***Assignment # 4***

***Faraz Ahmed***

***1. Hypothesis:***

A hypothesis is a proposed explanation or statement that can be tested through experimentation or observation. It is a fundamental concept in the scientific method and is used to make predictions about the outcome of experiments or observations. A hypothesis typically takes the form of a testable statement that can be either supported or refuted based on empirical evidence.

***In-Simple Words:***

A hypothesis is a concept or idea that you test through research and experiments. In other words, it is a prediction that is can be tested by research. Most researchers come up with a hypothesis statement at the beginning of the study

***With Example:***

A simple hypothesis suggests only the relationship between two variables: one independent and one dependent.

**Examples:** If you stay up late, then you feel tired the next day. Turning off your phone makes it charge faster.

There are two main types of hypotheses:

1. **Null Hypothesis (H0):** This hypothesis suggests that there is no significant relationship or effect between the variables being studied. In other words, it assumes that any observed differences or correlations are due to chance or random variation. Researchers typically try to disprove the null hypothesis to establish the presence of a meaningful relationship or effect.
2. **Alternative Hypothesis (Ha or H1):** The alternative hypothesis is the opposite of the null hypothesis. It proposes that there is a significant relationship or effect between the variables under investigation. Researchers aim to provide evidence in support of the alternative hypothesis to demonstrate a meaningful finding.

Hypotheses are essential in the scientific process because they guide research and experimentation. Researchers formulate hypotheses based on prior knowledge, observations, or theories, and then design experiments or studies to test these hypotheses. The results of these tests can either support the hypothesis, leading to further understanding, or reject the hypothesis, prompting the need for new explanations or theories.

It is important to note that a hypothesis is not a definitive statement of fact but rather a tentative and testable proposition. Scientific inquiry involves gathering empirical evidence to either support or reject a hypothesis, which can then contribute to the advancement of knowledge in a particular field.

***Characteristics of Hypothesis***

Following are the characteristics of the hypothesis:

* The hypothesis should be clear and precise to consider it reliable.
* If the hypothesis is a relational hypothesis, then it should be stating the relationship between variables.
* The hypothesis must be specific and should have scope for conducting more tests.
* The way of explanation of the hypothesis must be very simple and it should be understood that the simplicity of the hypothesis is not related to its significance.

***Sources of Hypothesis***

Following are the sources of hypothesis:

* The resemblance between the phenomenon.
* Observations from past studies, present-day experiences and from the competitors.
* Scientific theories.
* General patterns that influence the thinking process of people.

***Types of Hypothesis***

There are six forms of hypothesis and they are:

* Simple hypothesis
* Complex hypothesis
* Directional hypothesis
* Non-directional hypothesis
* Null hypothesis
* Associative and casual hypothesis

***Simple Hypothesis***

It shows a relationship between one dependent variable and a single independent variable. For example – If you eat more vegetables, you will lose weight faster. Here, eating more vegetables is an independent variable, while losing weight is the dependent variable.

***Complex Hypothesis***

It shows the relationship between two or more dependent variables and two or more independent variables. Eating more vegetables and fruits leads to weight loss, glowing skin, and reduces the risk of many diseases such as heart disease.

***Directional Hypothesis***

It shows how a researcher is intellectual and committed to a particular outcome. The relationship between the variables can also predict its nature. For example- children aged four years eating proper food over a five-year period are having higher IQ levels than children not having a proper meal. This shows the effect and direction of the effect.

***Non-directional Hypothesis***

It is used when there is no theory involved. It is a statement that a relationship exists between two variables, without predicting the exact nature (direction) of the relationship.

***Null Hypothesis***

It provides a statement, which is contrary to the hypothesis. It is a negative statement, and there is no relationship between independent and dependent variables. “HO” denotes the symbol.

***Associative and Causal Hypothesis***

Associative hypothesis occurs when there is a change in one variable resulting in a change in the other variable. Whereas, the causal hypothesis proposes a cause and effect interaction between two or more variables.

