UNIT – 3 : INTRODUCTORY HYPOTHESIS TESTING AND STATISTICAL INFERENCE

Introduction to Hypothesis Testing:

Hypothesis testing is a statistical method that is used to make a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Example: You say an average height in the class is 30 or a boy is taller than a girl. All of these is an assumption that we are assuming, and we need some statistical way to prove these. We need some mathematical conclusion whatever we are assuming is true.

Defining Hypotheses:

- **Null hypothesis (H0):** In statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.

  Example: A company's mean production is 50 units/per i.e. H0: $\mu = 50$.

- **Alternative hypothesis (H1):** The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

  Example: A company's production is not equal to 50 units/per day i.e. H1: $\mu \neq 50$.

Key Terms of Hypothesis Testing:

- **Level of significance**: It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we, therefore, select a level of significance that is usually 5%. This is normally denoted with $\alpha$and generally, it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.
- **P-value:** The P value, or calculated probability, is the probability of finding the observed/extreme results when the null hypothesis(H0) of a study-given problem is true. If your P-value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample claims to support the alternative hypothesis.
- **Test Statistic:** The test statistic is a numerical value calculated from sample data during a hypothesis test, used to determine whether to reject the null hypothesis. It

is compared to a critical value or p-value to make decisions about the statistical significance of the observed results.

- **Critical value**: The critical value in statistics is a threshold or cutoff point used to determine whether to reject the null hypothesis in a hypothesis test.
- **Degrees of freedom:** Degrees of freedom are associated with the variability or freedom one has in estimating a parameter. The degrees of freedom are related to the sample size and determine the shape.

Why do we use Hypothesis Testing?

Hypothesis testing is an important procedure in statistics. Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data. When we say that the findings are statistically significant, thanks to hypothesis testing.

One-Tailed and Two-Tailed Test:

One tailed test focuses on one direction, either greater than or less than a specified value. We use a one-tailed test when there is a clear directional expectation based on prior knowledge or theory. The critical region is located on only one side of the distribution curve. If the sample falls into this critical region, the null hypothesis is rejected in favor of the alternative hypothesis.

One-Tailed Test:

There are two types of one-tailed test:

- **Left-Tailed (Left-Sided) Test:** The alternative hypothesis asserts that the true parameter value is less than the null hypothesis.

  Example: H0: $\mu \geq 50$ and H1: $\mu < 50$

- **Right-Tailed (Right-Sided) Test**: The alternative hypothesis asserts that the true parameter value is greater than the null hypothesis.

  Example: H0: $\mu \leq 50$ and H1: $\mu > 50$

Two-Tailed Test:

A two-tailed test considers both directions, greater than and less than a specified value.We use a two-tailed test when there is no specific directional expectation, and want to detect any significant difference.

Example: H0: $\mu = 50$ and H1: $\mu \neq 50$

What are Type 1 and Type 2 errors in Hypothesis Testing?

In hypothesis testing, Type I and Type II errors are two possible errors that researchers can make when drawing conclusions about a population based on a sample of data. These errors are associated with the decisions made regarding the null hypothesis and the alternative hypothesis.

- **Type I error:** When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha (α).
- **Type II errors:** When we accept the null hypothesis, but it is false. Type II errors are denoted by beta (β).

	Null Hypothesis is True	**Null Hypothesis is False**
Null Hypothesis is True (Accept)	Correct Decision	Type II Error (False Negative)
Alternative Hypothesis is True (Reject)	Type I Error (False Positive)	Correct Decision

How does Hypothesis Testing work?

Step 1: Define Null and Alternative Hypothesis

State the null hypothesis (H_0), representing no effect, and the alternative hypothesis (H_1), suggesting an effect or difference.

We first identify the problem about which we want to make an assumption keeping in mind that our assumption should be contradictory to one another, assuming Normally distributed data.

Step 2: Choose significance level

Select a significance level (α), typically 0.05, to determine the threshold for rejecting the null hypothesis. It provides validity to our hypothesis test, ensuring that we have sufficient data to back up our claims. Usually, we determine our significance level beforehand of the test. The p-value is the criterion used to calculate our significance value.

Step 3: Collect and Analyze data.

Gather relevant data through observation or experimentation. Analyze the data using appropriate statistical methods to obtain a test statistic.

Step 4: Calculate Test Statistic

The data for the tests are evaluated in this step we look for various scores based on the characteristics of data. The choice of the test statistic depends on the type of hypothesis test being conducted.

There are various hypothesis tests, each appropriate for various goal to calculate our test. This could be a Z-test, Chi-square, T-test, and so on.

1. **Z-test**: If population means and standard deviations are known. Z-statistic is commonly used.
2. **t-test**: If population standard deviations are unknown. and sample size is small than t-test statistic is more appropriate.
3. **Chi-square test**: Chi-square test is used for categorical data or for testing independence in contingency tables
4. **F-test**: F-test is often used in analysis of variance (ANOVA) to compare variances or test the equality of means across multiple groups.

We have a smaller dataset, So, T-test is more appropriate to test our hypothesis.

T-statistic is a measure of the difference between the means of two groups relative to the variability within each group. It is calculated as the difference between the sample means divided by the standard error of the difference. It is also known as the t-value or t-score.

Step 5: Comparing Test Statistic

In this stage, we decide where we should accept the null hypothesis or reject the null hypothesis. There are two ways to decide where we should accept or reject the null hypothesis.

Method A: Using Crtical values

Comparing the test statistic and tabulated critical value we have,

- If Test Statistic>Critical Value: Reject the null hypothesis.
- If Test Statistic≤Critical Value: Fail to reject the null hypothesis.

Note: Critical values are predetermined threshold values that are used to make a decision in hypothesis testing. To determine critical values for hypothesis testing, we typically refer to a statistical distribution table, such as the normal distribution or t-distribution tables based on.

Method B: Using P-values

We can also come to a conclusion using the p-value,

- If the p-value is less than or equal to the significance level i.e. ($p \leq \alpha$), you reject the null hypothesis. This indicates that the observed results are unlikely to have occurred by chance alone, providing evidence in favor of the alternative hypothesis.
- If the p-value is greater than the significance level i.e. ($p \geq \alpha$), you fail to reject the null hypothesis. This suggests that the observed results are consistent with what would be expected under the null hypothesis.

Note: The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed in the sample, assuming the null hypothesis is true. To determine p-value for hypothesis testing, we typically refer to a statistical distribution table, such as the normal distribution or t-distribution tables based on.

Step 7- Interpret the Results

At last, we can conclude our experiment using method A or B.

Calculating Test Statistic:

To validate our hypothesis about a population parameter we use statistical functions. We use the z-score, p-value, and level of significance(alpha) to make evidence for our hypothesis for normally distributed data.

1. Z-statistics:

When population means and standard deviations are known.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where,

- $\bar{x}$ is the sample mean,
- μ represents the population mean,
- σ is the standard deviation
- and n is the size of the sample.

2. T-Statistics:

T test is used when n<30,

t-statistic calculation is given by:

$$t = \frac{x -}{s/\sqrt{n}}$$

where,

- t = t-score,
- $\bar{x}$ = sample mean
- μ = population mean,
- s = standard deviation of the sample,
- n = sample size

3. Chi-Square Test:

Chi-Square Test for Independence categorical Data (Non-normally distributed) using:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where,

- O_{ij} is the observed frequency in cell ij
- i,j are the rows and columns index respectively.
- E_{ij} is the expected frequency in cell ij, calculated as:

$$\frac{\text{Row total} \times \text{Column total}}{\text{Total observations}}$$

Real life Hypothesis Testing Example:

Let's examine hypothesis testing using two real life situations,

Case A: Does a New Drug Affect Blood Pressure?

Imagine a pharmaceutical company has developed a new drug that they believe can effectively lower blood pressure in patients with hypertension. Before bringing the drug to market, they need to conduct a study to assess its impact on blood pressure.

Data:

- Before Treatment: 120, 122, 118, 130, 125, 128, 115, 121, 123, 119
- After Treatment: 115, 120, 112, 128, 122, 125, 110, 117, 119, 114

Step 1: Define the Hypothesis

- **Null Hypothesis**: (H0)The new drug has no effect on blood pressure.
- **Alternate Hypothesis**: (H1)The new drug has an effect on blood pressure.

