks for the two

variables. A rank-correlation coefficient of 0 indicates no association between the ranks for the two variables.

But, we are interested in using the **sample rank correlation (r_s)**, to make an inference about the **population rank correlation coefficient (ρ_s)**

To do this, construct following Null and Alternate Hypothesis.
H_0: $\rho_s = 0$
H_a: $\rho_s \neq 0$

The sampling distribution of r_s is approximated by a normal distribution, So the standard normal random variable z becomes the test statistic, with the following formula:

$$ z = \frac{r_s- \mu_{r_s}}{ \sigma_{r_s}} $$

Where:

Mean of sample, $\mu_{r_s} = 0$
Standard Deviation of sample, $\sigma_{r_s} = \sqrt{ \frac{1}{n-1} }$
Using the equation, we get z = 2.20

Using the standard normal probability table and z =2.20, we find the two-tailed p-value = 0.0278. With a 0 .05 level of significance, we see that p-value ≤ α. Thus, we reject the null hypothesis that the population rank-correlation coefficient is 0.

Thus, the test result shows that there is a significant rank correlation between potential at the time of hiring and actual sales performance.

---

1. Introduction

Surveys serve as structured tools to gather information from a wide population, employing standardized questions for collecting data. They are distributed through various channels like online platforms, emails, or phone calls, covering diverse topics and primarily useful for quantitative data collection.

A survey, defined as a method of collecting primary data, communicates with a representative sample of individuals, providing a snapshot of a particular time. This type of survey relies on respondents' answers presumed to represent the larger target population.

Survey participation involves direct engagement by research respondents, either through questionnaires or interacting with interviewers, making it obtrusive compared to unobtrusive methods where subjects are unaware of data collection, as seen in disguised observation.

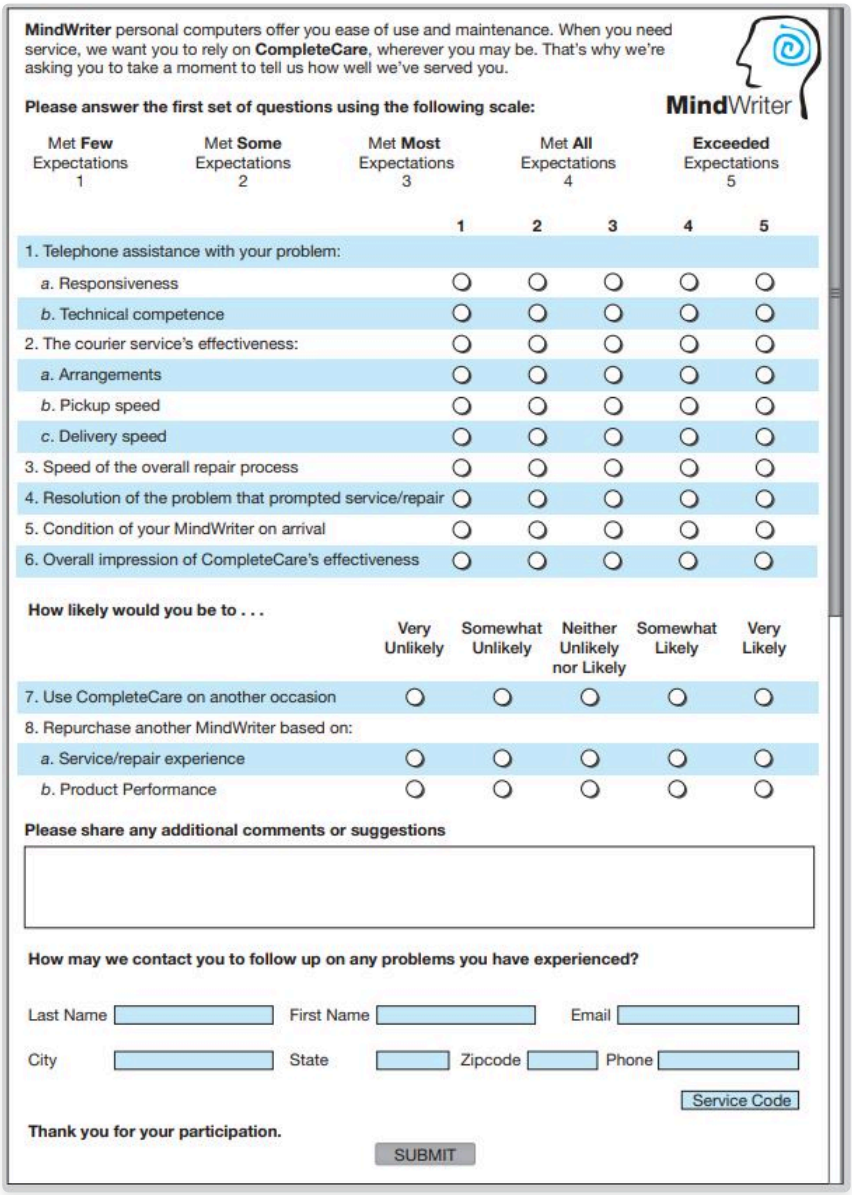
Surveys can be conducted with or without an administrator/interviewer. For instance, placing questionnaires in stores/malls or using mail or fax where respondents read and answer questions independently, termed self-administered questionnaires.

The survey method uses structured **questionnaires** administered to a sample of the target population, inquiring about behavior, intentions, attitudes, awareness, motivations, demographics, and lifestyle through verbal, written, or digital means. The 'structured' aspect implies standardized data collection via a formal questionnaire with questions asked in a predetermined sequence, making the project's purpose evident to participants. This structured direct survey, involving questionnaire administration, stands as the most common data-collection method.

---

2. Questionnaire Design

A questionnaire is a tool containing a series of written questions intended to gather information or opinions from respondents. It can function within a survey or be distributed independently to collect specific information. Questionnaires follow a standardized format and enable quantitative data analysis.



MindWriter personal computers offer you ease of use and maintenance. When you need service, we want you to rely on **CompleteCare**, wherever you may be. That's why we're asking you to take a moment to tell us how well we've served you.

MindWriter

Please answer the first set of questions using the following scale:

Met **Few** Expectations 1	Met **Some** Expectations 2	Met **Most** Expectations 3	Met **All** Expectations 4	**Exceeded** Expectations 5

	1	2	3	4	5
1. Telephone assistance with your problem:					
a. Responsiveness	○	○	○	○	○
b. Technical competence	○	○	○	○	○
2. The courier service's effectiveness:	○	○	○	○	○
a. Arrangements	○	○	○	○	○
b. Pickup speed	○	○	○	○	○
c. Delivery speed	○	○	○	○	○
3. Speed of the overall repair process	○	○	○	○	○
4. Resolution of the problem that prompted service/repair	○	○	○	○	○
5. Condition of your MindWriter on arrival	○	○	○	○	○
6. Overall impression of CompleteCare's effectiveness	○	○	○	○	○

How likely would you be to . . .

	Very Unlikely	Somewhat Unlikely	Neither Unlikely nor Likely	Somewhat Likely	Very Likely
7. Use CompleteCare on another occasion	○	○	○	○	○
8. Repurchase another MindWriter based on:					
a. Service/repair experience	○	○	○	○	○
b. Product Performance	○	○	○	○	○

Please share any additional comments or suggestions

How may we contact you to follow up on any problems you have experienced?

Last Name First Name Email

City State Zipcode Phone

Service Code

Thank you for your participation.

SUBMIT

Questionnaire Design Process

The process of designing a questionnaire is built upon the generation of information that effectively aids decision makers. The following are the steps involved in crafting a questionnaire:

1. Defining the Required Information

The initial stage of questionnaire design involves specifying the necessary information. This aligns with the initial phase of research design. It's beneficial to review the research problem components, approach, particularly research questions, hypotheses, and factors influencing research design.

2. Determining the Interview Method

The choice of interview method significantly influences questionnaire design and question content. *Self-administered surveys* like online and postal questionnaires require straightforward questions and clear instructions. In contrast, *face-to-face interviews* allow for complex and varied questions, while telephone interviews restrict questions to be brief and simple.

3. Crafting Individual Question Content

This phase involves addressing various considerations such as:

- Is the question essential?
- Could multiple questions be necessary instead of one?

Once necessity is established, ensuring questions are sufficient to gather desired information is vital. Multiple questions may be required at times for unambiguous data collection. *Double-barrelled question* is a single question that attempts to cover two issues. Such questions can be confusing to respondents and result in ambiguous responses.

4. Addressing Participants' Limitations and Willingness to Respond

Certain factors limit participants' ability to provide information, including lack of information, memory issues, or inability to articulate certain responses. Considerations include:

- Is the participant informed? (Filter questions can screen uninformed participants.)
- Can the participant recall? (Errors like *omission* (the inability to recall an event that actually took place), *telescoping* (when an individual telescopes or compresses time by remembering an event as occurring more recently than it actually occurred), and *creation* (when a participant 'remembers' an event that did not actually occur) may occur.)
- Can the participant articulate their thoughts and feelings?
- What level of effort is required from participants?
- Are participants willing to share sensitive information?

5. Selecting Question Structure

Questions can be unstructured or structured:

- *Unstructured questions* are open-ended, allowing participants to respond in their own words.
- *Structured questions* specify response alternatives and format, such as multiple-choice, dichotomous (yes/no), or scales.

Several guidelines can be employed to create an effective questionnaire:

- Incorporate counterbiasing statements, which serve as introductory remarks to potentially sensitive questions, reducing reluctance among respondents by suggesting that certain behaviors are common.
- Implement filter questions, which screen out respondents who are not qualified to answer a subsequent question.
- Utilize pivot questions, a form of filter question, to gather information like income or other sensitive data that respondents might be hesitant to provide.
- Employ the split ballot technique, which involves presenting two alternative phrasings of the same question to different halves of the sample, yielding a more accurate overall response compared to a single phrasing.

6. Choosing Question Wording

Question wording translates the desired content and structure into understandable terms for participants. Poorly worded questions can lead to refusal or inaccurate responses. Guidelines include defining the issue, using clear and ordinary words, avoiding implicit assumptions, and steering clear of leading or loaded questions.

A leading question is one that clues the participant to what answer is desired or leads the participant to answer in a certain way. Some participants have a tendency to agree with whatever way the question is leading them to answer. This tendency is known as yea-saying and results in a bias called acquiescence bias. A loaded question suggests a socially desirable answer or is emotionally charged.

7. Sequencing Questions Appropriately

Question order is as crucial as wording. Factors to consider:

- Opening questions should be engaging and simple.
- Sequence typically follows basic information, classification, and identification data.
- Difficult or sensitive questions are better placed later.
- Transition from general to specific questions is recommended.
- Logical order is essential, and branching questions should be well-structured.

8. Determining Form and Layout

Formatting, spacing, and positioning of questions impact results, especially in self-administered surveys. Numbering questions aids in coding responses, and pre-coding is useful for postal surveys. Pre-coding involves assigning a code to every conceivable response before data collection.

9. Preparing the Questionnaire for Publication

Guidelines for questionnaire publication include making it visually appealing, placing instructions near questions, considering booklet formats for longer surveys, and using pre-coded questionnaires when applicable.

10. Quality Assurance via Pilot Testing

Pilot testing is essential for identifying and resolving potential issues in the questionnaire. Focus on testing all aspects, analyzing pilot-test data, and ensuring that participants in the pilot-test mirror the actual survey population.

---

2. Questionnaire Design

While evaluating the quality of a survey, the researcher must estimate its accuracy. There are different types of error that can take place during a survey. These are similar to the error types that can occur in any research design or experimental methods.

1. Random Sampling Error

Most surveys try to portray a representative cross-section of a particular target population. Even with technically proper random probability samples, however, statistical errors will occur because of chance variation in the elements selected for the sample. Such errors are called as random sampling errors.

2. Non Sampling/ systematic error

These result from some imperfect aspect of the research design or from a mistake in the execution of the research. These errors include all sources of error other than those introduced by the random sampling procedure.

Non Sampling Errors can further be categorized into Respondent Error and Administrative Errors.

A *non-response error* arises when some of the respondents included in the sample do not respond. The primary causes of non-response are refusals and not-at-homes. Non-response will cause the net or resulting sample to be different in size or composition from the original sample. Non-response error is defined as the variation between the true mean value of the variable in the original sample and the true mean value in the net sample.

Response error arises when respondents give inaccurate answers or their answers are mis-recorded or mis-analysed. Response error is defined as the variation between the true mean value of the variable in the net sample and the observed mean value obtained in the research project. Response errors can be made by researchers, interviewers or respondents.