***Examples of Hypothesis***

Following are the examples of hypotheses based on their types:

* Consumption of sugary drinks every day leads to obesity is an example of a simple hypothesis.
* All lilies have the same number of petals is an example of a null hypothesis.
* If a person gets 7 hours of sleep, then he will feel less fatigue than if he sleeps less. It is an example of a directional hypothesis.

***Functions of Hypothesis***

Following are the functions performed by the hypothesis:

* Hypothesis helps in making an observation and experiments possible.
* It becomes the start point for the investigation.
* Hypothesis helps in verifying the observations.

***2. Difference between t and z test in regression:***

**Z-tests and T-tests** are the two statistical methods that involve data analysis, which has applications in science, business, and many other disciplines. The T-test is a univariate hypothesis test based on T-statistics, wherein the mean, i.e., the average, is known, and population variance, i.e., the standard deviation, is approximated from the sample. On the other hand, Z-test is also a univariate test based on a [standard normal distribution](https://www.wallstreetmojo.com/standard-normal-distribution-formula/).

Z-test is used to test a Null Hypothesis if the population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population variance is unknown.

1. **Sample Size and Population Variance:**

* **T-test:** T-tests are suitable when you have a relatively small sample size (typically less than 30) and when the population variance is unknown. T-tests use the sample standard error to estimate the standard error of the coefficient.
* **Z-test:** Z-tests are used when you have a large sample size (typically greater than 30) and when the population variance is known. They rely on the known population standard deviation to estimate the standard error of the coefficient. However, in practical regression analysis, knowledge of the population variance is rare.

1. **Test Statistic Formulas:**

* **T-test:** The t-test statistic is calculated as (estimate - population parameter) / (standard error), where the standard error is estimated from the sample.
* **Z-test:** The z-test statistic is calculated in a similar way, but it uses the known population standard deviation to estimate the standard error.

1. **Distribution of Test Statistic:**

* **T-test:** The t-test statistic follows a Student's t-distribution, which has heavier tails compared to the normal distribution. The shape of the t-distribution depends on the degrees of freedom, which is related to the sample size.
* **Z-test:** The z-test statistic follows a standard normal distribution (normal distribution with a mean of 0 and a standard deviation of 1).

1. **Common Use Cases:**

* **T-test:** T-tests are commonly used in regression analysis, particularly for testing the significance of individual coefficients (e.g., slope coefficients). They are well suited when dealing with small to moderately sized samples.
* **Z-test:** Z-tests are less common in regression analysis because they require knowledge of the population standard deviation, which is often not available in practice.

1. **Robustness:**

* **T-test:** T-tests are more robust when dealing with small sample sizes or when the population variance is unknown because they rely on the sample standard deviation.
* **Z-test:** Z-tests are more suitable when dealing with large samples and known population variances, but this situation is less common in practice.

|  |  |  |
| --- | --- | --- |
| **Points** | **T-Test** | **Z-Test** |
| **1. Purpose** | Compare means of small samples (n < 30) | Compare means of large samples (n ≥ 30) |
| **2. Assumptions** | Normally distributed data, approximate normality | Normally distributed data, known population standard deviation |
| **3. Population Standard Deviation** | Unknown | Known |
| **4. Sample Size** | Small (n < 30) | Large (n ≥ 30) |
| **5. Test Statistic** | T-distribution | Standard normal distribution (Z-distribution) |
| **6. Degrees of Freedom** | n1 + n2 - 2 | Not applicable |
| **7. Use Case** | Small sample analysis, comparing means between groups | Large sample analysis, population mean comparisons |
| **8. One-sample vs. Two-sample** | Both | Usually two-sample |
| **9. Data Requirement** | Raw data | Raw data |
| **10. Complexity** | Relatively more complex | Relatively simpler |

***Conclusion:***

T-tests and z-tests are statistical methods employed in regression analysis to assess the significance of coefficients in a model. T-tests are the more common choice, suitable for small to moderately sized samples and situations where the population variance is unknown. They rely on the sample standard deviation to estimate the standard error and are used to evaluate the significance of coefficients by examining associated p-values. In contrast, z-tests are employed in rare cases with large samples and known population variances, making them less frequently used in regression analysis. The selection between these tests depends on sample size and data availability, but t-tests are typically more practical and widely applied in regression modeling.