Step 2: Define the Significance level

Let's consider the Significance level at 0.05, indicating rejection of the null hypothesis.

If the evidence suggests less than a 5% chance of observing the results due to random variation.

Step 3: Compute the test statistic

Using paired T-test analyze the data to obtain a test statistic and a p-value.

The test statistic (e.g., T-statistic) is calculated based on the differences between blood pressure measurements before and after treatment.

$$\mathbf{t = m/(s/\sqrt{n})}$$

Where:

- **m** = mean of the difference i.e Xafter, Xbefore
- **s** = standard deviation of the difference (d) i.e $di=X$after,$i-X$before,
- **n** = sample size,

then, m= -3.9, s= 1.8 and n= 10

we, calculate the, T-statistic = -9 based on the formula for paired t test

Step 4: Find the p-value

The calculated t-statistic is -9 and degrees of freedom $df = 9$, you can find the p-value using statistical software or a t-distribution table.

thus, p-value = 8.538051223166285e-06

Step 5: Result

- If the p-value is less than or equal to 0.05, the researchers reject the null hypothesis.
- If the p-value is greater than 0.05, they fail to reject the null hypothesis.

Conclusion: Since the p-value (8.538051223166285e-06) is less than the significance level (0.05), the researchers reject the null hypothesis. There is statistically significant evidence that the average blood pressure before and after treatment with the new drug is different.

Case B: Cholesterol level in a population

Data: A sample of 25 individuals is taken, and their cholesterol levels are measured.

Cholesterol Levels (mg/dL): 205, 198, 210, 190, 215, 205, 200, 192, 198, 205, 198, 202, 208, 200, 205, 198, 205, 210, 192, 205, 198, 205, 210, 192, 205.

Populations Mean = 200

Population Standard Deviation (σ): 5 mg/dL(given for this problem)

Step 1: Define the Hypothesis

- **Null Hypothesis (H0):** The average cholesterol level in a population is 200 mg/dL.
- **Alternate Hypothesis (H1):** The average cholesterol level in a population is different from 200 mg/dL.

Step 2: Define the Significance level

As the direction of deviation is not given , we assume a two-tailed test, and based on a normal distribution table, the critical values for a significance level of 0.05 (two-tailed) can be calculated through the z-table and are approximately -1.96 and 1.96.

Step 3: Compute the test statistic

The test statistic is calculated by using the z formula $Z=(203.8 - 200)/(5 \div \sqrt{25})$ and we get accordingly, Z=2.039999999999992.

Step 4: Result

Since the absolute value of the test statistic (2.04) is greater than the critical value (1.96), we reject the null hypothesis. And conclude that, there is statistically significant evidence that the average cholesterol level in the population is different from 200 mg/dL

Limitations of Hypothesis Testing:

- Although a useful technique, hypothesis testing does not offer a comprehensive grasp of the topic being studied. Without fully reflecting the intricacy or whole context of the phenomena, it concentrates on certain hypotheses and statistical significance.
- The accuracy of hypothesis testing results is contingent on the quality of available data and the appropriateness of statistical methods used. Inaccurate data or poorly formulated hypotheses can lead to incorrect conclusions.
- Relying solely on hypothesis testing may cause analysts to overlook significant patterns or relationships in the data that are not captured by the specific hypotheses being tested. This limitation underscores the importance of complimenting hypothesis testing with other analytical approaches.

Central Limit Theorem:

The **Central Limit Theorem (CLT)** is a fundamental concept in statistics. It states that when you take a large enough sample size from any population with any distribution, the distribution of the sample means will be approximately normal, regardless of the original distribution of the population.

In simpler terms, it suggests that if you repeatedly take samples from any population and calculate the average of each sample, those averages will tend to follow a normal (bell-shaped) distribution, even if the original data doesn't follow a normal distribution.

Central Limit Theorem Formula:

Let us assume we have a random variable X. Let σ be its standard deviation and μ is the mean of the random variable. Now as per the Central Limit Theorem, the sample mean $\bar{X}$ will approximate to the normal distribution which is given as $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. The Z-Score of the random variable $\bar{X}$ is given as

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

$$\text{Sample Mean} = \text{Population Mean} = \mu$$

$$\text{Sample Standard Deviation} = \frac{\text{Standard Deviation}}{n}$$

OR

$$\text{Sample Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

Assumptions Behind the Central Limit Theorem:

- The data must adhere to the randomization rule. It needs to be sampled at random.
- The samples should be unrelated to one another. One sample should not impact the others.
- When taking samples without replacement, the sample size should not exceed 10% of the population.
- When the population is symmetric, a sample size of 30 is generally considered reasonable.
- CLT only holds for a population with finite variance.

Why n ≥ 30 Samples?

The sample size of 30 is considered sufficient to see the effect of the CLT. If the population distribution is closer to the normal distribution, you will need fewer samples to demonstrate the central limit theorem. On the other hand, if the population distribution is highly skewed, you will need a large number of samples to understand the CLT.

Steps to Solve Problems on Central Limit Theorem:

Problems of Central Limit Theorem that involves >, < or between can be solved by the following steps:

- **Step 1:** First identify the >, < associated with sample size, population size, mean and variance in the problem. Also, there can be 'between; associated with range of two numbers.
- **Step 2:** Draw a Graph with Mean as Centre
- **Step 3:** Find the Z-Score using the formula
- **Step 4:** Refer to the Z table to find the value of Z obtained in the previous step.
- **Step 5:** If the problem involves '>' subtract the Z score from 0.5; if the problem involves '<' add 0.5 to the Z score and if the problem involves 'between' then perform only step 3 and 4.
- **Step 6:** The Z score value is found along $\bar{X}$X
- **Step 7:** Convert the decimal value obtained in all three cases to decimal.

Central Limit Theorem Applications:

Central Limit Theorem is generally used to predict the characteristics of a population from a set of samples. It can be applied in various fields. Some of the applications Central Limit Theorem are mentioned below:

- Central Limit Theorem is used by Economist and Data Scientist to draw conclusion about population to make a statistical model.
- Central Limit Theorem is used by Biologists to make accurate predictions about the characteristics of the population from set of samples.
- Manufacturing Industries use Central Limit Theorem to predict overall defective items produced by selecting random products from a sample.

- Central Limit Theorem is used in surveys to predict the characteristics of the population or to predict the average response of the population by analyzing a sample of obtained responses.
- CLT can be used in Machine Learning to make conclusion about the performance of the model.

Examples on Central Limit Theorem:

Example 1:

20 students are selected at random from a clinical psychology class; find the probability that their mean GPA is more than 5. If the average GPA scored by the entire batch is 4.91, the standard deviation is 0.72.

Solution:

Here,

Population mean = μ = 4.91

Population standard deviation= σ = 0.72

Sample size = n = 20 (which is less than 30)

Since the sample size is smaller than 30, use the t-score instead of the z-score, even though the population standard deviation is known.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Substituting the values, we have $\sigma_{\bar{x}} = \frac{0.72}{\sqrt{20}} = 0.161$

Now, find the t-score:

$$t = \frac{x - \mu}{\sigma_{\bar{x}}}$$

For this problem, the raw score x = 5. So, $t = \frac{5 - 4.91}{0.161} = 0.559$

Find the probability for the t value using the t-score table. The degree of freedom here would be:

Df = 20 – 1 = 19

P(t ≤ 0.559) = 0.7087

P(t > 0.559) = 1 – 0.7087 = 0.2913

Thus, the probability that the score is more than 5 is 9.13 %.

Example 2:

The average weight of a water bottle is 30 kg, with a standard deviation of 1.5 kg. If a sample of 45 water bottles is selected at random from a consignment and their weights are measured, find the probability that the mean weight of the sample is less than 28 kg.

Solution:

Population mean: μ = 30 kg

Population standard deviation: σ = 1.5 Kg

Sample size: n = 45 (which is greater than 30)

Using the z-score, we have

The sample standard deviation: $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

And,

$\sigma_{\overline{x}} = \frac{1.5}{\sqrt{45}}$ = 6.7082

Find the z-score for the raw score of x = 28 kg

$z = \frac{x-\mu}{\sigma_{\overline{x}}}$ = (28 – 30)(6.7082) = -0.2981

Using the z-score table OR normal CDF function on a statistical calculator,

P(z < -0.2981) = 0.3828

Thus, the probability that the weight of the cylinder is less than 28 kg is 38.28%.