Errors made by Researcher

Errors made by the researcher include:

- *Surrogate information error* may be defined as the variation between the information needed for the research problem and the information sought by the researcher. For example, instead of obtaining information on consumer choice of a new brand, the researcher obtains information on consumer preferences because the choice process cannot be easily observed.
- *Measurement error* may be defined as the variation between the information sought and information generated by the measurement process employed by the researcher. While seeking to measure consumer preferences, the researcher employs a scale that measures perceptions rather than preferences.
- *Population definition error* may be defined as the variation between the actual population relevant to the problem at hand and the population as defined by the researcher.
- *Sampling frame error* may be defined as the variation between the population defined by the researcher and the population as implied by the sampling frame (list) used. For example, the telephone directory used to generate a list of telephone numbers does not accurately represent the population of potential consumers due to unlisted, disconnected and new numbers in service.
- *Data analysis error* encompasses errors that occur while raw data from questionnaires are transformed into research findings. For example, an inappropriate statistical procedure is used, resulting in incorrect interpretation and findings.
- The *experimenter effect* refers to unintentional bias or influence exerted by the researcher on the participants or the experiment's outcome due to their actions, expectations, or mannerisms.

Response errors made by Interviewer

Response errors made by the interviewer include respondent selection, questioning, recording and cheating errors.

- *Respondent selection error* occurs when interviewers select respondents other than those specified by the sampling design or in a manner inconsistent with the sampling design. For example, in a readership survey, a non-reader is selected for the interview but classified as a reader of The Time of India in the 19–25 year-old category in order to meet a difficult quota requirement.
- *Questioning error* denotes errors made in asking questions of the respondents or in not probing, when more information is needed. For example, while asking questions an interviewer does not use the exact wording given in the questionnaire.
- *Recording error* arises due to errors in hearing, interpreting and recording the answers given by the respondents. For example, a respondent indicates a neutral response (undecided) but the interviewer misinterprets that to mean a positive response (would buy the new brand).
- *Cheating error* arises when the interviewer fabricates answers to a part or the whole of the interview. For example, an interviewer does not ask the sensitive questions related to a respondent's debt but later fills in the answers based on personal

assessment.

Response errors made by Respondent

Response errors made by the respondent comprise inability and unwillingness errors.

- *Inability error* results from the respondent's inability to provide accurate answers. Respondents may provide inaccurate answers because of unfamiliarity, fatigue, boredom, faulty recall, question format, question content and other factors. For example, a respondent cannot recall the brand of toothpaste purchased four weeks ago.
- *Unwillingness error* arises from the respondent's unwillingness to provide accurate information. Respondents may intentionally misreport their answers because of a desire to provide socially acceptable answers, to avoid embarrassment, or to please the interviewer. For example, to impress the interviewer, a respondent intentionally says that they read The Economist magazine.
- *Self-consciousness effect* refers to a situation where a person's awareness of being observed or studied alters their behavior or responses. This phenomenon influences how individuals act or react in a given situation due to their self-awareness, potentially leading to modified or distorted behavior, especially in research or observational settings.

Other Errors

Some of other key terms are:

- *Voluntary response bias:* the sampling bias that often occurs when the sample is volunteers.
- *Self-interest study:* bias that can occur when the researchers have an interest in the outcome.
- *Perceived lack of anonymity:* when the responder fears giving an honest answer might negatively affect them.
- *Loaded questions:* when the question wording influences the responses.
- *Satisficing:* Satisficing occurs when respondents use a suboptimal amount of cognitive effort to answer questions. Instead, satisficers will typically pick what they consider to be the first acceptable response alternative.
- *Acquiescence Bias:* When presented with agree/disagree, yes/no, or true/false statements, some respondents are more likely to concur with the statement independent of its substance.
- *Social Desirability:* Social desirability occurs when respondents answer questions in a manner they feel will be positively perceived by others.
- *Response Order Bias:* Response order bias is the tendency to select the items toward the beginning (i.e., primacy effect) or the end (i.e., recency effect) of an answer list or scale.
- *Question Order Bias:* Order effects also apply to the order of the questions in surveys. Each question in a survey has the potential to bias each subsequent question by priming respondents.

---

1. Introduction

In the dynamic landscape of business and management, effective communication of statistical analyses is a crucial skill for decision-makers. The process of transforming statistical findings into meaningful insights requires the ability to craft clear and concise reports.

Managerial decisions often hinge on accurate data interpretation, and statistical analyses play a pivotal role in providing the necessary insights. A well-written statistical report ensures that these insights are not only accurate but also accessible to a diverse audience, ranging from executives with a non-technical background to data analysts well-versed in statistical methods.

Statistics, as a tool, enables managers to make informed decisions based on data-driven evidence. A well-constructed statistical report serves as a bridge between raw data and managerial decision-making. It allows decision-makers to understand the nuances of the data, draw meaningful conclusions, and implement strategies that align with organizational goals.

In the subsequent sections, we will delve into the specific elements that constitute a well-structured statistical report. Understanding the foundational concepts outlined in this chapter will pave the way for constructing reports that resonate with the managerial audience, fostering effective decision-making based on robust statistical analyses.

---

2. Structure of A report

A well-structured statistical report is crucial for effectively communicating research findings and insights. The structure not only enhances readability but also guides the reader through the logical flow of information. In managerial statistics, where precision and clarity are paramount, a thoughtfully organized report ensures that decision-makers can readily grasp the implications of statistical analyses.

Here is a recommended structure for a statistical report:

1. Title Page
Clearly state the title of the report, providing a concise yet informative description of the research. Include the names of the authors, their affiliations, and the date of publication. Mention any relevant institutional or departmental affiliations.

2. Abstract
Offer a brief summary of the report, providing an overview of the research objectives, methodology, key findings, and conclusions. Aim for clarity and conciseness, allowing readers to quickly assess the report's relevance to their interests.

3. Table of Contents
Provide a detailed table of contents to guide readers through the report's structure. Clearly list sections, subsections, figures, tables, and appendices, along with corresponding page numbers.

4. Introduction
Introduce the research problem or question that motivated the study. Clearly state the objectives and the significance of the research. Provide a brief overview of the methodology employed.

5. Literature Review (if applicable)
Review relevant literature that informs the research. Highlight gaps in existing knowledge and explain how the current study contributes to the field.

6. Methodology
Detail the research design, including the sampling method, data collection procedures, and statistical techniques employed. Clearly describe any statistical software or tools used for analysis.

7. Data Presentation
Present the data in an organized and understandable manner. Use visualizations, tables, and graphs to illustrate key patterns and trends in the data. Provide detailed descriptions and interpretations of the presented data.

8. Statistical Analysis
Conduct a thorough statistical analysis, explaining the rationale behind chosen methods. Present statistical tests, models, or algorithms used, along with relevant parameters and results.

9. Results
Summarize the key results of the statistical analysis. Use clear and concise language to convey the findings, referencing visual elements when necessary.

10. Discussion
Interpret the results within the context of the research question and objectives. Discuss implications, limitations, and potential areas for future research.

11. Conclusion
Provide a concise summary of the main findings and their implications. Revisit the research objectives and highlight the contributions of the study.

12. References
Cite all sources, including literature, data sources, and statistical methods. Follow a consistent citation style (APA, MLA, Chicago, etc.).

13. Appendices (if necessary)
Include supplementary materials, such as detailed data tables, additional charts, or complex mathematical derivations. By adhering to this structured approach, a statistical report becomes a cohesive and accessible document that supports effective

decision-making. Each section plays a crucial role in guiding the reader through the research process, from the introduction of the problem to the presentation and interpretation of statistical findings.

---

3. Writing Style and Clarity

Effective communication is paramount in Report Writing. A statistical report can house the most robust analyses, but without clear and coherent writing, the message may be lost. This section delves into the key elements of writing style and clarity in statistical reporting:

Precision in Language
Utilize precise and unambiguous language to convey ideas. Avoid unnecessary jargon or overly complex terms that may hinder understanding.

Clarity of Expression
Craft sentences that are clear, concise, and to the point. Clearly articulate the purpose of each section to guide the reader through the report.

Consistent Terminology
Maintain consistency in the use of terminology throughout the report. Define any specialized terms to ensure a shared understanding with the audience.

Logical Flow
Structure the report with a logical flow of information. Ensure that each section seamlessly connects to the next, facilitating a smooth reading experience.

Transitions
Use effective transitions between paragraphs and sections to enhance coherence. Guide the reader through the progression of ideas, methods, results, and conclusions.

Active Voice
Prefer the use of the active voice to enhance clarity and directness. Clearly attribute actions to specific actors or variables in the study.

Avoid Redundancy
Eliminate unnecessary repetition of information. Ensure that each sentence contributes meaningfully to the overall narrative.

Engage the Reader
Craft engaging introductions and conclusions to capture and maintain the reader's interest. Use descriptive language where applicable to create a vivid and compelling narrative.

Visual Clarity
Accompany statistical analyses with clear visualizations, such as charts and graphs. Ensure that visual elements are appropriately labeled and explained in the text.

Proofreading
Conduct thorough proofreading to catch grammatical errors, typos, and inconsistencies. Consider seeking feedback from colleagues to gain insights into the clarity of the writing.

Audience Consideration
Tailor the writing style to the target audience, whether it be executives, colleagues, or a broader readership. Clarify complex concepts for those who may not have an extensive statistical background.

Conciseness
Strive for brevity without sacrificing clarity. Eliminate unnecessary words or sentences to keep the report focused and impactful.

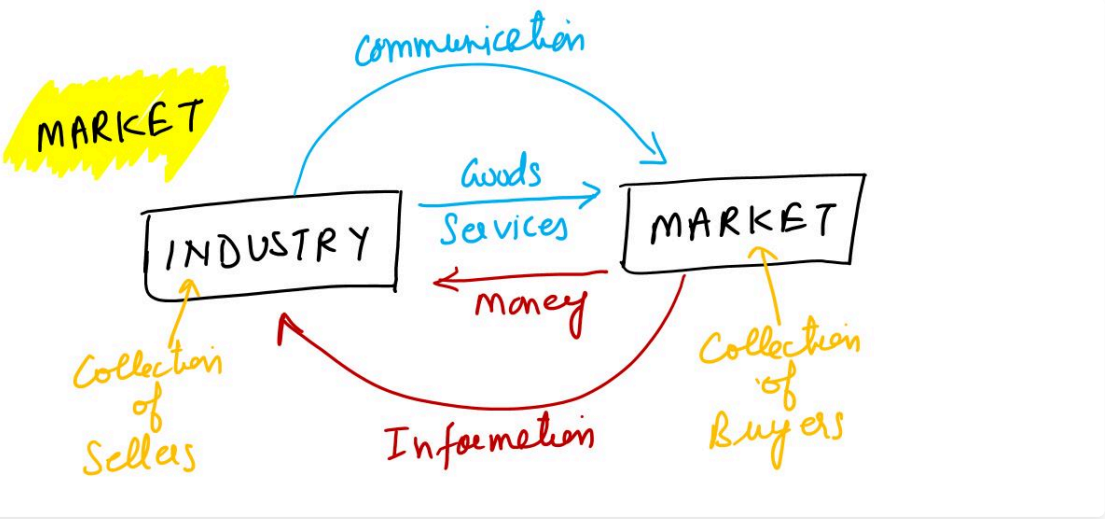
Accessibility
Ensure that the report is accessible to readers with varying levels of statistical expertise. Use footnotes or appendices for in-depth explanations without cluttering the main text.

By prioritizing clarity and adopting a reader-centric approach to writing, a statistical report becomes a powerful tool for conveying complex information in an understandable and impactful manner. Effective communication enhances the utility of statistical insights, facilitating informed decision-making in managerial contexts.

---

1. Market

Traditionally, *market* was a physical place where buyers and sellers gathered to exchange goods. Now marketers view the sellers as the industry and the buyers as the market. The sellers send goods and services and communications (ads, direct mail, e-mail messages) to the market; in return they receive money and information (attitudes, sales data).



We can distinguish between a Marketplace and a Marketspace. **Marketplace** is physical, as when one goes shopping in a store. **Marketspace** is digital, as when one goes shopping on the Internet.

Metamarket, a concept proposed by Mohan Sawhney, describes a cluster of complementary products and services that are closely related in the minds of consumers but are spread across a diverse set of industries. For example, the automobile Metamarket consists of automobile manufacturers, new and used car dealers, financing companies, insurance companies, mechanics, spare parts dealers, service shops, auto magazines, classified auto ads in newspapers, and auto sites on the Internet.