***3. Normal Distribution and its types:***

**Normal distribution**, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

A normal distribution, also known as a Gaussian distribution or bell curve, is a probability distribution that is symmetric and bell-shaped. It is characterized by several key properties:

1. **Symmetry:** The normal distribution is symmetric, which means that the left and right sides of the distribution are mirror images of each other. The mean, median, and mode of a normal distribution are all located at the center of the distribution.
2. **Bell-Shaped:** The distribution has a distinctive bell-shaped curve, with the majority of data points clustered around the mean, and fewer data points in the tails.
3. **Mean and Variance:** The mean (average) and variance (a measure of spread) completely describe a normal distribution. The mean determines the center of the distribution, while the variance controls the width of the distribution. Larger variances result in wider distributions.
4. **Empirical Rule:** The empirical rule (also known as the 68-95-99.7 rule) states that in a normal distribution:

* Approximately 68% of the data falls within one standard deviation of the mean.
* Approximately 95% of the data falls within two standard deviations of the mean.
* Approximately 99.7% of the data falls within three standard deviations of the mean.

***Types of Normal Distributions:***

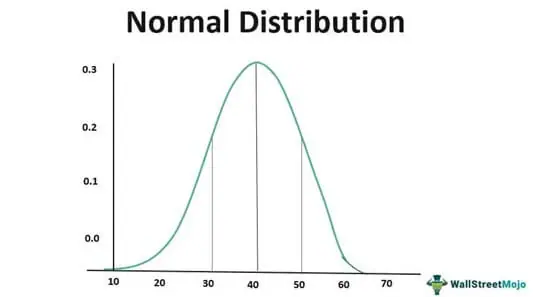
1. **Standard Normal Distribution:** This is a specific type of normal distribution where the mean (μ) is 0, and the standard deviation (σ) is 1. It is often denoted as N (0, 1). Standardization is often used to convert any normal distribution into the standard normal distribution using z-scores.
2. **Standardized Normal Distribution:** This is the same as the standard normal distribution and is used when you want to transform any normal distribution into one with a mean of 0 and a standard deviation of 1 using z-scores.
3. **Non-Standard Normal Distribution:** This refers to any normal distribution that does not have a mean of 0 and a standard deviation of 1. In practice, most normal distributions encountered in real-world data analysis are non-standard, and their properties are described by specific mean and standard deviation values.

Normal distributions play a fundamental role in statistics and data analysis because they describe the distribution of many naturally occurring phenomena, such as heights, test scores, and measurement errors. They are also essential in hypothesis testing and inferential statistics, as many statistical tests assume data is normally distributed or can be approximated as such.

***Normal Distribution Curve***

The curve takes the shape of a bell due to the symmetrical arrangement of the values that are concentrated towards the [central tendency](https://www.wallstreetmojo.com/central-tendency/). At the same time, the tail consists of an insignificant number of values.