Example 3: The record of weights of the female population follows a normal distribution. Its mean and standard deviation are 65 kg and 14 kg, respectively. If a researcher considers the records of 50 females, then what would be the standard deviation of the chosen sample?

Solution:

Mean of the population μ = 65 kg

The standard deviation of the population = 14 kg

Sample size n = 50

Standard deviation is given by $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

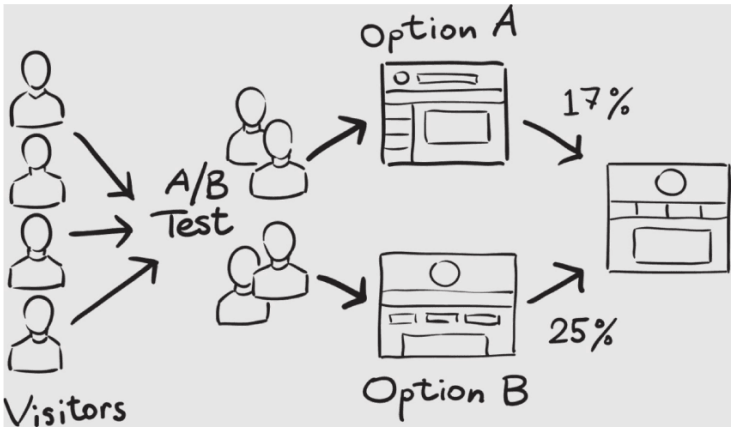
= 14/√50 = 14/7.071

= 1.97

A/B Testing:

A/B testing, also known as **split testing** or **bucket testing**, is a method used in data science to compare two versions of a webpage, app feature, or any other item to determine which one performs better in terms of a specific metric (e.g., conversion rate, click-through rate).

This technique is widely used in various fields such as marketing, product development, web design, and software engineering to optimize user experiences and improve outcomes.



When to Use A/B Testing in Data Science:

Testing incremental changes, such as UX adjustments, new features, ranking, and page load times, is where A/B testing excels. Here, you may compare the outcomes before and after the modifications to determine whether the adjustments are having the desired effect.

When testing significant changes, such as new products, new branding, or entirely new user experiences, A/B testing doesn't function effectively. In certain situations, there might be impacts that promote stronger-than-usual engagement or emotional reactions that might influence users' behavior.

How does A/B Testing Work?

A/B testing works wonders, but only if the steps are followed meticulously. Here are some critical steps in designing a successful A/B test:

1. Formulate a Hypothesis:

A hypothesis states how the change of a test variable impacts a performance metric on a population. An example of a hypothesis is the following:

“Changing the color of the add-to-cart button from blue to red (the test variable) will increase the conversion rate (the performance metric) on all desktop users (the population)”.

One common pitfall is having multiple test variables in a single experiment, making it difficult to tease out the impact of each change on the metric.

2. Create control and treatment versions of your test variable:

The term “A/B” in A/B testing refers to the two versions of the thing you’re testing.

Colloquially, the control is “Version A” of the product or the existing version of the test variable you’re testing. Whereas “Version B” is the treatment or the new version of the test variable you’re testing.

Using the same example of the add-to-cart button, the control (Version A) is the existing blue add-to-cart button, while the treatment (Version B) is the new red button.

3. Determine the sample size for statistical significance:

Depending on the use case and the number of users a service has, it can be impossible to run an A/B test on all the population. The next best alternative is to run the A/B test on a subset or sample of users. To do this, practitioners usually determine a statistically significant sample of users that is large enough for them to make conclusions about the population.

For example, when revisiting our example of the add-to-cart button, the A/B test would be run on a fraction of desktop users instead of all desktop users.

4. Select randomized groups for control and treatment:

Each user in the experiment is shown either version A or version B. How do we decide whether a user is given the control or the treatment?

To ensure that the test is fair and square, practitioners usually split the samples into the treatment and control groups randomly (as in, each user has the same probability of being in treatment or control) and equally (as in, the treatment and control groups are of the same size).

5. Run the test, and analyze the results:

With the groundwork done, the A/B test is ready to go. Once a large enough sample is reached, the results of an A/B test can be analyzed.

To analyze the results, we calculate the difference in the test metric–conversion rate–between the treatment and control groups. If the difference is significant enough, we can confidently conclude that one version is indeed better than the other.

6. Iterate, iterate, iterate

If a clear winner emerges from the A/B test, the superior version can now go live! Practitioners also perform a deep dive into the data to better understand users' behavior. An A/B test is but a small part of the optimization process. Learnings extracted from running A/B tests can inspire new ideas and hypotheses whose validity can, in turn, be tested.

Mistakes We Must Avoid While Conducting A/B Testing in Data Science:

There are a few crucial errors that data science experts make. Here, let me explain them to you:

1. Invalid Hypothesis: The hypothesis is the only thing on which the entire experiment is predicated. What needs to be altered? What justifies the change, what results are anticipated, etc.? The likelihood that the test will be successful diminishes if you begin with the incorrect hypothesis.

2. Testing too many components at once: Run as few tests as possible at once, industry experts advise. It might be challenging to determine which aspect contributed to success or failure when too many variables are tested simultaneously. As a result, prioritizing tests is crucial for effective A/B testing.

3. Ignoring Statistical Significance: Your opinion of the test is irrelevant. Allow the test to run its full course, whether it is successful or not, so that it obtains statistical significance.

4. Not taking external factors into account: To get significant findings, tests should be run during comparable times. For instance, comparing website traffic on days with the highest traffic to days with the lowest traffic due to outside reasons like sales or holidays is unfair.

Advantages Of A/B Testing:

1. Enhanced Content: For instance, while testing marketing content, users must be shown a list of potential upgrades. The simple act of **developing, considering, and analyzing** these lists eliminates unproductive language and improves the usability of the final products for consumers.

2. Reduces Costs: Companies can save money by using A/B testing to find procedures that produce better results. One marketing effort will always be superior to the other; no two campaigns will ever yield comparable results.Businesses can use A/B testing to identify the option that provides **greater returns, eliminate the procedure that provides lower returns, and invest money** where it pays off more.

3. Low Risks: You can lower risks by using A/B tests. You can run an A/B test to observe how a new update or component on your product affects your system and how users respond to it if you're unsure of how it will perform. You may instantly roll back the code if it has a significant negative effect by utilizing a feature flag to run your A/B test.

4. More Engagement: The fact that 69 percent of businesses do A/B tests on emails is not surprising given that firms seek highly **engaged customers and followers**. Businesses can use it to determine the types of content that are most effective so they can focus more on those types.

Identifying Potential Data Sources:

Identifying potential data sources is a crucial step in any data science project or analysis. Having access to relevant, high-quality data is essential for drawing meaningful insights and making informed decisions.

Here are some common approaches and considerations for identifying potential data sources:

1. Internal Data Sources: This is data collected within your organization.

- Transactional data (e.g., sales records, purchase histories)
- Customer data (e.g., demographics, profiles, feedback)
- Operational data (e.g., logistics, production, inventory)
- Financial data (e.g., accounts, budgets, revenue streams)
- Web analytics data (e.g., website traffic, user behavior)
- Sensor data (e.g., IoT devices, equipment monitoring)

2. External Data Sources: This data comes from outside sources.

- Public data repositories (e.g., government websites, open data initiatives)
- Commercial data providers (e.g., market research firms, data brokers)
- Social media data (e.g., Twitter, Facebook, LinkedIn)
- Web scraping (e.g., extracting data from websites)
- Crowdsourced data (e.g., surveys, user-generated content)
- Affiliate or partner data (e.g., shared data from collaborators)

3. Data Format:

- **Structured Data:** This data is well-organized and follows a defined format, often stored in relational databases or spreadsheets. Examples include sales figures, customer demographics, or sensor readings.
- **Unstructured Data:** This data is less organized and can include text documents, emails, social media posts, images, audio, or video. Extracting insights from unstructured data often requires techniques like natural language processing (NLP) or computer vision.

4. Define the Problem or Objectives:

- Clearly define the problem you're trying to solve or the questions you're trying to answer.

- Identify the key variables or features that are relevant to your analysis.

- Understand the scope and context of your project.

5. Leverage Domain Expertise:

- Consult with subject matter experts within your organization or industry.

- Identify individuals or teams that work closely with the data or processes of interest.