---

1. Market

Customer markets represent different segments of consumers or organizations that purchase goods or services. Here are the four main types:

1. Consumer Markets

Consumer markets consist of individuals or households who purchase goods and services for personal use or consumption. Examples are Buying groceries, clothing, electronics, or entertainment services.

Consumer markets are vast and diverse, with purchases driven by personal preferences, needs, and desires. Marketers often employ targeted advertising and promotions to appeal to specific consumer segments.

2. Business Markets

Business markets involve organizations that purchase goods and services for use in their operations or for resale. Examples are Procuring raw materials, machinery, office supplies, or services like consulting or IT support.

Business markets are characterized by larger transaction sizes, longer sales cycles, and a focus on value and efficiency. Purchase decisions often involve multiple stakeholders and may be based on factors such as quality, price, and supplier reputation.

3. Global Markets

Global markets refer to transactions and exchanges that occur across national borders, involving buyers and sellers from different countries. Examples are Exporting goods to foreign markets, importing products from overseas, or participating in international trade agreements.

Global markets offer opportunities for businesses to expand their reach and access new customer bases. They require an understanding of international regulations, cultural differences, and market dynamics to navigate successfully.

4. Nonprofit and Governmental Markets

Nonprofit and governmental markets include organizations that do not aim to generate profits or operate for public welfare and governance purposes. Examples are Government agencies purchasing equipment or services, nonprofit organizations acquiring resources for charitable activities or community services.

Nonprofit and governmental markets often involve unique procurement processes, compliance requirements, and funding sources. Suppliers may need to align with the mission and values of these organizations to secure contracts.

---

2. What is Marketed

Marketers market 10 main types of entities: goods, services, events, experiences, persons, places, properties, organizations, information, and ideas.

Goods

Goods refer to tangible, physical products that are produced and marketed. The marketing efforts for goods typically focus on highlighting product features, quality, and value for money.

Example: Maruti Suzuki in India extensively markets its range of cars, emphasizing factors like fuel efficiency, safety features, and innovative technology to appeal to different consumer segments.

Services

Services are intangible offerings that fulfill a customer's needs. Marketing services involves promoting the benefits, reliability, and unique features of the service to potential customers.

Example: Airtel, one of India's leading telecom service providers, uses marketing to showcase the reliability of its network, innovative service packages, and customer support to attract and retain subscribers.

Events

Event marketing involves promoting time-based occurrences, such as conferences, festivals, or trade shows. The goal is to attract attendees, sponsors, and media attention.

Example: The Indian Premier League (IPL) is a cricket league that heavily relies on marketing to attract a global audience, sponsors, and advertisers. The event is marketed for its entertainment value and as a platform for brand visibility.

Experiences

Experiential marketing involves creating and promoting a memorable, holistic experience for consumers by combining goods and services. The focus is on the overall customer journey and emotional connection.

Example: FabIndia, a retail brand in India, markets an experience by offering a range of ethnic products along with the ambiance of its stores. The marketing emphasizes a connection to Indian traditions and crafts.

Persons

Personal branding involves marketing individuals, typically public figures, to enhance their image and influence. Marketers work on strategies to manage perceptions and build a positive association.

Example: Virat Kohli, the captain of the Indian cricket team, collaborates with various brands. Marketers assist him in endorsing products, managing social media presence, and building a personal brand that extends beyond cricket.

Places

Place marketing involves promoting a location to attract residents, businesses, and tourists. It focuses on showcasing the unique features and advantages of a place.

Example: Kerala Tourism engages in marketing campaigns that highlight the state's natural beauty, cultural heritage, and Ayurvedic traditions to attract tourists and position itself as a must-visit destination.

Properties

Property marketing involves promoting real estate, either physical (real property) or financial (stocks and bonds). The goal is to attract buyers, investors, or tenants.

Example: DLF, a real estate developer in India, markets its residential and commercial properties by emphasizing factors such as location, infrastructure, and amenities to potential buyers and investors.

Organizations

Organizational marketing involves promoting the overall image and reputation of a company or institution. It encompasses activities to enhance public perception and attract stakeholders.

Example: Infosys, an IT services company in India, utilizes marketing to communicate its innovation, global presence, and corporate social responsibility initiatives. This helps in attracting clients, investors, and top talent.

Information

Information refers to knowledge or data that is produced, packaged, and distributed for educational purposes. This includes textbooks, educational materials, and online resources created by schools and universities.

Example: NCERT (National Council of Educational Research and Training) in India produces and markets textbooks for school students. These textbooks contain essential information aligned with the national curriculum and are distributed to students and schools across the country.

Ideas

Ideas are fundamental to any market offering. They represent the conceptual foundation or core message behind a product or service. The idea helps shape the product's identity and value proposition.

Example: Apple Inc.'s marketing is often centered around the idea of innovation and user-friendly technology. The basic idea is to offer products that are not just technologically advanced but also seamlessly integrated into users' lives, creating a distinct market identity for the brand.

---

3. Marketing Management

Marketing is an organizational function and a set of processes for creating, communicating, and delivering value to customers and for managing customer relationships in ways that benefit the organization and its stakeholders.

In other words, it is the process by which companies engage customers, build strong customer relationships, and create customer value in order to capture value from customers in return.

Peter Drucker, a leading management theorist, said that "the aim of marketing is to make selling superfluous. The aim of marketing is to know and understand the customer so well that the product or service fits him and sells itself. Ideally, marketing should result in a customer who is ready to buy."

Marketing management is the art and science of choosing target markets and building profitable relationships with them. The marketing manager's aim is to engage, keep, and grow target customers by creating, delivering, and communicating superior customer value.

The marketing manager must answer two important questions: What customers will we serve (*what's our target market*)? And How can we serve these customers best (*what's our value proposition*)?

Example

Let us illustrate the concept of Marketing using the example of a smartphone company.

Creating Value: Company focuses on creating value for customers by developing innovative smartphones with advanced features, sleek designs, and reliable performance. Through research and development, the company continuously enhances its products to meet evolving customer needs and preferences.

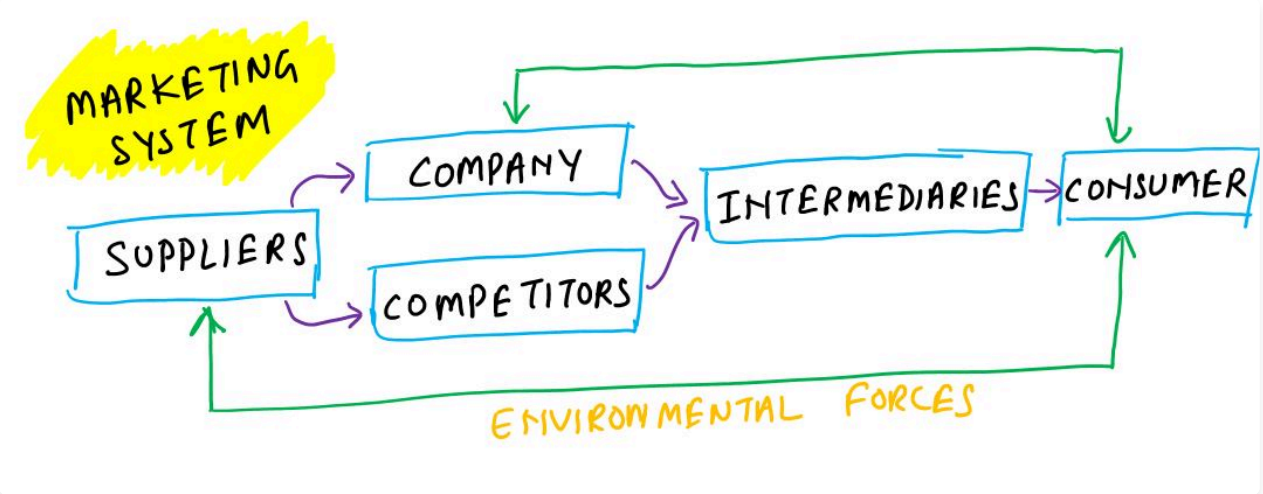
Communicating Value: Company effectively communicates the value of its smartphones through various marketing channels such as television commercials, online advertisements, and influencer partnerships. Engaging campaigns highlight the unique features, benefits, and user experiences offered by devices, compelling customers to consider purchasing.

Delivering Value: Company ensures the seamless delivery of value to customers through its distribution channels, retail partners, and online platforms. From the moment a customer purchases a smartphone to the post-purchase support provided through customer service channels, the company prioritizes delivering an exceptional user experience.

Managing Customer Relationships: Company understands the importance of managing customer relationships to foster loyalty and repeat business. Through personalized email newsletters, exclusive offers, and social media engagement, the company stays connected with customers, solicits feedback, and addresses concerns promptly to maintain positive relationships.

4. Marketing System

The figure shows the main elements in a marketing system.



Marketing involves serving a market of final consumers in the face of competitors. The company and competitors research the market and interact with consumers to understand their needs. Then they create and send their market offerings and messages to consumers, either directly or through marketing intermediaries.

Each party in the system is affected by major environmental forces (demographic, economic, natural, technological, political, and social/cultural).

Each party in the system adds value for the next level. The arrows represent relationships that must be developed and managed. Thus, a company's success at building profitable relationships depends not only on its own actions but also on how well the entire system serves the needs of final consumers.

Example

Let's use the example of a fast-food chain, "Bite Delight," to illustrate the concept of marketing.

Consumers

Bite Delight focuses on serving a market of final consumers, individuals who purchase its fast-food products for personal consumption. These consumers are the end-users of Bite Delight's offerings, ranging from burgers and fries to salads and beverages.

Competitors

Bite Delight conducts research and analysis to identify competitors within the fast-food industry. By studying competitors' offerings, pricing strategies, and marketing tactics, Bite Delight gains insights into the competitive landscape and adjusts its own strategies accordingly to stand out in the market.

Market Offerings

Based on consumer insights and competitive analysis, Bite Delight creates market offerings such as new menu items, promotional deals, and value-added services to meet the evolving needs and preferences of its target audience.

Messages

Bite Delight communicates with consumers through various marketing channels, including advertisements on television, social media platforms, and mobile apps. These messages highlight the quality, convenience, and affordability of Bite Delight's offerings, enticing consumers to visit its outlets or place orders for delivery.

Intermediaries

Bite Delight collaborates with marketing intermediaries such as food delivery services, advertising agencies, and franchise partners to extend its reach and amplify its marketing efforts. These intermediaries help Bite Delight connect with consumers effectively and drive sales growth.

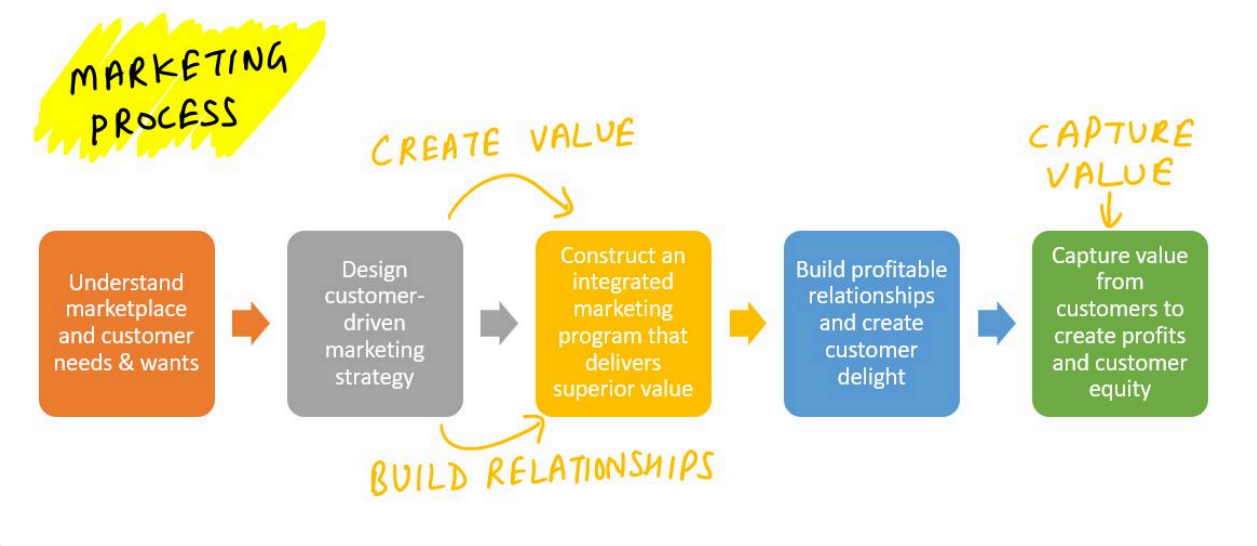
Environmental Forces

Bite Delight operates within a dynamic environment influenced by various factors, including demographic shifts, economic

trends, technological advancements, and social/cultural changes. By monitoring and adapting to these environmental forces, Bite Delight remains agile and responsive to market dynamics, ensuring sustained success in the competitive fast-food industry.