Have a look at the curve below to understand its shape better:



### *Normal Distribution Formula*

The Probability Density Function (PDF) of a random variable (X) is given by:



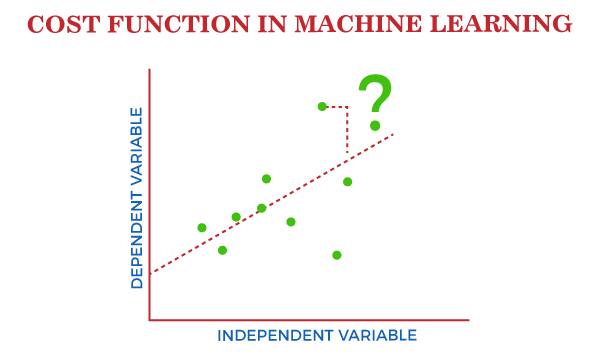
Where;

* -∞ < x < ∞; -∞ < µ < ∞; σ > 0
* F(x) = Normal probability Function
* x = Random variable
* µ = Mean of distribution
* σ = Standard deviation
* of the distribution
* π = 3.14159
* e = 2.71828

***4. Cost function in regression:***

Cost function measures the performance of a machine-learning model for a data set. Cost function quantifies the error between predicted and expected values and presents that error in the form of a single real number. Depending on the problem, cost function can be formed in many different ways. The purpose of cost function is to be either minimized or maximized. For algorithms relying on gradient descent to optimize model parameters, every function has to be differentiable.

A Machine Learning model should have a very high level of accuracy in order to perform well with real-world applications. However, how to calculate the accuracy of the model, i.e., how good or poor our model will perform in the real world? In such a case, the Cost function comes into existence. It is an important machine learning parameter to correctly estimate the model.



Cost function also plays a crucial role in understanding that how well your model estimates the relationship between the input and output parameters.In this topic, we will explain the cost function in Machine Learning, Gradient descent, and types of cost functions.

***Cost Function:***

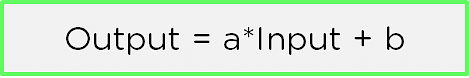
A cost function is an important parameter that determines how well a machine-learning model performs for a given dataset. It calculates the difference between the expected value and predicted value and represents it as a single real number.

**In machine learning**, once we train our model, then we want to see how well our model is performing. Although there are various accuracy functions that tell you how your model is performing, but will not give insights to improve them. Therefore, we need a function that can find when the model is most accurate by finding the spot between the undertrained and over-trained model.

**In simple**, "**Cost function is a measure of how wrong the model is in estimating the relationship between X (input) and Y (output) Parameter."** A cost function is sometimes also referred to as Loss function, and it can be estimated by iteratively running the model to compare estimated predictions against the known values of Y.

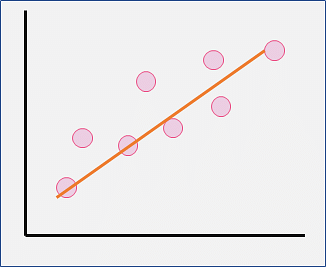
The main aim of each ML model is to determine parameters or weights that can minimize the cost function.

**A**[**Linear Regression**](https://www.simplilearn.com/tutorials/machine-learning-tutorial/linear-regression-in-python) model uses a straight line to fit the model. This is done using the equation for a straight line as shown:



In the equation, you can see that two entities can have changeable values (variable) a, which is the point at which the line intercepts the x-axis, and b, which is how steep the line will be, or slope.

At first, if the variables are not properly optimized, you get a line that might not properly fit the model. As you optimize the values of the model, for some variables, you will get the perfect fit. The perfect fit will be a straight line running through most of the data points while ignoring the noise and outliers. A properly fit Linear Regression model looks as shown below:



              For the Linear regression model, the cost function will be the minimum of the Root Mean Squared Error of the model, obtained by subtracting the predicted values from actual values. The cost function will be the minimum of these error values.

**Regression analysis**, the cost function (also known as the loss function or objective function) is a crucial component used to evaluate and optimize the performance of a regression model. The cost function quantifies the error or the disparity between the predicted values generated by the model and the actual observed values in the dataset. The goal is to minimize this cost function to obtain the best-fitting model. The specific form of the cost function depends on the type of regression being performed. Here are some common examples:

1. **Mean Squared Error (MSE):**

* For linear regression and many other regression techniques, the MSE is commonly used as the cost function.
* Formula: MSE=n1​∑i=1n​(yi​−y^​i​)2
* Here, yi represents the actual observed values, y^​i​represents the predicted values generated by the model, and n is the number of data points.
* The goal is to minimize the MSE, which is equivalent to minimizing the sum of squared errors.