- Gather insights and recommendations on potential data sources.

6. Conduct Literature Reviews or Market Research:

- Review academic publications, industry reports, or case studies related to your domain.

- Explore data sources or methodologies used in similar projects or analyses.

- Identify commonly used or recommended data sources for your field.

7. Evaluate Data Quality and Accessibility:

- Assess the completeness, accuracy, and reliability of potential data sources.

- Consider factors such as data format, documentation, and ease of integration.

- Evaluate any legal, ethical, or privacy concerns related to the data sources.

8. Explore Data Partnerships or Collaborations:

- Identify organizations or entities that may have access to relevant data.
- Investigate opportunities for data sharing or collaboration agreements.
- Engage with industry associations, research institutions, or community groups.

9. Leverage Existing Data Catalogs or Inventories:

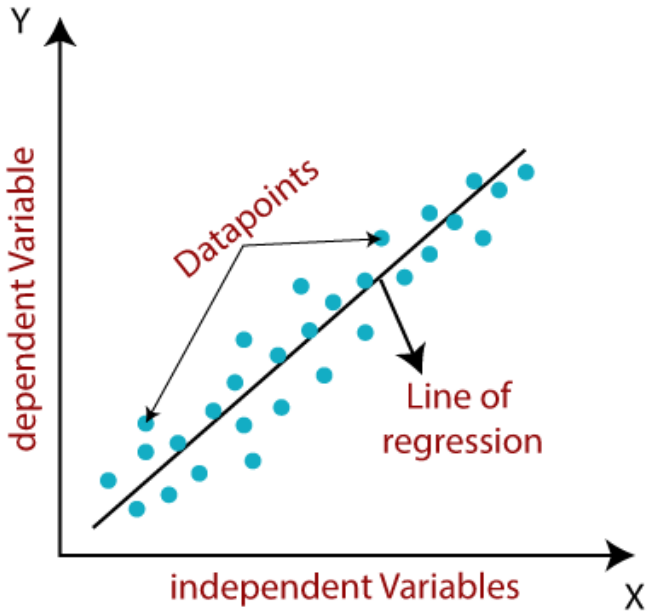
- Many organizations maintain data catalogs or inventories of their internal data sources.
- Consult with data governance teams or data stewards to access these resources.
- Identify potentially relevant datasets or data assets.

Linear Regression:

Linear Regression is a commonly used type of predictive analysis. Linear Regression is a statistical approach for modelling the relationship between a dependent variable and a given set of independent variables.

It is predicted that a straight line can be used to approximate the relationship. The goal of linear regression is to identify the line that minimizes the discrepancies between the observed data points and the line's anticipated values.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



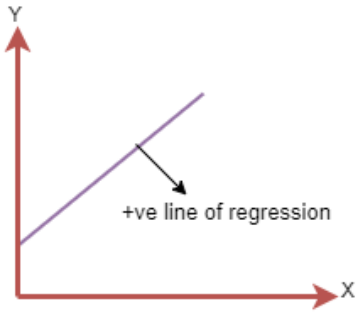
There are two types of linear regression -

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line:

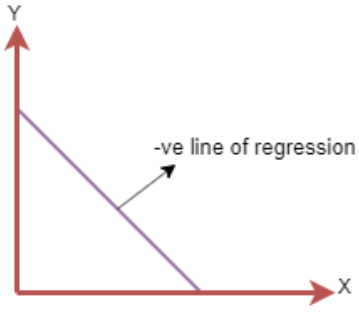
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:** If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- **Negative Linear Relationship:** If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

Simple Linear Regression:

Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.

The key point in Simple Linear Regression is that the ***dependent variable must be a continuous/real value***. However, the independent variable can be measured on continuous or categorical values.

Simple Linear regression algorithm has mainly two objectives:

- **Model the relationship between the two variables.** Such as the relationship between Income and expenditure, experience and Salary, etc.
- **Forecasting new observations.** Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

Simple Linear Regression Model:

The Simple Linear Regression model can be represented using the below equation:

$$y= a_0+a_1x+ \varepsilon$$

Where,

Y = Dependent Variable (Target Variable)

X = Independent Variable (predictor Variable)

a0 = intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Assumptions of Linear Regression:

1. Linearity: The relationship between the independent variable (x) and the dependent variable (y) is linear. This means that the change in the dependent variable is constant for a unit change in the independent variable. If the relationship is non-linear, linear regression may not provide an accurate model.

2. Independence of observations: The observations in the dataset should be independent of each other. This means that the value of one observation does not depend on the value of another observation.

3. Homoscedasticity: The variance of the residuals (the difference between the observed y values and the predicted y values) should be constant across all values of the independent variable (x). If the residuals have a pattern or trend, it violates the assumption of homoscedasticity, which can lead to biased estimates and invalid statistical inference.

4. Normality of residuals: The residuals should be normally distributed. This assumption is important for hypothesis testing and constructing confidence intervals. If the residuals are not normally distributed, the statistical inferences based on linear regression may be inaccurate.

5. No multicollinearity: In the case of multiple linear regression (when there are multiple independent variables), the independent variables should not be highly correlated with each other. High multicollinearity can lead to unstable and unreliable coefficient estimates.

6. No autocorrelation: The residuals should not be correlated with each other. Autocorrelation can occur in time series data or when observations are not independent, and it can lead to biased standard errors and incorrect inferences.

Multiple Linear Regression:

Multiple linear regression is used to estimate the relationship between **two or more independent variables** and **one dependent variable**. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Assumptions of multiple linear regression

Multiple linear regression makes all of the same assumptions as simple linear regression:

Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.

Normality: The data follows a normal distribution.

Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

Multiple Linear Regression Formula:

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The *t* statistic of the overall model.
- The associated *p* value (how likely it is that the *t* statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the *t* statistic and *p* value for each regression coefficient in the model.

Difference Between Simple Linear and Multiple Linear Regression:

Aspect	Simple Linear Regression	Multiple Linear Regression
Number of Independent Variables	One	Two or more
Model Equation	$y = \beta0 + \beta1x + \varepsilon$	$y = \beta0 + \beta1x1 + \beta2x2 + ... + \betanxn + \varepsilon$
Relationship Modeled	Relationship between one independent variable (x) and the dependent variable (y)	Relationship between multiple independent variables (x1, x2, ..., xn) and the dependent variable (y)
Interpretation of Coefficients	$\beta1$ represents the change in y for a unit change in x	$\betai$ represents the change in y for a unit change in xi, holding all other independent variables constant
Assumptions	Linearity, normality, homoscedasticity, independence of errors	Linearity, normality, homoscedasticity, independence of errors, no multicollinearity
Multicollinearity	Not applicable	Multicollinearity among independent variables should be avoided
Model Complexity	Simple and easy to interpret	More complex and may require careful interpretation
Explanatory Power	Limited by having only one independent variable	Potentially higher explanatory power with more independent variables
Applications	Suitable for simple relationships or when there is only one relevant independent variable	Suitable for complex relationships involving multiple factors or independent variables

Least Squares Principle:

The **least square method** is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve.

During the process of finding the relation between two variables, the trend of outcomes are estimated quantitatively. This process is termed as **regression analysis**. The method of curve fitting is an approach to regression analysis. This method of fitting equations which approximates the curves to given raw data is the least squares.

It is quite obvious that the fitting of curves for a particular data set are not always unique. Thus, it is required to find a curve having a minimal deviation from all the measured data points. This is known as the best-fitting curve and is found by using the least-squares method.

Least Square Method Formula:

The least-square method states that the curve that best fits a given set of observations, is said to be a curve having a minimum sum of the squared residuals (or deviations or errors) from the given data points. Let us assume that the given points of data are (x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n) in which all x's are independent variables, while all y's are dependent ones. Also, suppose that f(x) is the fitting curve and d represents error or deviation from each given point.

Now, we can write:

$d_1 = y_1 - f(x_1)$

$d_2 = y_2 - f(x_2)$

$d_3 = y_3 - f(x_3)$

…..

$d_n = y_n - f(x_n)$

The least-squares explain that the curve that best fits is represented by the property that the sum of squares of all the deviations from given values must be minimum, i.e:

$$S = \sum_{i=1}^{n} d_i^2$$

$$S = \sum_{i=1}^{n} [y_i - f_{x_i}]^2$$

$$S = d_1^2 + d_2^2 + d_3^2 + \cdots + d_n^2$$

Sum = Minimum Quantity

Suppose when we have to determine the equation of line of best fit for the given data, then we first use the following formula.