---

5. Process of Marketing



The diagram presents a simple, five-step model of the marketing process for creating and capturing customer value. In the first four steps, companies work to understand consumers, create customer value, and build strong customer relationships. In the final step, companies reap the rewards of creating superior customer value. By creating value for consumers, they in turn capture value from consumers in the form of sales, profits, and long-term customer equity.

1. Understand the marketplace and customer needs and wants

Successful marketing begins with a deep understanding of customer needs, wants, and demands. This involves grasping the core concepts of needs, wants, and demands, as well as developing a keen awareness of the market dynamics. The goal is to design market offerings that not only satisfy immediate desires but also build long-term relationships.

Example: An Indian fashion retailer invests in market research to understand the evolving preferences of its target audience. By staying attuned to changing fashion trends and consumer preferences, the company tailors its product offerings to meet the specific desires of its customers, fostering loyalty and repeated business.

2. Design a customer-driven marketing strategy

Crafting an effective marketing strategy involves making strategic decisions about which customer segments to target and how to position the company in the marketplace. The chosen market orientation (production, product, selling, marketing, or societal) guides these decisions, shaping the company's overall approach to meeting customer needs.

Example: An Indian mobile app developer adopts a marketing concept by focusing on understanding the needs and preferences of its target users. By consistently delivering apps that address specific user requirements and preferences, the company builds a loyal user base and differentiates itself in a competitive market.

3. Construct an integrated marketing program that delivers superior value

Transforming the marketing strategy into action involves developing an integrated marketing program. This program includes the marketing mix, encompassing product, price, place, and promotion. Each element of the marketing mix must work cohesively to deliver the intended value proposition to the target audience.

Example: A leading Indian soft drink manufacturer strategically combines product innovation, competitive pricing, an efficient distribution network, and engaging promotional campaigns. This integrated approach ensures that the brand consistently delivers superior value to consumers, contributing to its market success.

4. Build profitable relationships and create customer delight

Customer relationship management focuses on cultivating and maintaining profitable relationships by consistently delivering superior value and satisfaction. This involves engaging customers in brand experiences and conversations, ultimately leading to high customer equity.

Example: A prominent Indian tech company actively involves its users in shaping the features of its software products. By incorporating user feedback, providing excellent customer support, and offering personalized experiences, the company builds

strong relationships, fostering customer loyalty and advocacy.

5. Capture value from customers to create profits and customer equity

The final step involves capturing value, such as sales, market share, and profits, by creating and delivering superior customer value. The outcomes include customer loyalty, increased market share, and enhanced customer equity, contributing to the long-term success of the company.

Example: An Indian online streaming platform not only attracts subscribers with a vast content library but also retains them through personalized recommendations and exclusive content. This strategy not only drives profits but also increases the platform's share of each customer's entertainment spending over time.

---

6. Needs, Wants and Demands

Philip Kotler, defines marketing as "A human activity directed at satisfying needs and wants through exchange processes." Thus, the most fundamental concept for the basis of all marketing activities is the existence of human needs.

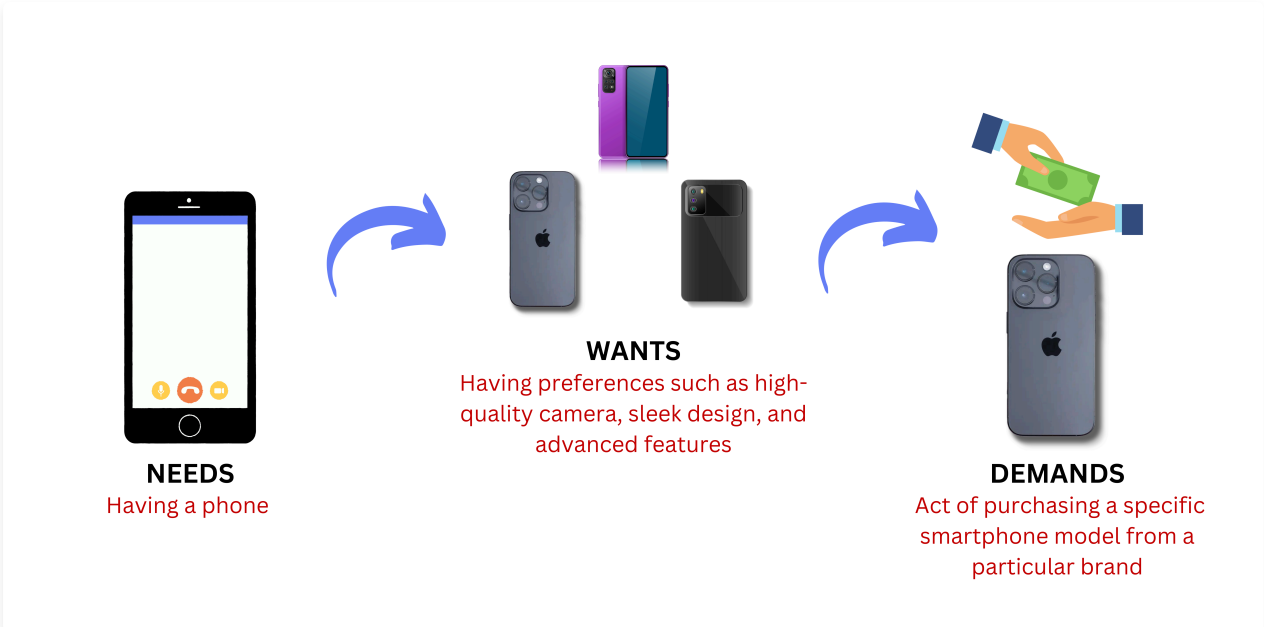
The human need is a state in which a person feels deprived of something. There are many human needs described in many ways. It is important to understand that at any time some needs in a human being are dormant and unsatisfied whereas others are active and are being satisfied. A marketing man may thus devise a product or service aimed at satisfying a certain dormant need and thus provide satisfaction to the user. This is why a man is often described as `*a bundle of dormant wants*'. The need exists but these have to be converted into 'wants' by a marketing strategy.

The successful marketer will try to understand the target market's needs, wants, and demands.

Needs describe basic human requirements such as food, air, water, clothing, and shelter. People also have strong needs for recreation, education, and entertainment.

These needs become **wants** when they are directed to specific objects (culture and personality of individual) that might satisfy the need. Average human needs food but wants a hamburger, french fries, and a soft drink.

Demands are wants for specific products backed by an ability to pay. Many people want a Mercedes; only a few are able and willing to buy one. Companies must measure not only how many people want their product, but also how many would actually be willing and able to buy it.



---

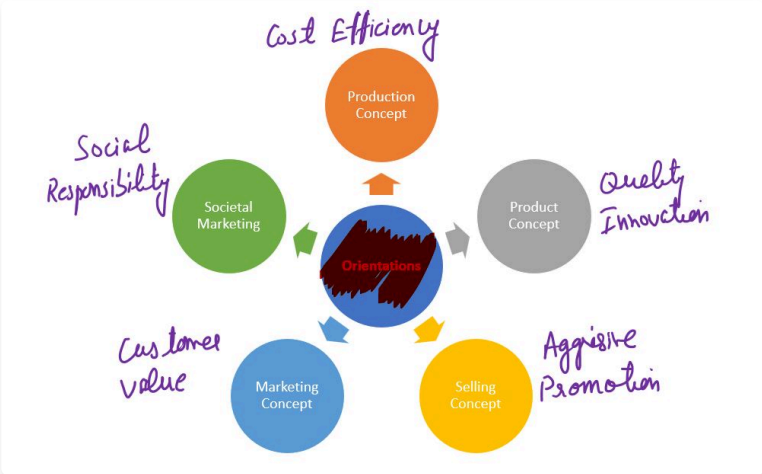
6. Needs, Wants and Demands

The marketers can distinguish 5 types of needs:

1. **Stated Needs:** Stated needs are the explicit and openly expressed requirements or desires that a customer communicates.
   Example: The customer wants a car.
2. **Real Needs:** Real needs go beyond the explicit statements and address the underlying requirements that customers may not express directly. It involves understanding the core necessity behind the stated need.
   Example: The customer wants a car whose operating cost, not initial price, is low.
3. **Unstated Needs:** Unstated needs are expectations or desires that customers have but do not explicitly communicate. These needs are often inferred through careful observation and understanding of customer behavior.
   Example: The customer expects good service from the dealer.
4. **Delight Needs:** Delight needs refer to additional features or services that, if provided, would exceed customer expectations and create a positive emotional response. These are not essential but add an extra layer of satisfaction.
   Example: The customer would like the dealer to include an onboard GPS navigation system.
5. **Secret Needs:** Secret needs are desires that customers may not openly admit but play a role in their decision-making. These needs are often tied to self-image, social status, or personal identity.
   Example: The customer wants friends to see him or her as a savvy consumer.

---

7. Marketing Orientations



There are five competing concepts that guide how organizations conduct their business: (i) the production concept, (ii) the product concept, (iii) the selling concept, (iv) the marketing concept, and (v) the societal marketing concept.

Let's examine each of these concepts individually.

---

7. Marketing Orientations

The production concept, rooted in traditional business practices, posits that consumers prioritize products that are readily available and priced affordably. Under this concept, managers of production-oriented businesses emphasize maximizing production efficiency, minimizing production costs, and achieving widespread distribution.

Henry Ford's implementation of the production concept revolutionized the automotive industry. By pioneering assembly line production methods, Ford was able to mass-produce the affordable Model T automobile, making car ownership accessible to the average consumer and transforming transportation worldwide.

IKEA, the Swedish furniture retailer, embodies the production concept in the retail sector. Through standardized designs, efficient manufacturing processes, and flat-packaging for easy transportation, IKEA delivers affordable furniture solutions to customers globally. By focusing on cost-effective production and streamlined distribution, IKEA maintains its position as a leader in the furniture industry while catering to consumer demand for affordable home furnishings.

---

7. Marketing Orientations

The product concept is based on the belief that consumers prioritize products that offer the highest quality, superior performance, or innovative features. Managers in organizations following the product concept focus on continuously improving their products over time, assuming that consumers can recognize and appreciate superior quality and performance.

Product-oriented companies often invest heavily in research and development to design and produce exceptional products without significant input from customers. They trust their engineers and designers to create innovative solutions that exceed consumer expectations.

However, solely relying on the product concept can lead to Marketing Myopia, a term coined by Theodore Levitt. This narrow focus on product excellence may result in overlooking broader consumer needs and market trends.

Apple exemplifies the product concept with its iPhone lineup. Each new iteration of the iPhone aims to deliver cutting-edge technology, sleek design, and innovative features. Apple's relentless pursuit of product excellence has cultivated a dedicated customer base willing to pay a premium for the latest iPhone model.

Tesla embodies the product concept in the automotive industry with its electric vehicles (EVs). Tesla's EVs are known for their exceptional performance, advanced technology, and sustainability features. By prioritizing innovation and quality, Tesla has disrupted the traditional automotive market and garnered widespread acclaim for its products.

---

7. Marketing Orientations

The selling concept revolves around the idea that consumers and businesses may not naturally purchase enough of a company's products unless the company actively engages in aggressive selling and promotion efforts. This concept assumes that consumers need to be persuaded or "coaxed" into making purchases, prompting companies to deploy extensive selling and promotional strategies to stimulate buying behavior.

Many companies adhere to the selling concept, particularly when marketing unsought goods—products that consumers typically do not actively seek out, such as insurance or burial plots. Industries selling such products often focus on locating potential customers and persuading them to recognize the benefits of the offering.

However, the selling concept carries inherent risks. It prioritizes short-term sales transactions over long-term customer relationships and profitability. By focusing solely on selling what the company produces, rather than addressing consumer needs and preferences, companies may fail to build sustainable customer loyalty. Additionally, assuming that customers coerced into buying will develop positive perceptions of the product is often unrealistic and can lead to disappointment and dissatisfaction.