1. **Mean Absolute Error (MAE):**

* Similar to MSE, but it measures the average absolute difference between actual and predicted values.
* Formula: MAE=n1​∑i=1n​∣yi​−y^​i​∣
* MAE is less sensitive to outliers compared to MSE.

1. **Huber Loss:**

* A hybrid of MSE and MAE that balances the robustness of MAE with the differentiability of MSE.
* It switches between MSE and MAE based on a user-defined parameter.

1. **Log-Loss (for Logistic Regression):**

* Used in logistic regression to measure the accuracy of binary classification.
* Formula: LogLoss=−n1​∑i=1n​(yi​log(y^​i​)+(1−yi​)log(1−y^​i​))
* Here, yi is the true binary label, and y^i\hat{y}\_iy^​i​ is the predicted probability of belonging to class 1.

1. **Poisson Deviance (for Poisson Regression):**

* Used in Poisson regression to measure the fit of the model to count data.
* It is based on the likelihood function of the Poisson distribution.

The choice of the cost function depends on the specific regression problem and the assumptions made about the distribution of the data. Different cost functions may emphasize different aspects of the model's performance, such as the treatment of outliers or the sensitivity to prediction errors. The regression model aims to find the parameter values that minimize the chosen cost function, often using optimization techniques like gradient descent or analytical solutions in the case of linear regression.

***5. Model Evaluation in Regression:***

The part in which we evaluate and test our model is where the loss functions come into play. Evaluation metric is an integral part of regression models. Loss functions take the model's predicted values and compare them against the actual values.

**Model evaluation in regression** is a critical step in assessing how well a regression model performs in predicting or explaining a dependent variable based on independent variables. Several metrics and techniques can be used to evaluate the performance of a regression model:

1. **Mean Absolute Error (MAE):**

* MAE measures the average absolute difference between the predicted values and the actual values.
* Formula: MAE=n1​∑i=1n​∣yi​−y^​i​∣
* Lower MAE indicates better model performance.

1. **Mean Squared Error (MSE):**

* MSE measures the average squared difference between the predicted values and the actual values.
* Formula: MSE=n1​∑i=1n​(yi​−y^​i​)2
* Lower MSE indicates better model performance, but it penalizes larger errors more than MAE.

1. **Root Mean Squared Error (RMSE):**

* RMSE is the square root of MSE and is often used because it's in the same units as the dependent variable.
* Formula: RMSE=MSE​
* Like MSE, lower RMSE indicates better performance.

1. **R-squared (R²) or Coefficient of Determination:**

* R² measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.
* Values range from 0 to 1, where 1 indicates a perfect fit.
* R² = 1 - (MSE of the model / MSE of a simple baseline model)
* Higher R² indicates a better fit.

1. **Adjusted R-squared (Adjusted R²):**

* Adjusted R² adjusts R² for the number of independent variables in the model, penalizing the inclusion of irrelevant variables.
* It provides a more realistic estimate of model fit in multiple regression.
* Higher adjusted R² indicates better model fit.

1. **Residual Analysis:**

* Plotting the residuals (the differences between actual and predicted values) can help identify patterns or outliers that the model might have missed.
* Residual plots, such as scatterplots of residuals against predicted values or independent variables, can be insightful.

1. **Hypothesis Tests:**

* Hypothesis tests can assess the significance of individual coefficients in the regression model. Common tests include t-tests or F-tests, depending on the type of regression.

1. **Cross-Validation:**

* Techniques like k-fold cross-validation can provide a more robust assessment of model performance by partitioning the data into multiple subsets for training and testing.