The equation of least square line is given by $Y = a + bX$

Normal equation for ‘a’:

$\sum Y = na + b\sum X$

Normal equation for ‘b’:

$\sum XY = a\sum X + b\sum X^2$

Solving these two normal equations we can get the required trend line equation.

Thus, we can get the line of best fit with formula $y = ax + b$

Question:

Consider the time series data given below:

x_i	8	3	2	10	11	3	6	5	6	8
y_i	4	12	1	12	9	4	9	6	1	14

Use the least square method to determine the equation of line of best fit for the data. Then plot the line.

Solution:

Mean of x_i values = $(8 + 3 + 2 + 10 + 11 + 3 + 6 + 5 + 6 + 8)/10 = 62/10 = 6.2$

Mean of y_i values = $(4 + 12 + 1 + 12 + 9 + 4 + 9 + 6 + 1 + 14)/10 = 72/10 = 7.2$

Straight line equation is $y = a + bx$.

The normal equations are

$\sum y = an + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

x	y	x^2	xy
8	4	64	32
3	12	9	36
2	1	4	2
10	12	100	120
11	9	121	99
3	4	9	12
6	9	36	54
5	6	25	30
6	1	36	6
8	14	64	112
$\sum x = 62$	$\sum y = 72$	$\sum x^2 = 468$	$\sum xy = 503$

Substituting these values in the normal equations,

$10a + 62b = 72$….(1)

$62a + 468b = 503$….(2)

$(1) \times 62 – (2) \times 10$,

$620a + 3844b – (620a + 4680b) = 4464 – 5030$

$-836b = -566$

$b = 566/836$

$b = 283/418$

$b = 0.677$

Substituting $b = 0.677$ in equation (1),

$10a + 62(0.677) = 72$

$10a + 41.974 = 72$

$10a = 72 – 41.974$

10a = 30.026

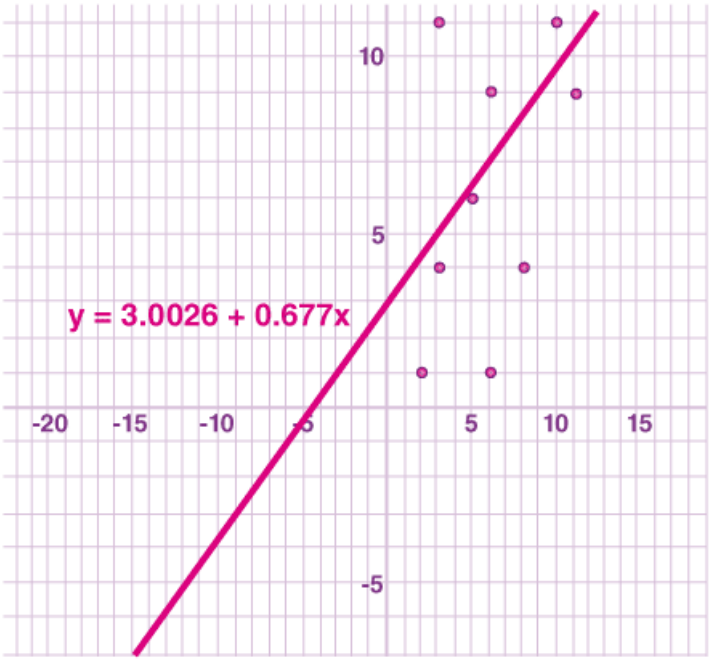
a = 30.026/10

a = 3.0026

Therefore, the equation becomes,

y = a + bx

y = 3.0026 + 0.677x



This is the required trend line equation.

Now, we can find the sum of squares of deviations from the obtained values as:

d_1 = [4 – (3.0026 + 0.677*8)] = (-4.4186)

d_2 = [12 – (3.0026 + 0.677*3)] = (6.9664)

d_3 = [1 – (3.0026 + 0.677*2)] = (-3.3566)

d_4 = [12 – (3.0026 + 0.677*10)] = (2.2274)

d_5 = [9 – (3.0026 + 0.677*11)] =(-1.4496)

d_6 = [4 – (3.0026 + 0.677*3)] = (-1.0336)

d_7 = [9 – (3.0026 + 0.677*6)] = (1.9354)

d_8 = [6 – (3.0026 + 0.677*5)] = (-0.3876)

d_9 = [1 – (3.0026 + 0.677*6)] = (-6.0646)

$d_{10} = [14 – (3.0026 + 0.677*8)] = (5.5814)$

$\sum d^2 = (-4.4186)^2 + (6.9664)^2 + (-3.3566)^2 + (2.2274)^2 + (-1.4496)^2 + (-1.0336)^2 + (1.9354)^2 + (-0.3876)^2 + (-6.0646)^2 + (5.5814)^2 = 159.27990$

Limitations for Least-Square Method:

The least-squares method is a very beneficial method of curve fitting. Despite many benefits, it has a few shortcomings too. One of the main limitations is discussed here.

In the process of regression analysis, which utilizes the least-square method for curve fitting, it is inevitably assumed that the errors in the independent variable are negligible or zero. In such cases, when independent variable errors are non-negligible, the models are subjected to measurement errors. Therefore, here, the least square method may even lead to hypothesis testing, where parameter estimates and confidence intervals are taken into consideration due to the presence of errors occurring in the independent variables.

Exploratory vs Inferential Viewpoints:

Exploratory Viewpoint:

The exploratory viewpoint is primarily concerned with understanding and gaining insights from data through visual and analytical techniques. It is often the first step in a data analysis process and is used to uncover patterns, trends, and relationships within the data.

The key aspects of the exploratory viewpoint include:

- **Data visualization:** Creating various plots, charts, and graphical representations to visually explore the data and identify potential patterns or outliers.
- **Descriptive statistics:** Calculating summary statistics, such as mean, median, standard deviation, and quantiles, to summarize and describe the characteristics of the data.
- **Unsupervised learning:** Applying techniques like clustering, dimensionality reduction, and association rule mining to discover inherent structures or relationships within the data without relying on predefined target variables.
- **Data cleaning and preparation:** Identifying and handling missing values, outliers, and inconsistencies in the data to ensure its quality and suitability for further analysis.

The exploratory viewpoint is particularly useful when dealing with new or unfamiliar datasets, as it helps researchers understand the data, generate hypotheses, and formulate questions for further investigation.

Examples:

- Plotting the distribution of a variable to check for normality.
- Creating scatter plots to identify potential correlations between variables.
- Using box plots to compare distributions across different categories.

Inferential Viewpoint:

The inferential viewpoint aims to draw conclusions or make predictions about a broader population or process based on a sample of data. It involves using statistical methods and models to test hypotheses, estimate parameters, and quantify the uncertainty associated with the findings.

The key aspects of the inferential viewpoint include:

- **Hypothesis testing:** Formulating null and alternative hypotheses, and using statistical tests (e.g., t-tests, ANOVA, chi-square tests) to evaluate the evidence against the null hypothesis.
- **Estimation and confidence intervals:** Estimating population parameters (e.g., means, proportions, regression coefficients) and constructing confidence intervals to quantify the uncertainty around these estimates.
- **Predictive modeling:** Building and evaluating models (e.g., regression, classification, time series) to predict or forecast future values or outcomes based on the available data.
- **Supervised learning:** Applying machine learning algorithms to train models on labeled data, with the goal of making accurate predictions or classifications on new, unseen data.
- **Experimental design:** Designing and analyzing controlled experiments to establish causal relationships and evaluate the effects of interventions or treatments.

The inferential viewpoint is crucial when the goal is to make generalizations or draw conclusions that extend beyond the specific dataset being analyzed, and when quantifying the uncertainty associated with these conclusions is important.

Examples:

- Conducting a t-test to compare the means of two groups.
- Estimating the proportion of a population that exhibits a certain characteristic using a sample proportion and constructing a confidence interval around it.
- Using regression analysis to infer the relationship between independent and dependent variables.

It's important to note that these viewpoints are not mutually exclusive; they are often used in a complementary manner. The exploratory viewpoint can inform the inferential viewpoint by identifying potential relationships or patterns that can be further investigated using inferential methods. Conversely, the inferential viewpoint can guide the exploratory process by suggesting specific hypotheses or areas of interest to explore in the data.