Telemarketing firms frequently operate based on the selling concept. They engage in outbound sales calls and promotional efforts to persuade consumers to buy products or services they may not have initially considered purchasing. While these tactics may yield short-term sales, they often lead to customer frustration and negative perceptions of the brand, undermining long-term relationships and profitability.

Companies in the timeshare industry often rely heavily on the selling concept. They invest significant resources in aggressive sales tactics, such as high-pressure sales presentations and persuasive marketing campaigns, to convince consumers to purchase timeshare properties. These companies prioritize closing deals rather than focusing on the long-term satisfaction and retention of customers.

7. Marketing Orientations

The marketing concept represents a comprehensive approach to business that places the customer at the center of all organizational activities. It's founded on the principle that a company's success is contingent upon its ability to meet and exceed customer expectations by delivering superior value compared to competitors.

This concept rests on 4 key pillars:

(i) **Target Market:** Companies adopting the marketing concept identify specific groups of customers, known as target markets, whose needs and preferences align with the company's offerings. By focusing resources on these target markets, companies can tailor their marketing efforts to effectively reach and engage with the right audience.

(ii) **Customer Needs:** Understanding and addressing customer needs form the cornerstone of the marketing concept. Rather than solely relying on product features or functionalities, companies delve deep into customer insights to discern unmet needs, desires, and pain points. By aligning product development and marketing strategies with customer needs, companies can create offerings that resonate with their target market.

(iii) **Integrated Marketing:** The marketing concept emphasizes the integration of all marketing activities to deliver a cohesive and consistent message to customers. This holistic approach spans various channels and touchpoints, including advertising, promotions, sales, customer service, and product development. Integrated marketing ensures that every interaction with the customer reinforces the brand's value proposition and strengthens customer relationships.

(iv) **Profitability:** While customer satisfaction is paramount, the ultimate goal of the marketing concept is to drive profitability for the organization. By delivering superior value to customers, companies can command premium prices, capture market share, and generate sustainable profits over the long term. Profitability serves as a metric of the effectiveness of the marketing efforts in creating and retaining satisfied customers.

P&G, a global consumer goods company, exemplifies the marketing concept through its portfolio of household brands. From Tide detergent to Pampers diapers, P&G prioritizes consumer insights to develop products that address specific needs and deliver tangible benefits to customers. Through targeted advertising, product innovation, and brand-building initiatives, P&G maintains its market leadership by continuously adapting to evolving consumer preferences and market dynamics.

Nike exemplifies the marketing concept in the athletic footwear and apparel industry. Through extensive market research and branding efforts, Nike identifies and responds to consumer preferences for performance, style, and innovation. By delivering high-quality products and engaging marketing campaigns that resonate with its target audience, Nike establishes strong emotional connections with consumers, driving brand loyalty and market leadership.

---

7. Marketing Orientations

The Societal Marketing Concept advocates for organizations to not only meet the needs and desires of target markets but also to do so in a manner that preserves or enhances both consumer well-being and societal welfare. Unlike traditional marketing approaches that prioritize profit maximization and consumer satisfaction, the societal marketing concept emphasizes the integration of social and ethical considerations into marketing practices.

Under this concept, marketers are tasked with identifying and addressing the needs, wants, and interests of target markets while also taking into account broader societal implications. This involves balancing the often-conflicting objectives of company profitability, consumer satisfaction, and public interest. The goal is to deliver value to customers in a way that contributes positively to both individual consumers and society as a whole.

The societal marketing concept prompts marketers to question whether fulfilling immediate consumer wants aligns with long-term consumer welfare and societal well-being. It emphasizes the importance of sustainable marketing practices that not only meet present consumer needs but also safeguard the ability of future generations to fulfill their needs.

Patagonia, an outdoor apparel company, exemplifies the societal marketing concept through its commitment to environmental sustainability and social responsibility. The company not only produces high-quality outdoor gear but also prioritizes environmentally friendly manufacturing processes, ethical sourcing of materials, and advocacy for environmental conservation. By aligning its business practices with environmental stewardship, Patagonia not only attracts environmentally conscious consumers but also contributes to the broader goal of preserving the planet's natural resources.

TOMS Shoes is renowned for its one-for-one giving model, where for every pair of shoes purchased, the company donates a pair to a child in need. This philanthropic approach integrates social impact into TOMS' marketing strategy, appealing to consumers who seek to make a positive difference through their purchasing decisions. By addressing both consumer needs for footwear and societal needs for access to basic necessities, TOMS demonstrates the principles of the societal marketing concept in action.

Some companies practice a form of the societal marketing concept called **Cause-Related marketing**. This is a form of marketing in which a company tackles a social or environmental problem and create business value for the company at the same time. Typically, in cause-related marketing campaigns, a brand is affiliated with a cause and a portion of the proceeds from the sales of the brand is donated to the cause.

---

7. Marketing Orientations

Marketing myopia occurs when a company becomes excessively focused on its products rather than understanding the broader needs and desires of its customers. Instead of identifying and addressing the underlying benefits and experiences sought by consumers, the company fixates on promoting its own offerings, potentially overlooking opportunities for innovation and growth.

Consider a company that manufactures and sells traditional incandescent light bulbs. Despite advancements in lighting technology and shifting consumer preferences towards energy-efficient alternatives, the company remains steadfast in promoting its existing product line. It invests heavily in advertising campaigns highlighting the features of its incandescent bulbs, such as brightness and longevity, without considering the broader benefits and experiences desired by consumers.

Meanwhile, competitors in the lighting industry recognize the growing demand for energy-efficient solutions and pivot their focus towards developing and marketing LED bulbs. These competitors understand that consumers are not simply seeking light bulbs but are looking for cost-effective, eco-friendly lighting solutions that enhance their living spaces and reduce energy consumption.

As a result, the company entrenched in marketing myopia fails to adapt to changing market dynamics and experiences declining sales and relevance over time. In contrast, competitors who prioritize understanding and fulfilling customer needs thrive by offering innovative products that align with consumer preferences and deliver meaningful benefits and experiences.

---

8. New Marketing Realities

New Marketing Realities refer to the major changes in the marketplace that influence how marketing is practiced today. These are typically grouped under three broad categories: Technology, Globalization, and Social Responsibility. Let's explore each category:

1. Technology
Technology is transforming the way marketers interact with consumers.

- Digital Transformation: Shift from traditional to digital platforms (e.g., Tata Neu, Paytm, Zomato marketing online).
- Big Data & Analytics: Marketers now use consumer data to predict behavior and personalize campaigns.
- Artificial Intelligence & Automation: Chatbots, email automation, AI-driven customer segmentation (e.g., HDFC Bank uses AI for customer queries).
- Mobile Marketing: High mobile penetration in India has made mobile-first strategies essential.
- E-commerce & D2C: Rise of platforms like Flipkart, Meesho, and Nykaa has changed distribution and promotion models.

2. Globalization
Markets and competitors are now global.

- Global Brands in Local Markets: McDonald's or IKEA adapting products to Indian tastes.
- Cross-Border Competition: Indian brands like Mamaearth and boAt now compete globally.
- Global Supply Chains: Marketing must consider international pricing, logistics, and branding.
- Access to International Consumers: Indian MSMEs selling via Amazon Global.

3. Social Responsibility
Consumers expect ethical, transparent, and socially conscious behavior.

- Sustainability Marketing: Emphasis on eco-friendly packaging, green products (e.g., Tata Tea's "Jaago Re" campaign).
- Cause-Related Marketing: Brands supporting social causes (e.g., Surf Excel's "Daag Achhe Hain" aligns with learning through play).
- Corporate Social Responsibility (CSR): Statutory requirement in India influences brand image.
- Consumer Privacy & Data Protection: Companies must handle data ethically (especially after Indian Data Protection laws).

---

9. New Marketing World

Marketing is no longer about just selling products—it's about understanding changing consumers, adapting to dynamic organizations, and surviving in an evolving competitive landscape.

1. How the Consumer is Changing

Today's consumer is more informed, connected, and empowered than ever before.

- Digital-first mindset: Consumers now research online before purchasing—even in small towns.
  Example: Consumers check reviews on Amazon or Flipkart before buying products.
- Demand for personalisation: They expect products, ads, and experiences tailored to their preferences.
  Example: Netflix and Spotify curate personalized content based on viewing/listening history.
- Value-conscious & quality-focused: Consumers seek best value for money, not just lowest price.
  Example: Rise in demand for products like Mamaearth (natural ingredients) over cheaper chemical alternatives.
- Socially aware & ethical: They prefer brands aligned with environmental and social causes.
  Example: Tata Tea's "Jaago Re" or FabIndia's sustainable fashion.
- Connected 24x7: Consumers constantly interact with brands via Instagram, WhatsApp, YouTube, etc.

2. How Companies are Changing

To keep pace with changing consumer expectations and technology, companies are evolving rapidly.

- Customer-centric business models: More companies are using customer data to shape offerings.
  Example: Swiggy uses order patterns to improve delivery time and menu choices.
- Agile & tech-enabled: Firms are becoming flexible, digital-first, and data-driven.
  Example: Infosys and TCS offer digital transformation to clients worldwide.
- From Products to Experiences: Focus is shifting from selling goods to delivering customer experiences.
  Example: Starbucks India sells ambiance and experience, not just coffee.
- Omnichannel integration: Integration of physical and online presence.
  Example: Reliance Retail combines JioMart (online) with Reliance Smart stores (offline).
- Sustainability as strategy: Companies integrate sustainability into core strategy.
  Example: ITC's e-Choupal for rural empowerment and sustainable sourcing.

3. How the Competitive Environment is Changing

The competitive landscape has been reshaped by economic reforms, technological change, and evolving market structures.

- Deregulation: Reduction of government control has increased competition.
  Example: Telecom and aviation sectors opened up, leading to more consumer choices.
- Privatization: Government sectors are now competing with private firms.
  Example: LIC competing with private insurers like HDFC Life.
- Retail Transformation: Modern formats like malls, e-commerce, and organised retail chains are dominating.
  Example: D-Mart, Reliance Trends, and Big Bazaar replacing local kirana dominance.
- Disintermediation: Middlemen (wholesalers, retailers) are being bypassed due to direct-to-consumer (D2C) models.
  Example: boAt, Wow Skin Science selling directly through their websites.
- Private Labels: Retailers are launching their own brands to increase margins.
  Example: AmazonBasics, Reliance Select.
- Mega Brands: Conglomerates are consolidating under umbrella brands for stronger identity.
  Example: Hindustan Unilever's multiple products like Dove, Surf Excel, Lipton, etc., being clubbed under a unified strategy.

---

10. Key Terms

Some types of marketing, which have lately emerged, are:

Micro Marketing: Micro Marketing is the practice of tailoring products and marketing programs according to the needs and wants of specific individuals or local customer groups. It includes Local Marketing and Individual Marketing. Individual marketing is also called One to One Marketing or Customized Marketing.

Macro Marketing: Macro Marketing is a term that has been recently used to refer to the study of marketing within context of the entire economic system with special emphasis on its aggregate performance.

Mega Marketing: Mega Marketing is defined as the strategic coordination of economic, psychological, political and public relations skill to gain the co-operation of a number of parties in order to enter the Indian marketing. This approach is adopted by those large companies who are trying to enter blocked markets.

De-Marketing: De-Marketing is state of affairs under which the demand far exceeds the supply. It is a reverse step is marketing.

Over marketing: Over marketing is the striving efforts by a firm to generate increased sales while neglecting quality control, production efficiency and cash flow management.

Re-Marketing: Re-Marketing refers to the process of finding or creating new users for an existing product. It is associated with "*faultering demand*". Re marketing helps in creating and providing new satisfaction levels to the customer.