1. **Out-of-Sample Testing:**

* After training a model on one dataset, it's important to evaluate its performance on a separate, unseen dataset to assess how well it generalizes to new data.

1. **Domain-Specific Metrics:**

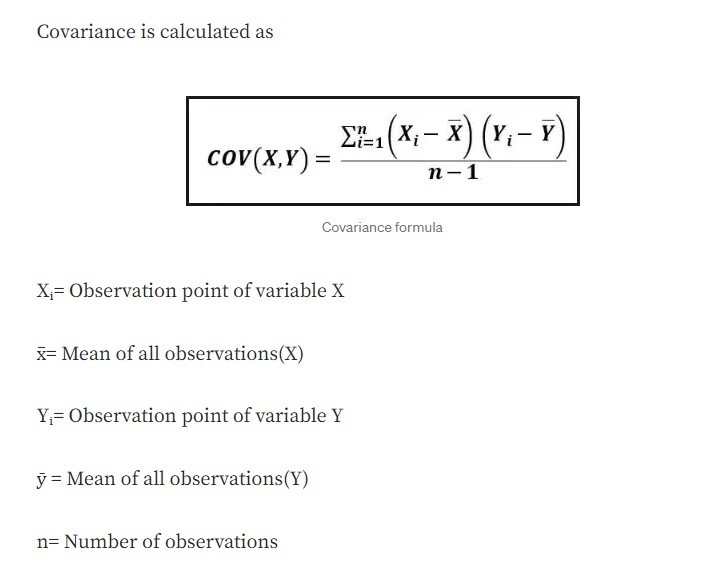
* Depending on the specific application, domain-specific metrics may be relevant. For example, in finance, metrics like Mean Absolute Percentage Error (MAPE) may be used.

The choice of evaluation metrics depends on the goals and characteristics of the regression problem. It is often a good practice to use a combination of metrics to get a comprehensive view of how well the model is performing. Additionally, visualizations and residual analysis can provide valuable insights into the model's strengths and weaknesses.

***6. Correlation | Causation | Co- variance:***

**Covariance** measures the direction of a relationship between two variables, while correlation measures the strength of that relationship. Both correlation and covariance are positive when the variables move in the same direction and negative when they move in opposite directions.

**Covariance** is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.



* The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.
* The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.
* A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables. Covariance is excellent for defining the type of relationship, but it's terrible for interpreting the magnitude.

**Correlation** is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

* Correlation is limited to values between the range -1 and +1.If two variables are closely correlated; we can then predict one variable from the other.
* One example of a common problem is that with small samples, correlations can be unreliable. The smaller the sample size, the more likely we are to observe a correlation that is further from 0, even if the true correlation (obtained if we had data for the entire population) was 0. The standard way of quantifying this is to use p-values. In academic research, a common rule of thumb is that when p is greater than 0.05, the correlation should not be trusted.
* Another problem with *correlation*is that it summarizes a linear relationship. If the true relationship is nonlinear, then this may be missed.

**Causation** indicates that one event results from the occurrence of the other event; i.e., there is a causal relationship between the two events. This is also referred to as cause and effect.

For Example: *After I exercise, I feel physically exhausted.*This is cause-and-effect because I’m purposefully pushing my body to physical exhaustion when doing exercise. The muscles I used to exercise are exhausted (effect) after I exercise (cause). This cause-and-effect IS confirmed.

**All causations are correlations, but not all correlations are causations.**

Examples of correlation, NOT causation: “On days where I go running, I notice more cars on the road. “ I, personally, am not CAUSING more cars to drive outside on the road when I go running. It is just that because I go running outside, I see more cars than when I stay at home. This relationship is not cause-and-effect because neither the cars nor I are impacting each other.

In Conclusions, while covariance identifies how two variables vary simultaneously, correlation determines how change in one variable affects the change in another variable.

The two variables are correlated with each other and there is a causal link between them. A correlation does not imply causation, but causation always implies correlation.