Effective data science projects often involve an iterative cycle of exploration and inference, combining the strengths of both viewpoints to gain a comprehensive understanding of the data and draw meaningful conclusions.

Aspect	**Exploratory Data Analysis (EDA)**	**Inferential Data Analysis**
Purpose	Understand the data set, identify patterns, anomalies, and generate hypotheses.	Make generalizations about a population, test hypotheses, and estimate reliability.
Objective	Data summarization, visualization, and pattern discovery.	Hypothesis testing, parameter estimation, and drawing conclusions about a population.
Methods	Descriptive statistics, various plots (histograms, scatter plots, box plots).	Statistical tests (t-tests, chi-square tests, ANOVA), regression analysis.
Assumptions	Non-parametric; fewer assumptions about data distribution.	Often relies on specific assumptions (e.g., normality, independence).
Flexibility	Highly flexible, iterative, and informal.	More formal and structured with predefined procedures.
Hypothesis	Generation of hypotheses.	Testing and validation of hypotheses.
Visualization	Extensive use of graphical representations.	Less emphasis on visualization, more on numerical results.
Examples	Plotting distributions, identifying correlations, detecting outliers.	Conducting significance tests, constructing confidence intervals, estimating population parameters.
Outcome	Insights, patterns, potential hypotheses.	Statistical evidence, generalizations, and conclusions.
Typical Questions	What does the data look like? Are there any patterns or outliers?	Are the observed effects statistically significant? What is the estimated parameter for the population?

Model Generalizability:

Model generalizability refers to the ability of a predictive model to perform well on new, unseen data, not just on the data used to create the model. A model with good generalizability effectively captures the underlying patterns in the data rather than just memorizing the training data (a problem known as overfitting).

Key Aspects of Model Generalizability:

1. **Training vs. Testing Performance**:
   - **Training Performance**: How well the model performs on the data it was trained on.
   - **Testing Performance**: How well the model performs on a separate set of data that it hasn't seen before (test data).
2. **Overfitting**:
   - **Definition**: A model is overfitting if it performs well on the training data but poorly on the test data.
   - **Symptoms**: High accuracy on training data and low accuracy on test data.
   - **Causes**: Too complex a model relative to the amount of training data (e.g., too many parameters).
3. **Underfitting**:
   - **Definition**: A model is underfitting if it performs poorly on both the training and test data.
   - **Symptoms**: Low accuracy on both training and test data.
   - **Causes**: Too simple a model to capture the underlying patterns in the data.

Strategies to Improve Model Generalizability:

1. **Cross-Validation**:
   - **K-Fold Cross-Validation**: The data is divided into K subsets, and the model is trained K times, each time using a different subset as the test set and the remaining as the training set.
   - **Leave-One-Out Cross-Validation**: A special case of K-fold where K equals the number of data points. Each point is used once as the test set.

2. **Regularization**:
    - **L1 Regularization (Lasso)**: Adds a penalty equal to the absolute value of the magnitude of coefficients.
    - **L2 Regularization (Ridge)**: Adds a penalty equal to the square of the magnitude of coefficients.
    - **Elastic Net**: Combines both L1 and L2 regularization.
3. **Model Complexity**:
    - **Simpler Models**: Start with a simple model and gradually increase complexity.
    - **Pruning**: In decision trees, remove sections of the tree that provide little power to classify instances.
4. **Feature Selection**:
    - **Removing Irrelevant Features**: Exclude features that do not contribute to the model's predictive power.
    - **Dimensionality Reduction**: Techniques like PCA (Principal Component Analysis) to reduce the number of features while retaining variance.
5. **Ensemble Methods**:
    - **Bagging**: Combines predictions from multiple models to reduce variance (e.g., Random Forest).
    - **Boosting**: Sequentially builds models to correct errors from previous models (e.g., Gradient Boosting Machines).
6. **Validation and Test Splits**:
    - **Hold-Out Validation**: Splitting the data into training, validation, and test sets.
    - **Nested Cross-Validation**: Useful for hyperparameter tuning, where an outer loop cross-validation is used to assess the model and an inner loop is used to tune parameters.

Metrics to Assess Generalizability:

- **Accuracy**: Percentage of correct predictions.
- **Precision and Recall**: Useful for imbalanced datasets.
- **F1 Score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Area under the Receiver Operating Characteristic curve.
- **Mean Squared Error (MSE)**: For regression tasks.

Importance of Generalizability:

- **Real-World Performance**: Ensures the model will perform well on new, unseen data, which is crucial for real-world applications.
- **Reliability**: A model that generalizes well is more reliable and trustworthy.
- **Avoiding Overfitting**: Prevents the model from capturing noise or random fluctuations in the training data.

Cross Validation:

Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. ***We can also say that it is a technique to check how a statistical model generalizes to an independent dataset***.

In machine learning, there is always the need to test the stability of the model. It means based only on the training dataset; we can't fit our model on the training dataset. For this purpose, we reserve a particular sample of the dataset, which was not part of the training dataset. After that, we test our model on that sample before deployment, and this complete process comes under cross-validation. This is something different from the general train-test split.

Hence the basic steps of cross-validations are:

- Reserve a subset of the dataset as a validation set.
- Provide the training to the model using the training dataset.
- Now, evaluate model performance using the validation set. If the model performs well with the validation set, perform the further step, else check for the issues.

Methods used for Cross-Validation:

There are some common methods that are used for cross-validation. These methods are given below:

1. **Validation Set Approach**
2. **Leave-P-out cross-validation**
3. **Leave one out cross-validation**
4. **K-fold cross-validation**
5. **Stratified k-fold cross-validation**

1. Validation Set Approach:

We divide our input dataset into a training set and test or validation set in the validation set approach. Both the subsets are given 50% of the dataset.

But it has one of the big disadvantages that we are just using a 50% dataset to train our model, so the model may miss out to capture important information of the dataset. It also tends to give the underfitted model.

2. Leave-P-Out Cross-Validation:

In this approach, the p datasets are left out of the training data. It means, if there are total n datapoints in the original input dataset, then n-p data points will be used as the training dataset and the p data points as the validation set. This complete process is repeated for all the samples, and the average error is calculated to know the effectiveness of the model.

There is a disadvantage of this technique; that is, it can be computationally difficult for the large p.

3. Leave One Out Cross-Validation (LOOCV):

This method is similar to the leave-p-out cross-validation, but instead of p, we need to take 1 dataset out of training. It means, in this approach, for each learning set, only one datapoint is reserved, and the remaining dataset is used to train the model. This process repeats for each datapoint. Hence for n samples, we get n different training set and n test set.

It has the following features:

- In this approach, the bias is minimum as all the data points are used.
- The process is executed for n times; hence execution time is high.
- This approach leads to high variation in testing the effectiveness of the model as we iteratively check against one data point.

4. K-Fold Cross-Validation:

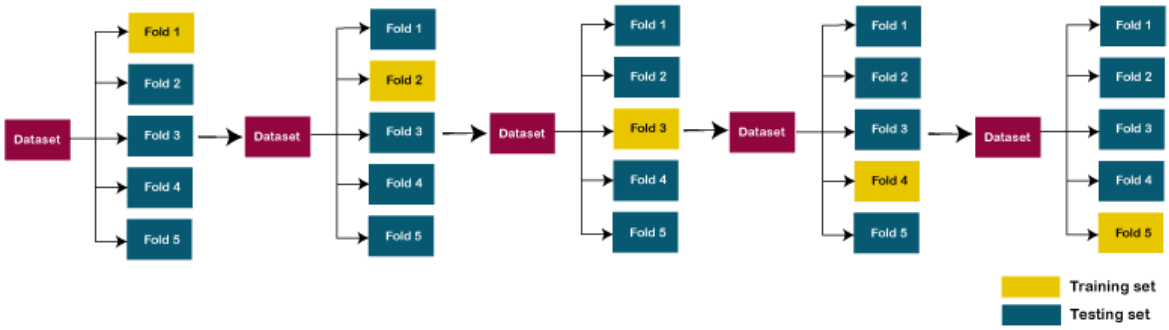
K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called **folds**. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set. This approach is a very popular CV approach because it is easy to understand, and the output is less biased than other methods.