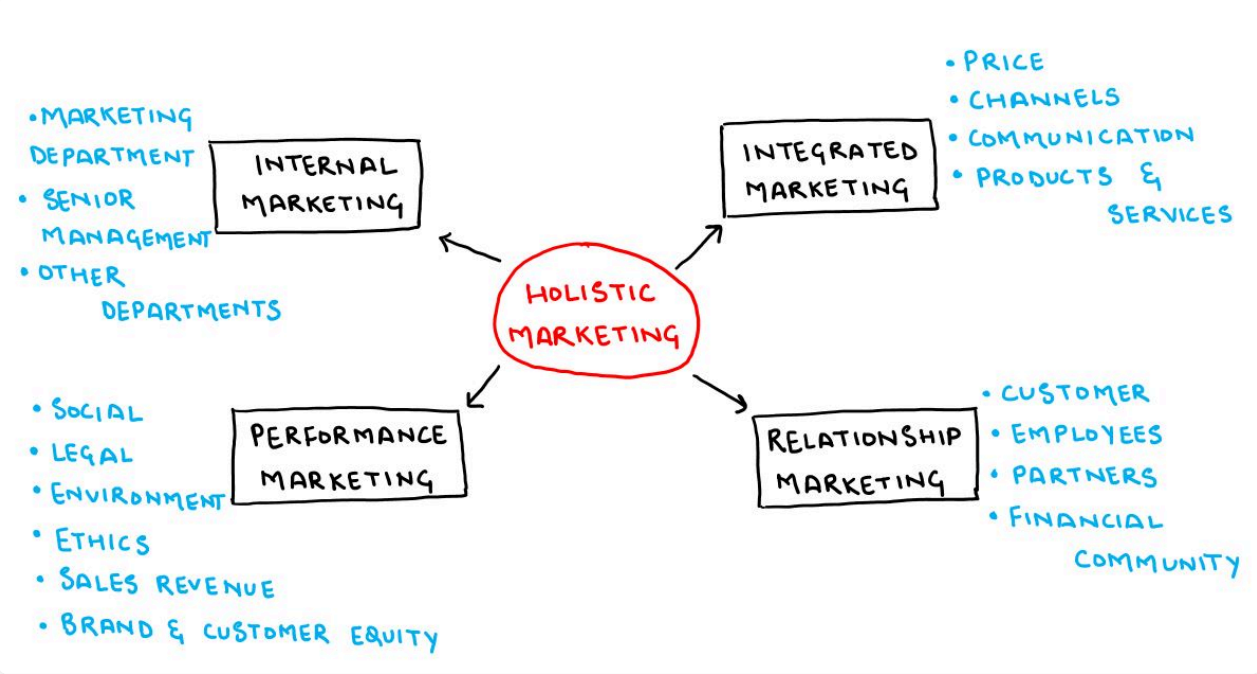
Reverse marketing: Reverse marketing is any marketing strategy that encourages consumers to seek out a company or a product on their own, rather than a company trying to sell specific products to consumers. Companies do this in a wide variety of ways, but the most common method is to provide valuable information to consumers without asking them to purchase anything. Service providers use reverse marketing to avoid what is known as "coercive" marketing. **Coercive marketing** tells customers that they should want a service, like getting a haircut, for a particular reason. A salon might use coercive advertising materials to suggest that people who don't take care of their hair cut at a salon are less attractive.

---

1. Holistic Marketing

The Holistic Marketing concept is based on the development, design, and implementation of marketing programs, processes, and activities that recognize their breadth and interdependencies. Holistic marketing recognizes that everything matters in marketing and that a broad, integrated perspective is often necessary.

Holistic marketing thus recognizes and reconciles the scope and complexities of marketing activities.



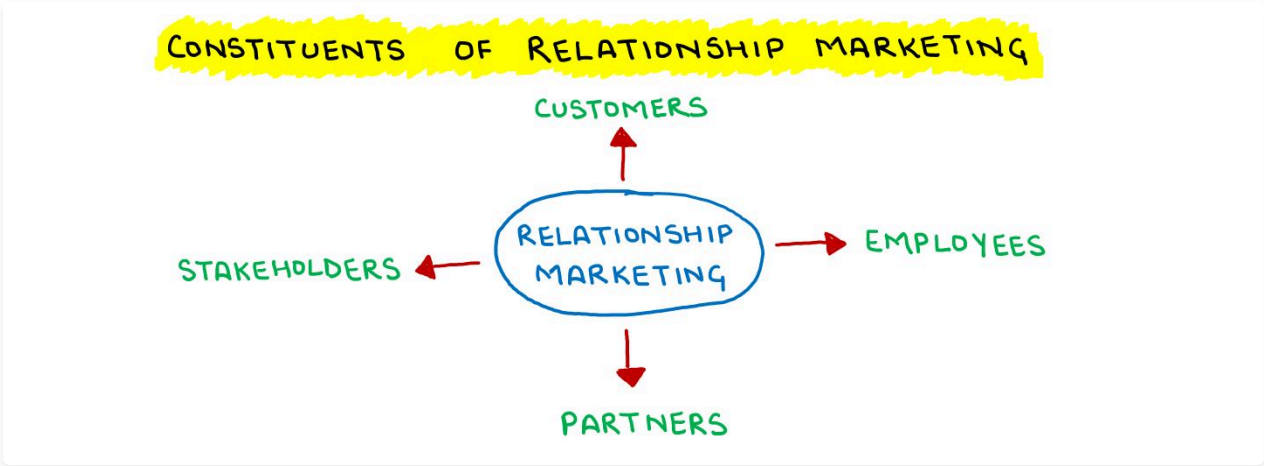
Holistics Marketing has 4 components:
1. Relationship Marketing
2. Integrated Marketing
3. Internal Marketing
4. Performance Marketing

Let us discuss them one by one.

---

1. Holistic Marketing

A key goal of marketing is to develop deep, enduring relationships with people and organizations that directly or indirectly affect the success of the firm's marketing activities. Relationship marketing aims to build mutually satisfying long-term relationships with key constituents in order to earn and retain their business.



Four key constituents for relationship marketing are-

1. Customers
2. Employees
3. Marketing partners (channels, suppliers, distributors, dealers, agencies)
4. Members of the financial community (shareholders, investors, analysts)

The ultimate outcome of relationship marketing is a unique company asset called a marketing network, consisting of the company and its supporting stakeholders—customers, employees, suppliers, distributors, retailers, and others—with whom it has built mutually profitable business relationships.

Marketing must skillfully conduct not only customer relationship management (CRM), but partner relationship management (PRM) as well. Companies are deepening their partnering arrangements with key suppliers and distributors, seeing them as partners in delivering value to final customers so everybody benefits.

Amazon, the e-commerce giant, is renowned for its relationship marketing strategies. Through its personalized recommendations, efficient customer service, and loyalty programs like Amazon Prime, the company fosters strong connections with its customers. Additionally, Amazon prioritizes its relationships with sellers, providing them with robust support, resources, and access to a vast customer base through its platform. By nurturing these relationships with both customers and partners, Amazon has built a loyal ecosystem that contributes to its sustained success and market dominance.

---

1. Holistic Marketing

It occurs when the marketer devises marketing activities and assembles marketing programs to create, communicate, and deliver value for consumers such that “the whole is greater than the sum of its parts.” Two key themes are that (i) many different marketing activities can create, communicate, and deliver value and (ii) marketers should design and implement any one marketing activity with all other activities in mind.

The company must develop an integrated channel strategy. It should assess each channel option for its direct effect on product sales and brand equity, as well as its indirect effect through interactions with other channel options.

Nike exemplifies integrated marketing through its multi-channel approach and consistent brand messaging. The company seamlessly integrates its advertising campaigns, social media presence, sponsorships, and retail experiences to create a cohesive brand narrative.

For instance, Nike's "Just Do It" campaign transcends traditional advertising platforms and extends to social media channels, in-store promotions, and sponsored events like marathons and sports competitions. By aligning all marketing activities with a unified brand identity and message, Nike maximizes its impact and creates a synergistic effect that enhances brand visibility and customer engagement across various touchpoints.

---

1. Holistic Marketing

It is an element of marketing, and is the task of hiring, training, and motivating able employees who want to serve customers well. It ensures that everyone in the organization embraces appropriate marketing principles, especially senior management.

Internal marketing requires vertical alignment with senior management and horizontal alignment with other departments so everyone understands, appreciates, and supports the marketing effort.

Zappos, the online shoe and clothing retailer, is renowned for its internal marketing practices. The company places a strong emphasis on hiring employees who align with its customer-centric culture and values. Zappos invests heavily in training programs to ensure that all employees, from customer service representatives to senior management, understand and embody the company's commitment to delivering exceptional customer experiences.

Additionally, Zappos fosters a supportive and inclusive work environment where employees feel valued and motivated to serve customers well. By prioritizing internal marketing efforts, Zappos cultivates a team of dedicated and enthusiastic employees who play a crucial role in delivering outstanding service and driving customer loyalty.

---

1. Holistic Marketing

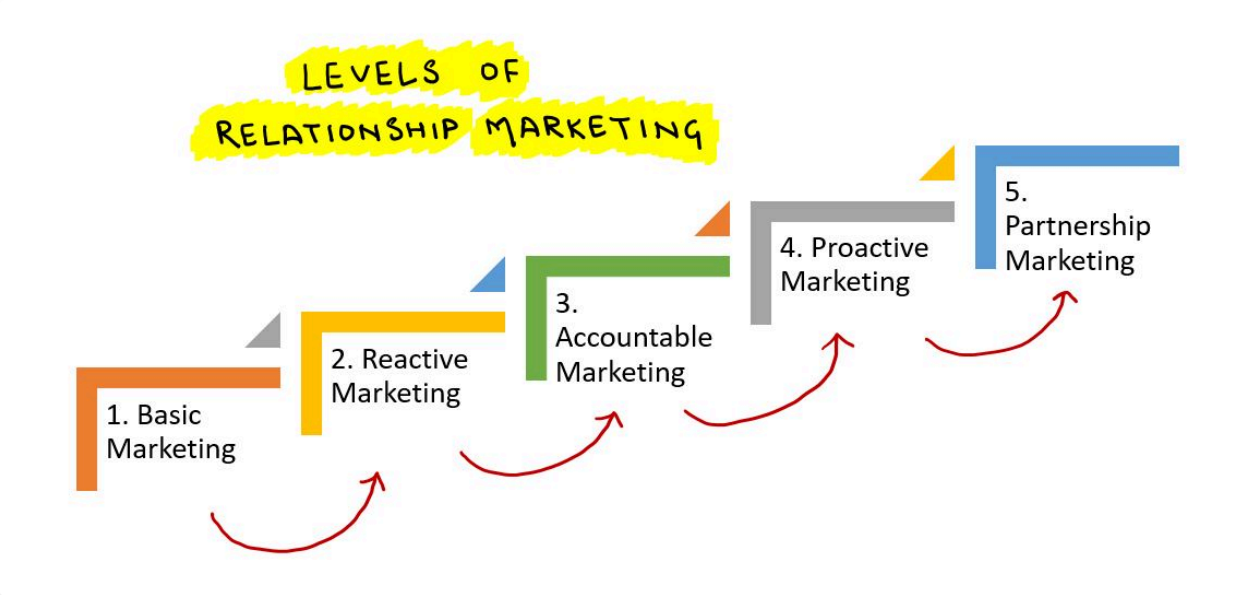
It requires understanding the financial and nonfinancial returns to business and society from marketing activities and programs. Smart marketers go beyond sales revenue to examine the marketing scorecard and interpret what is happening to market share, customer loss rate, customer satisfaction, product quality, and other measures. They also consider the legal, ethical, social and environmental effects of marketing activities and programs.

Google Ads, the online advertising platform, exemplifies performance marketing through its data-driven approach and focus on measurable outcomes. Advertisers using Google Ads have access to a wealth of analytics and performance metrics, allowing them to track the effectiveness of their campaigns in real-time. By analyzing key metrics such as click-through rates, conversion rates, and return on investment (ROI), advertisers can optimize their marketing efforts to maximize results and achieve their business objectives.

Furthermore, Google Ads considers the broader impacts of advertising activities, including their legal, ethical, and environmental implications, to ensure responsible marketing practices. Through its emphasis on data-driven decision-making and holistic performance evaluation, Google Ads enables marketers to drive tangible business outcomes while also considering societal and environmental factors.

---

2. Levels of Relationship Marketing



Different levels of relationship can be developed with customers. The extent to which an organization practices relationship marketing depends on its investment (time and money) in building the relationship.

There are 5 levels of relationship marketing:

1. Basic Marketing: The marketer simply sells the product without call backs or feed-back from the customers. Here, there is no effort to build the relationship.

2. Reactive Marketing: After selling the product, the marketer encourages customers to report their comments, complaints, and suggestions. Here, there is some effort to build the relationship.

3. Accountable Marketing: After selling the product, the marketer calls back the customer after a short time to find out if the customer is satisfied or not. Here, there is increased level of effort in building the relationship.

4. Proactive Marketing: After selling the product, the marketer remains in regular contact with the customer and provide suggestions for better use of the product, servicing, and repairs. The marketer also informs the customer about the firm's new products, improved quality and services. Here, the marketer is working harder to enhance the level of relationship.

5. Partnership Marketing: Under partnership marketing, the marketer continuously works with its customers and other stakeholders for mutual benefits. Here, the marketer works jointly with the customer to find ways and means to help customer perform better, effect savings, and increase the level of satisfaction. Thus, partnership marketing is the best form of relationship marketing.

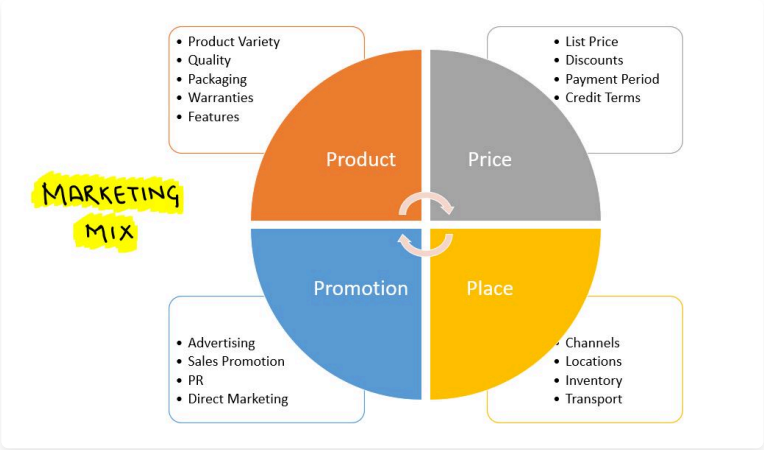
---

1. Marketing Mix

Marketing Mix is the set of marketing tools that the firm uses to pursue its marketing objectives in the target market. The concept was initially given by N H Borden.

Later *McCarthy* classified the concept into 4 broad groups that he called the "4 Ps of marketing": Product, Price, Place, and Promotion (This is the order of 4 Ps, which is used in marketing strategy).

For marketing of services, three additional P's have been identified, People, Process and Physical evidence.



The 4 P's of marketing represent the key elements that businesses need to consider when planning their marketing efforts. The 4 P's are:

1. Product

Product refers to the actual goods or services that a company offers to its target customers. It includes decisions about product design, features, branding, packaging, and quality. The goal is to create a product that meets the needs and desires of the target market and offers a unique value proposition.

2. Price

Price refers to the amount of money customers are willing to pay for the product or service. Setting the right price is crucial for a company's profitability and positioning in the market. Factors to consider when determining pricing include production costs, competition, perceived value, and pricing strategies such as premium pricing, penetration pricing, or discount pricing.

3. Place (Distribution)

This aspect focuses on getting the product into the hands of the target customers. It involves decisions about the distribution channels, logistics, and locations where the product will be available for purchase. The goal is to ensure that the product is easily accessible to the target market and reaches them through the most efficient and effective channels.

4. Promotion

Promotion encompasses all the activities aimed at communicating the value of the product to the target customers and persuading them to make a purchase. Key components of Promotion Mix are:

- Sales Promotion,
- Advertising,
- Sales Force,
- Public Relations and
- Direct Marketing.

The aim is to create awareness, generate interest, and ultimately drive sales of the product.

By effectively managing the 4 Ps of marketing, businesses can develop a well-rounded marketing strategy that aligns with their target market's needs and preferences, leading to increased sales, brand loyalty, and business success.

---

2. Marketing Offerings

Marketing is no longer limited to just the traditional product-centric view. In today's dynamic marketplace, companies offer not just products, but comprehensive value bundles called marketing offerings—a mix of tangible and intangible elements that satisfy customer needs and create long-term engagement.

Historically, marketers used the 4Ps framework—Product, Price, Place, Promotion—to plan their strategies. While effective for product-driven models, this framework has limitations when applied to modern markets involving services, digital platforms, and personalized solutions.

To address these limitations, modern marketers use a more inclusive framework comprising 7 Marketing Tactics (also called the 7 Elements of the Marketing Mix in some models). These better represent the full range of marketing offerings in today's business environment.

1. Product
The core physical item or digital solution offered to fulfill customer needs. This includes design, features, quality, and variety.
Example: iPhone, Tata Nexon EV.

2. Service
Support and experiences that accompany the product—like installation, delivery, customer care, and after-sales service. Especially critical in B2B and digital sectors.
Example: Flipkart's easy return service enhances the core product offering.

3. Brand
Branding gives identity, emotional appeal, and trust to an offering. A strong brand helps command premium pricing and customer loyalty.
Example: Amul, Tanishq, and LIC are brands deeply trusted in India.

4. Price
Represents the exchange value. Modern pricing goes beyond just setting a figure; it includes discounts, dynamic pricing, psychological pricing, etc.
Example: Ola uses surge pricing; D-Mart uses everyday low pricing.

5. Incentives
Includes offers, deals, referral bonuses, and loyalty rewards that create short-term motivation to buy or engage.
Example: Paytm cashback offers or Swiggy's "Buy 1 Get 1 Free" deals.

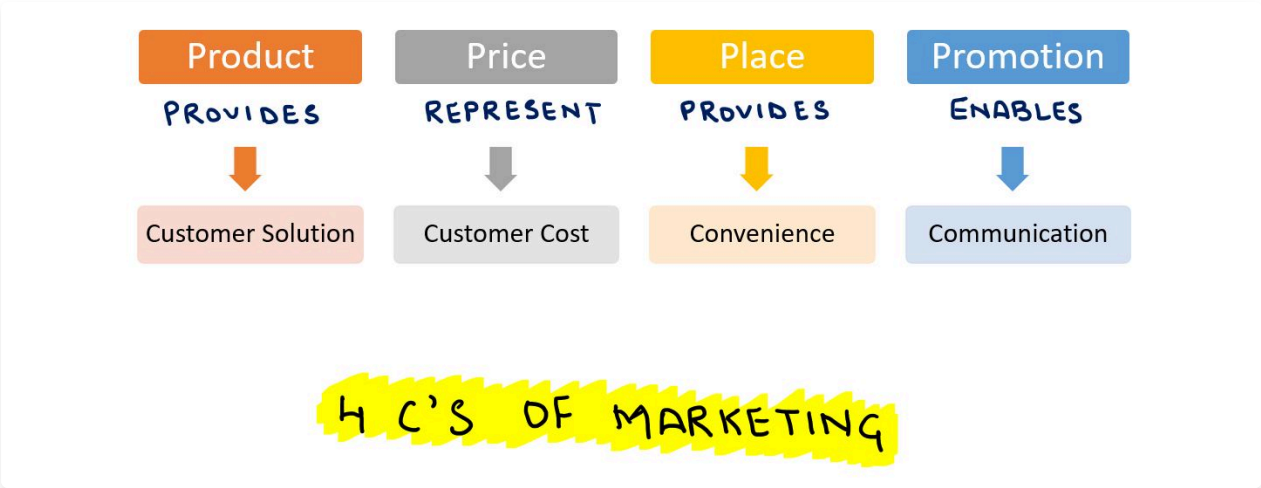
6. Communication
Covers all promotional efforts—advertising, PR, social media, influencer marketing, and customer engagement. It ensures awareness and brand recall.
Example: Surf Excel's emotionally driven ads like "Daag Achhe Hain."

7. Distribution
How the product or service is delivered to the customer—channels, locations, logistics, and accessibility. With the rise of e-commerce and omni-channel models, this has become a strategic area.
Example: JioMart integrates local kiranas with online ordering systems.

---

3. 4 C's of Marketing

Robert Lauterborn suggested that the sellers' 4 Ps correspond to the customers' 4 Cs, which include –



Product → Customer Solution

Instead of merely emphasizing the product, the focus shifts to providing a comprehensive solution that meets the specific needs and wants of the customer. It's about offering a solution rather than just a tangible item.

Example: If a company sells smartphones, the emphasis is not only on the features and specifications (Product) but also on how those features provide a solution to the customer's communication and entertainment needs.

Price → Customer Cost

The concept of cost expands beyond the monetary price to encompass the total cost to the customer, considering factors such as time, effort, and any additional expenses associated with obtaining and using the product or service.

Example: Instead of just promoting a low price, the marketing strategy considers the overall cost-effectiveness for the customer, including long-term value and any hidden costs.

Place → Convenience

Rather than focusing solely on distribution channels and physical locations, the emphasis shifts to providing convenience in terms of accessibility, ease of purchase, and the overall customer experience.

Example: In addition to traditional retail locations, marketers consider online platforms, home delivery options, and other elements that enhance the convenience of obtaining the product.

Promotion → Communication

Communication is not just about promoting the product; it involves engaging in a meaningful and interactive dialogue with the customer. It emphasizes two-way communication and building strong relationships.

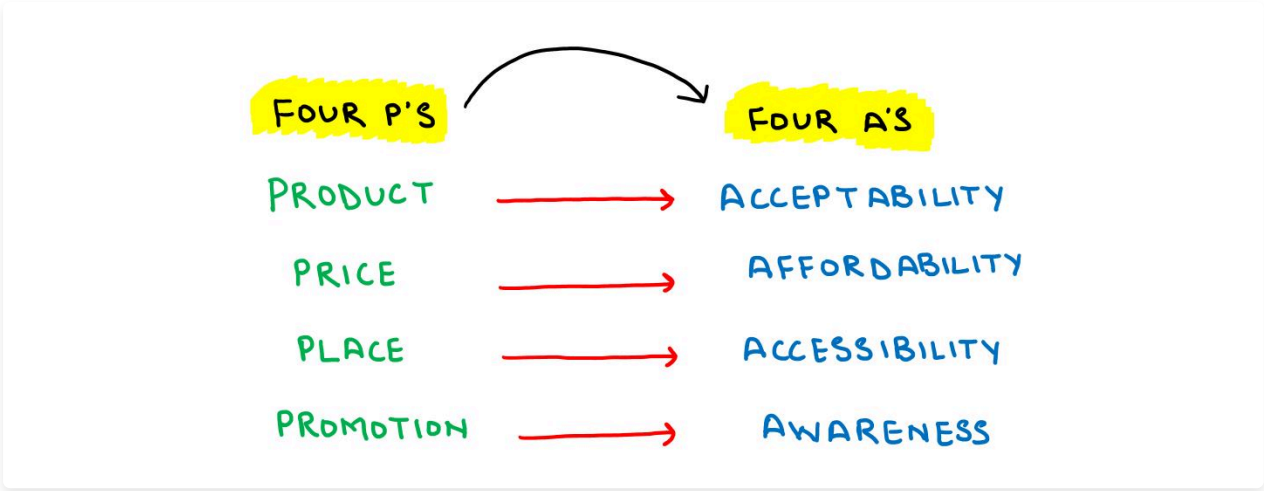
Example: Instead of one-way advertising, marketers use various channels for communication, such as social media, customer service interactions, and personalized messaging, to foster engagement and build a connection with customers.

Lauterborn's 4 Cs provide a customer-centric perspective, encouraging businesses to shift their focus from selling products to creating solutions, considering the overall cost to the customer, prioritizing convenience, and engaging in meaningful communication.

---

4. 4 A's of Marketing

Similarly, 4 P's have also been connected to 4 A's from buyer's view. The 4 As include:



Product → Acceptability

Acceptability is about the extent to which the product exceeds customer expectations. It goes beyond the basic features and quality, focusing on the overall satisfaction and how well the product meets or exceeds customer needs.

Example: If a company sells smartphones, the focus on acceptability means not just meeting technical specifications (Product) but ensuring the phone's overall user experience, features, and design are beyond customer expectations.

Price → Affordability

Affordability in the buyer's context considers the extent to which customers are willing and able to pay the product's price. It goes beyond the monetary value and includes the perceived value and willingness of customers to make the purchase.

Example: Rather than just setting a price based on production costs (Price), marketers consider the target market's willingness and ability to pay, ensuring that the product is affordable and perceived as valuable.

Place → Accessibility

Accessibility focuses on the extent to which customers can readily acquire the product. It goes beyond distribution channels and physical locations to consider the overall ease with which customers can access and obtain the product.

Example: Instead of merely placing products in retail locations (Place), marketers consider online platforms, delivery options, and other factors to enhance the overall accessibility of the product.

Promotion → Awareness

Awareness is about the extent to which customers are informed about the product's features, persuaded to try it, and reminded to repurchase. It extends beyond traditional promotion to encompass the entire customer journey, from awareness to post-purchase engagement.