The steps for k-fold cross-validation are:

- Split the input dataset into K groups
- For each group:
  - Take one group as the reserve or test data set.
  - Use remaining groups as the training dataset
  - Fit the model on the training set and evaluate the performance of the model using the test set.

Let's take an example of 5-folds cross-validation. So, the dataset is grouped into 5 folds. On 1st iteration, the first fold is reserved for test the model, and rest are used to train the model. On 2nd iteration, the second fold is used to test the model, and rest are used to train the model. This process will continue until each fold is not used for the test fold.

Consider the below diagram:



5. Stratified k-Fold Cross-Validation:

This technique is similar to k-fold cross-validation with some little changes. This approach works on stratification concept, it is a process of rearranging the data to ensure that each fold or group is a good representative of the complete dataset. To deal with the bias and variance, it is one of the best approaches.

It can be understood with an example of housing prices, such that the price of some houses can be much high than other houses. To tackle such situations, a stratified k-fold cross-validation technique is useful.

6. Holdout Method:

This method is the simplest cross-validation technique among all. In this method, we need to remove a subset of the training data and use it to get prediction results by training it on the rest part of the dataset.

The error that occurs in this process tells how well our model will perform with the unknown dataset. Although this approach is simple to perform, it still faces the issue of high variance, and it also produces misleading results sometimes.

Comparison of Cross-validation to train/test split:

- **Train/test split:** The input data is divided into two parts, that are training set and test set on a ratio of 70:30, 80:20, etc. It provides a high variance, which is one of the biggest disadvantages.
    - **Training Data:** The training data is used to train the model, and the dependent variable is known.
    - **Test Data:** The test data is used to make the predictions from the model that is already trained on the training data. This has the same features as training data but not the part of that.
- **Cross-Validation dataset:** It is used to overcome the disadvantage of train/test split by splitting the dataset into groups of train/test splits, and averaging the result. It can be used if we want to optimize our model that has been trained on the training dataset for the best performance. It is more efficient as compared to train/test split as every observation is used for the training and testing both.

Limitations of Cross-Validation:

- For the ideal conditions, it provides the optimum output. But for the inconsistent data, it may produce a drastic result. So, it is one of the big disadvantages of cross-validation, as there is no certainty of the type of data in machine learning.
- In predictive modeling, the data evolves over a period, due to which, it may face the differences between the training set and validation sets. Such as if we create a model for the prediction of stock market values, and the data is trained on the previous 5 years stock values, but the realistic future values for the next 5 years may drastically different, so it is difficult to expect the correct output for such situations.

Applications of Cross-Validation:

- This technique can be used to compare the performance of different predictive modeling methods.
- It has great scope in the medical research field.
- It can also be used for the meta-analysis, as it is already being used by the data scientists in the field of medical statistics.

Using Categorical Variables in Regression:

Categorical Variables are variables that can take on one of a limited and fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property. They are also known as a **factor, nominal or qualitative variables**.

The type of regression analysis that fits best with categorical variables is Logistic Regression. Logistic regression uses Maximum Likelihood Estimation to estimate the parameters. It derives the relationship between a set of variables(independent) and a categorical variable(dependent).

To include categorical variables in regression models, they need to be appropriately encoded or transformed into a suitable format.

There are several approaches to incorporating categorical variables in regression models:

1. One-Hot Encoding:

- One-hot encoding is a widely used technique for encoding categorical variables.

- It creates binary dummy variables for each category, where a value of 1 represents the presence of that category, and 0 represents its absence.

- For example, if a variable "Color" has three categories (Red, Green, Blue), it would be encoded as three binary variables: Color_Red, Color_Green, and Color_Blue.

- One category is typically dropped to avoid the dummy variable trap (multicollinearity issue).

- One-hot encoding allows the model to capture the non-linear effects of categorical variables.

2. Ordinal Encoding:

- If the categorical variable has an inherent order or ranking (e.g., education levels: high school, bachelor's, master's, Ph.D.), ordinal encoding can be used.

- In ordinal encoding, each category is assigned a unique integer value based on its order or rank.

- This approach assumes a linear relationship between the ordinal categories and the dependent variable.

3. Contrast Coding:

- Contrast coding is an alternative to one-hot encoding that can be useful when comparing specific categories or groups.

- It creates new variables that represent the contrasts or differences between categories.

- Common contrast coding schemes include treatment contrasts, sum-to-zero contrasts, and polynomial contrasts.

- Contrast coding can provide more interpretable coefficients and facilitate hypothesis testing regarding group differences.

4. Indicator/Dummy Variable Approach:

- This approach involves creating a single binary variable for each category, with one category serving as the reference or baseline category.

- The coefficients of the dummy variables represent the difference between the corresponding category and the reference category.

- This method is suitable when there is a natural reference category or when the focus is on comparing each category to a baseline.

5. Simple Coding: Similar to dummy coding, simple coding creates binary variables for each level of the categorical variable, but it compares each level directly to a specified reference level, rather than the omitted level.

6. Deviation Coding: In deviation coding, the coefficients represent the deviation of each level from the overall mean or grand mean of the dependent variable. The name "deviation" refers to the fact that the coefficients represent deviations from the grand mean.

7. Difference Coding: This method compares each level of the categorical variable to the mean of the preceding levels. The name "difference" comes from the fact that the coefficients represent the difference between a level and the mean of the previous levels.

8. Helmert Coding: Helmert coding compares each level of the categorical variable to the mean of the remaining levels. The name "Helmert" refers to the German mathematician and physicist, Friedrich Robert Helmert, who introduced this coding method.

9. Orthogonal Polynomial Coding: This method uses orthogonal polynomials to code the levels of the categorical variable. The name "orthogonal polynomial" refers to the use of polynomial functions that are orthogonal or uncorrelated with each other.

10. Repeated Coding: In repeated coding, each level of the categorical variable is compared to the adjacent levels. The name "repeated" comes from the fact that the contrasts are repeated for each pair of adjacent levels.

11. Special User-Defined Coding: This method allows users to specify their own custom contrasts or comparisons between levels of the categorical variable. The name "user-defined" indicates that the contrasts are defined by the user based on their specific research questions or hypotheses.

Once the categorical variables are encoded or transformed, they can be included in the regression model along with the continuous variables. It's important to note that the interpretation of the regression coefficients for categorical variables differs from that of continuous variables.

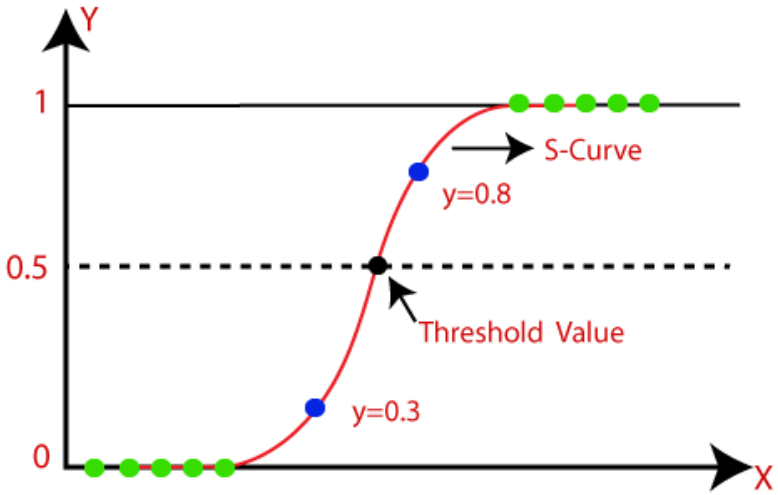
Additionally, it is advisable to check for potential multicollinearity issues when including multiple categorical variables or their interactions in the regression model. Techniques like variance inflation factors (VIF) can be used to detect and address multicollinearity.

The choice of encoding or transformation method for categorical variables depends on the nature of the variable, the assumptions of the regression model, and the specific research questions or goals of the analysis. It's also essential to consider the interpretability and domain knowledge when working with categorical variables in regression models.

Logistic Regression:

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions of Logistic Regression:

We will explore the assumptions of logistic regression as understanding these assumptions is important to ensure that we are using appropriate application of the model. The assumption include:

1. Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.
2. Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories SoftMax functions are used.
3. Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.
4. No outliers: There should be no outliers in the dataset.
5. Large sample size: The sample size is sufficiently large

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Correlation:

If two quantities vary in such a way that movements in one are accompanied by movements in the other, these quantities are correlated. For example, there exists some relationship between age of husband and age of wife.