Example: Instead of focusing solely on advertising (Promotion), marketers use various communication channels to inform customers about the product's features, persuade them to try it, and implement strategies for post-purchase reminders and engagement.

---

5. Macro Strategies

There are 4 macro strategies that focus on aspects of the marketing mix:

1. Customer Excellence

- **Focus:** Retaining loyal customers and providing excellent customer service.
- **Explanation:** This strategy centers around building strong, long-term relationships with customers. By focusing on customer satisfaction, businesses aim to create loyalty and positive word-of-mouth. Providing exceptional service and meeting or exceeding customer expectations are key components of customer excellence.
- **Example:** A luxury hotel chain may invest heavily in staff training, personalized services, and loyalty programs to enhance the overall experience and retain high-value customers.

2. Operational Excellence

- **Focus:** Achieved through efficient operations, excellent supply chain management, and human resource management.
- **Explanation:** Operational excellence is about optimizing internal processes to achieve efficiency and effectiveness. This includes streamlining operations, managing the supply chain efficiently, and ensuring that the workforce is well-trained and motivated. It often leads to cost savings and improved overall business performance.
- **Example:** A fast-food restaurant chain may focus on optimizing its supply chain to ensure fresh ingredients, efficient kitchen processes, and well-trained staff to enhance operational excellence.

3. Product Excellence

- **Focus:** Having products with high perceived value, effective branding, and positioning.
- **Explanation:** This strategy revolves around offering high-quality products that meet or exceed customer expectations. Effective branding and positioning play a crucial role in communicating the value of the products to the target audience. Creating a positive perception of the product's quality and uniqueness is essential for product excellence.
- **Example:** An electronic gadget company may invest in research and development, design, and branding to position its products as innovative, reliable, and of high quality.

4. Locational Excellence

- **Focus:** Having a good physical location and a strong Internet presence.
- **Explanation:** Locational excellence emphasizes the importance of being in the right place to reach the target market effectively. This includes having a physical presence in strategic locations and a strong online presence to cater to the growing importance of e-commerce. Accessibility and visibility are key considerations in locational excellence.
- **Example:** A retail brand may strategically choose locations with high foot traffic for its physical stores, and simultaneously invest in a user-friendly e-commerce platform to ensure locational excellence in both the physical and digital realms.

---

1. Introduction

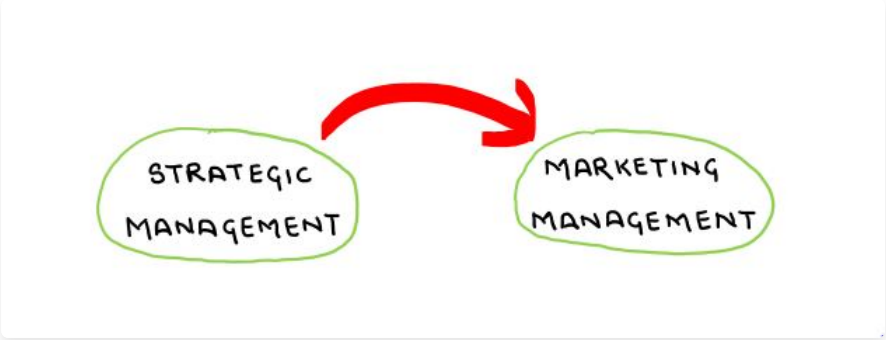
Imagine launching a great product — it's high quality, well-priced, and exactly what people need. But after months in the market, the results are disappointing. Customers aren't showing up, sales are flat, and competitors are getting ahead. You wonder — what went wrong? The answer often lies not in the product, but in the strategy behind it.

Today, companies operate in a crowded, fast-moving market. Consumer preferences change quickly. Competitors are aggressive. New technologies keep reshaping how businesses connect with customers. In such an environment, simply having a good product or service is not enough. You need a clear plan — a blueprint — that guides how you'll win in the market.

This is where developing a strong marketing strategy becomes essential. Before we get into the how, we'll first understand the bigger picture — how businesses make long-term decisions, what direction they choose, and how marketing becomes a powerful tool in making that vision a reality.

---

2. Strategic Management



Strategic management serves as the foundational framework guiding an organization towards its intended goals by orchestrating various components, including resource allocation, long-term planning, and decision-making processes. It encompasses a holistic view of the company's objectives, internal strengths and weaknesses, external opportunities, and threats.

Every successful business starts with a larger vision — what it wants to achieve, where it wants to go, and how it plans to get there. This process of setting direction, making choices, and ensuring all departments work in alignment is called **strategic management**. It's not about reacting to day-to-day events — it's about proactively shaping the future of the organization.

Within this broader direction, **marketing strategy** becomes the frontline driver of business success. While strategic management defines the overall mission, marketing strategy takes this mission to the market — identifying the right customers, positioning the brand, and creating value in a way that supports business goals.

You can think of strategic management as the "brain" of the organization — setting direction — and marketing strategy as the "voice and face" that connects with the outside world. When both work together, companies don't just operate — they compete, grow, and lead.

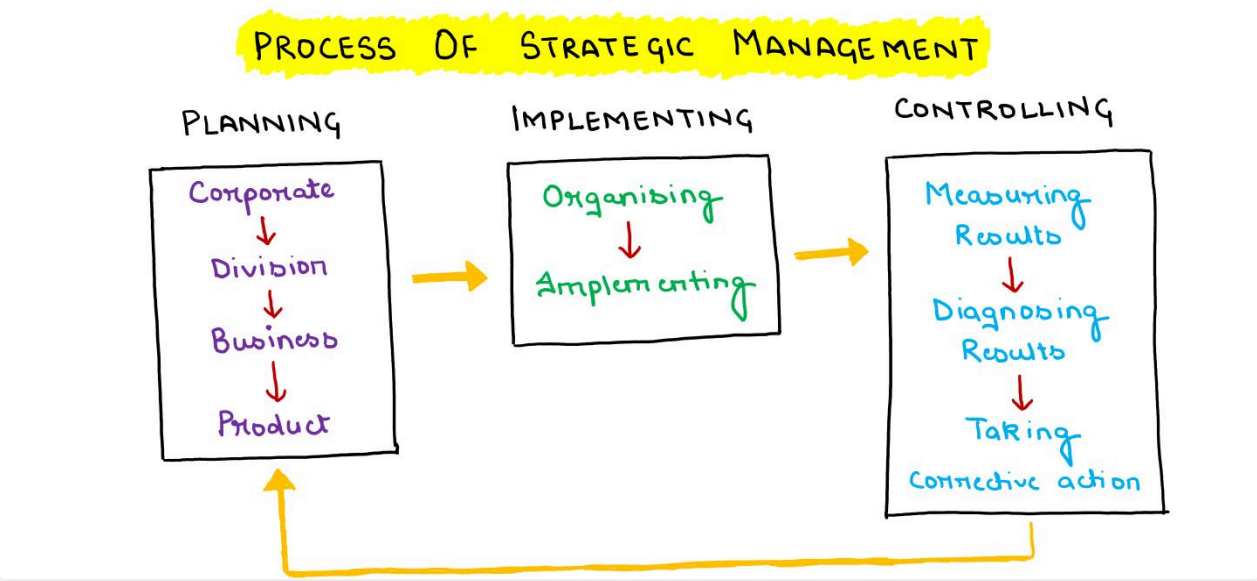
In essence, marketing strategy evolves from the overarching strategic management framework. Let us first understand the process of Strategic Management.

---

3. Process of Strategic Management

The process of strategic management involves 3 essential steps:

1. Strategy Planning,
2. Implementing, and
3. Controlling



Let's delve into each step and explore its components.

---

4. Strategy Planning

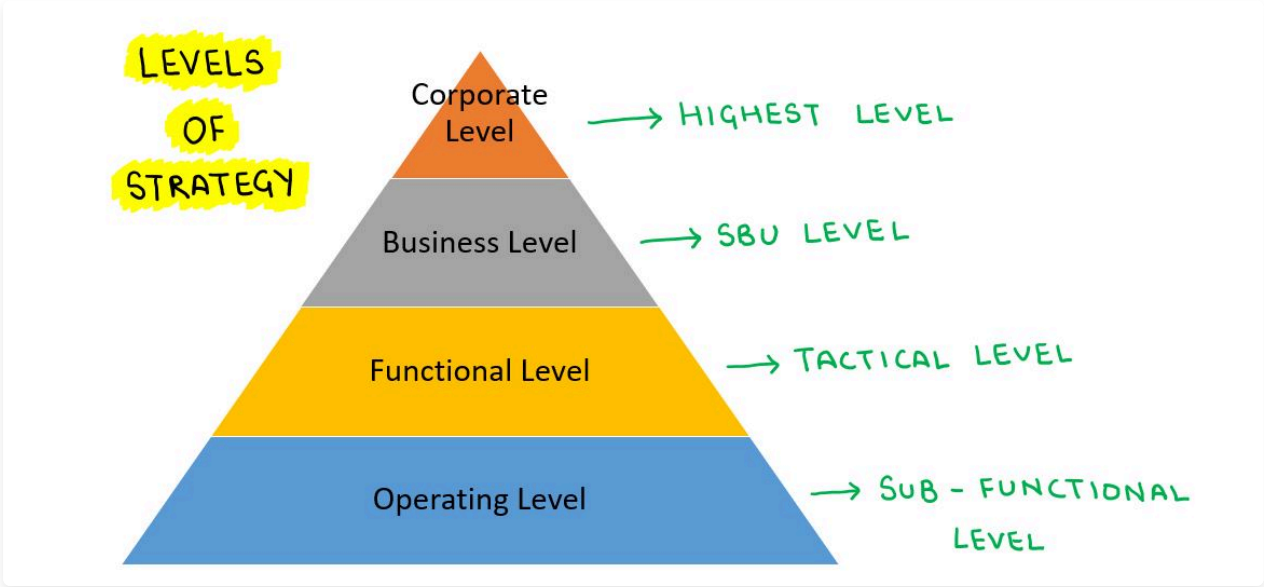
Strategy planning lays the foundation for a comprehensive marketing approach. It involves creating a roadmap to guide the organization at various levels.

Before we understand the process of strategy planning, let us understand the levels of Strategy Planning.

---

5. Levels of Strategy Planning

It is believed that strategic decision making is the responsibility of top management. However, it is considered useful to distinguish between the levels of operations of the strategy. Strategy operates at three levels: Corporate Level, Business Level, and Functional Level (this hierarchical model was proposed by C. Ronald Christensen). However sometimes, a fourth level is also created, known as Operating level.



These 4 levels are explained as follows:

1. Corporate Level Strategy

At the corporate level, strategic decision-making involves setting objectives for the organization based on its policies. These decisions, made by top management, carry higher risk, cost, and profit potential, with an emphasis on future-oriented, innovative, and pervasive strategies. Corporate strategies, often categorized as stability, growth, or retrenchment, encompass actions like acquisitions, diversification, and structural redesign. The Board of Directors and the Chief Executive Officer play pivotal roles in this strategic decision-making process, particularly in smaller or family-owned businesses where the entrepreneur serves as both the general and chief strategic manager.

2. Business Levels Strategy

Business strategy operates at the business unit or product level, focusing on enhancing the competitive position of a corporation's products or services within a specific industry or market segment. Two company categories exist: those with different businesses organized as profit centers or strategic business units (e.g., Reliance Industries Limited), and single-product companies (e.g., Ashok Leyland Ltd). The Strategic Business Unit (SBU) concept, introduced by General Electric Company, involves creating discrete product/market segments for effective management. Business-level strategies are formulated by each SBU to optimize resources and align with overall organizational goals, operating within the constraints and policies set by corporate strategy.

3. Functional Level Strategy

Functional strategy pertains to individual functional operations within an organization, addressing activities at the operational end. It aims to achieve corporate and business unit objectives by maximizing resource productivity and nurturing distinctive competencies. Tactical decisions at this level focus on how functions like marketing, finance, and manufacturing contribute to higher-level strategies. Various functional strategies include marketing, financial, R&D, operations, and logistics strategies, each tailored to specific objectives within their functional area. The orientation of a functional strategy aligns with the parent business unit's strategy, ensuring coherence with overarching goals.

4. Operational Level Strategy

The operational level, positioned below functional strategy, involves actions related to sub-functions within major functions. For instance, within the marketing function, operational levels may include marketing research and sales promotion. This level addresses detailed implementation and execution, ensuring alignment with the broader functional and corporate strategies.