The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.

Use: Most of the variables show some kind of relationship, for example there is relationship between price and supply, Income and expenditure etc. with the help of correlation analysis we can measure in one figure the degree of relationship existing between the variables.

Types of correlation:

1. Positive or negative
2. Simple, partial and Multiple
3. Linear and non-linear

Multiple correlation:

Multiple correlation is a statistical concept that measures the degree of linear association between a dependent variable and two or more independent variables in a multiple regression analysis. It is an extension of the simple correlation coefficient, which measures the linear relationship between two variables.

Multiple Correlation Coefficient (R):

The multiple correlation coefficient R measures the correlation between the observed values of YY and the values predicted by the multiple regression model.

Given variables x, y, and z, we define the **multiple correlation coefficient**

$$R_{z,xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$

where r_{xz}, r_{yz}, r_{xy} are defined as

$$r = cov(x, y)/s_x s_y$$

.

Here x and y are viewed as the independent variables and z is the dependent variable.

It ranges from 0 to 1:

- R=0: No linear relationship between the dependent variable and the independent variables.
- R=1: Perfect linear relationship.

Coefficient of Determination (R^2):

We also define the **multiple coefficient of determination** to be the square of the multiple correlation coefficient.

Often the subscripts are dropped and the multiple correlation coefficient and multiple coefficient of determination are written simply as R and R^2 respectively. These definitions may also be expanded to more than two independent variables. With just one independent variable the multiple correlation coefficient is simply r.

Unfortunately, R is not an unbiased estimate of the population multiple correlation coefficient, which is evident for small samples. A relatively unbiased version of R is given by R adjusted.

If R is $R_{z,xy}$ as defined above (or similarly for more variables) then the **adjusted** multiple coefficient of determination is

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where k = the number of independent variables and n = the number of data elements in the sample for z (which should be the same as the samples for x and y).

It ranges from 0 to 1:

- R^2=0: The model does not explain any of the variance in the dependent variable.
- R^2=1: The model explains all the variance in the dependent variable.

Example 1: From the following data, obtain $R_{1.23}$ and $R_{2.13}$

X_1	65	72	54	68	55	59	78	58	57	51
X_2	56	58	48	61	50	51	55	48	52	42
X_3	9	11	8	13	10	8	11	10	11	7

Solution: To obtain multiple correlation coefficients $R_{1.23}$ and $R_{2.13}$, we use following formulae

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \quad \text{and}$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}}$$

We need r_{12}, r_{13} and r_{23} which are obtained from the following table:

S. No.	X_1	X_2	X_3	$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	X_1X_2	X_1X_3	X_2X_3
1	65	56	9	4225	3136	81	3640	585	504
2	72	58	11	5184	3364	121	4176	792	638
3	54	48	8	2916	2304	64	2592	432	384
4	68	61	13	4624	3721	169	4148	884	793
5	55	50	10	3025	2500	100	2750	550	500
6	59	51	8	3481	2601	64	3009	472	408
7	78	55	11	6084	3025	121	4290	858	605
8	58	48	10	3364	2304	100	2784	580	480
9	57	52	11	3249	2704	121	2964	627	572
10	51	42	7	2601	1764	49	2142	357	294
Total	617	521	98	38753	27423	990	32495	6137	5178

Now we get the total correlation coefficient r_{12}, r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_2^2) - (\sum X_2)^2\}}}$$

$$r_{12} = \frac{(10 \times 32495) - (617) \times (521)}{\sqrt{\{(10 \times 38753) - (617) \times (617)\}\{(10 \times 27423) - (521) \times (521)\}}}$$

$$r_{12} = \frac{3493}{\sqrt{\{6841\} \times \{2789\}}} = \frac{3493}{4368.01} = 0.80$$

$$r_{13} = \frac{N(\sum X_1 X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{13} = \frac{(10 \times 6137) - (617) \times (98)}{\sqrt{\{(10 \times 38753) - (617 \times 617)\}\{(10 \times 990) - (98 \times 98)\}}}$$

$$r_{13} = \frac{904}{\sqrt{\{6841\}\{296\}}} = \frac{904}{1423.00} = 0.64$$

and

$$r_{23} = \frac{N(\sum X_2 X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{23} = \frac{(10 \times 5178) - (521) \times (98)}{\sqrt{\{(10 \times 27423) - (521 \times 521)\}\{(10 \times 990) - (98 \times 98)\}}}$$

$$r_{23} = \frac{722}{\sqrt{\{2789\}\{296\}}} = \frac{722}{908.59} = 0.79$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.80$, $r_{13} = 0.64$ and $r_{23} = 0.79$, then

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.80^2 + 0.64^2 - 2 \times 0.80 \times 0.64 \times 0.79}{1 - 0.79^2}$$

$$= \frac{0.64 + 0.41 - 0.81}{1 - 0.62}$$

$$R_{1.23}^2 = \frac{0.24}{0.38} = 0.63$$

Then

$$R_{1.23} = 0.79.$$

$$R_{2.13}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$$

$$= \frac{0.80^2 + 0.79^2 - 2 \times 0.80 \times 0.64 \times 0.79}{1 - 0.64^2}$$

$$= \frac{0.64 + 0.62 - 0.81}{1 - 0.49}$$

$$= \frac{0.45}{0.51} = 0.88$$

Thus,

$$R_{2.13} = 0.94$$

Partial Correlation:

Partial correlation is a statistical concept that measures the degree of association between two variables while controlling for or removing the effect of one or more additional variables. It allows researchers to examine the relationship between two variables after accounting for the influence of other variables that may be related to both.

The partial correlation coefficient is denoted by the symbol "r" with subscripts indicating the variables involved and a dot (.) separating the variables being partialed out or controlled for. For example, rxy.z represents the partial correlation between variables x and y, controlling for the effect of variable z.

The formula for calculating the partial correlation coefficient between two variables x and y, controlling for the effects of a third variable z, is:

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

where:

- r_{XY} is the correlation coefficient between X and Y.
- r_{XZ} is the correlation coefficient between X and Z.
- r_{YZ} is the correlation coefficient between Y and Z.

Interpretation

- $r_{XY.Z} = 0$: No partial correlation between X and Y when controlling for Z.
- $r_{XY.Z} > 0$: Positive partial correlation between X and Y when controlling for Z.
- $r_{XY.Z} < 0$: Negative partial correlation between X and Y when controlling for Z.

Assumptions for Partial Correlation:

Continuous: The variable that you care about must be continuous. Continuous means that the variable can take on any reasonable value.

Some good examples of continuous variables include age, weight, height, test scores, survey scores, yearly salary, etc.

Normally Distributed: The variable that you care about must be spread out in a normal way. In statistics, this is called being normally distributed (aka it must look like a bell curve when you graph the data). Only use an independent samples t-test with your data if the variable you care about is normally distributed.

Linearity: The variables that you care about must be related linearly. This means that if you plot the variables, you will be able to draw a straight line that fits the shape of the data.

No Outliers: The variables that you care about must not contain outliers. Pearson's correlation is sensitive to outliers, or data points that have unusually large or small values. You can tell if your variables have outliers by plotting them and observing if any points are far from all other points.

Similar Spread Across Range: In statistics this is called homoscedasticity, or making sure the variables have a similar spread across their ranges.

Covariate(s): You should only perform partial correlation if you have one or more covariates. A covariate is a variable whose effects you want to remove when examining the variable relationship of interest. For instance, if you're examining the relationship between age and memory performance, you may be interested in removing the effects of education level. This way, you can be sure that education level isn't influencing the results.

When to use Partial Correlation?

You should use Partial Correlation in the following scenario:

1. Relationship: You are looking for a statistical test to look at how two variables are related. Other types of analyses include testing for a difference between two variables or predicting one variable using another variable (prediction).

Continuous Data: Your variable of interest must be continuous. Continuous means that your variable of interest can basically take on any value, such as heart rate, height, weight, number of ice cream bars you can eat in 1 minute, etc.

Types of data that are NOT continuous include ordered data (such as finishing place in a race, best business rankings, etc.), categorical data (gender, eye color, race, etc.), or binary data (purchased the product or not, has the disease or not, etc.).

Two Groups: Pearson Correlation can only be used to compare two groups on your variable of interest.

If you have three or more groups, you should use [clustering] or [distance metrics] instead.