***Difference between Covariance and Correlation:***

Covariance measures the direction of a relationship between two variables, while correlation measures the strength of that relationship. Both correlation and covariance are positive when the variables move in the same direction and negative when they move in opposite directions. However, a correlation coefficient must always be between -1 and +1, with extreme values indicating a strong relationship.

**Correlation, causation, and covariance** are fundamental concepts in statistics and data analysis, but they represent different aspects of the relationships between variables:

1. **Correlation:**

* **Definition**: Correlation measures the statistical association or relationship between two variables. It quantifies how one variable's changes relate to changes in another variable.
* **Strength**: The strength of correlation is typically measured by a correlation coefficient, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value of 0 indicates no linear correlation.
* **Example**: If you find a high positive correlation between the number of hours spent studying and exam scores, it suggests that as study time increases, exam scores tend to increase as well. However, correlation does not imply causation.

1. **Causation:**

* **Definition**: Causation indicates a cause-and-effect relationship between two variables. It means that changes in one variable directly lead to changes in another variable.
* **Establishing Causation**: Establishing causation is more complex than finding correlation. It often requires controlled experiments, such as randomized controlled trials, to demonstrate that changes in one variable (the cause) lead to changes in another (the effect).
* **Example**: If a controlled study shows that administering a certain drug leads to a decrease in patient symptoms, you can conclude that the drug causes the symptom reduction. However, observing correlation alone does not prove causation because other factors may be at play.

1. **Covariance:**

* **Definition**: Covariance measures the degree to which two variables change together. It assesses the joint variability of two variables.
* **Strength**: Like correlation, covariance can be positive (variables tend to increase or decrease together) or negative (one increases as the other decreases). However, it doesn't provide a standardized measure like the correlation coefficient, so its magnitude is difficult to interpret directly.
* **Formula**: The covariance between two variables X and Y is given by: Cov(X,Y)=n∑(Xi​−Xˉ)(Yi​−Yˉ)​, where Xi and Yi are data points, Xˉ and Yˉ are the means of X and Y, and n is the number of data points.
* **Units**: The units of covariance are the product of the units of the two variables.

***Conclusion:***

Correlation measures the statistical association between two variables, causation implies a cause-and-effect relationship, and covariance measures the joint variability of two variables. While correlation is a useful tool for exploring relationships, it does not prove causation, which requires methods that are more rigorous. Covariance provides a measure of joint variability but lacks the standardized interpretation of correlation.

***7. Importance of P-value in Regression:***

The p values in regression **help determine whether the relationships that you observe in your sample also exist in the larger population.** The linear regression p value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable.

The P-value is a statistical number to conclude if there is a relationship between Average Pulse and Calorie\_Burnage.

We test if the true value of the coefficient is equal to zero (no relationship). The statistical test for this is called Hypothesis testing.

* A low P-value (< 0.05) means that the coefficient is likely not to equal zero.
* A high P-value (> 0.05) means that we cannot conclude that the explanatory variable affects the dependent variable (here: if Average Pulse affects Calorie\_Burnage).
* A high P-value is also called an insignificant P-value.

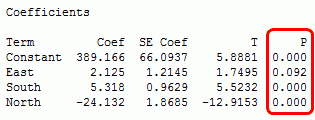
## *Interpreting P Values in Regression for Variables*

Regression analysis is a form of [inferential statistics](https://statisticsbyjim.com/basics/descriptive-inferential-statistics/). The p values in regression help determine whether the relationships that you observe in your [sample](https://statisticsbyjim.com/glossary/sample/) also exist in the larger [population](https://statisticsbyjim.com/glossary/population/). The linear regression p value for each independent variable tests the null hypothesis that the variable has no [correlation](https://statisticsbyjim.com/glossary/correlation/) with the dependent variable. If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is insufficient evidence to conclude that there is an [effect](https://statisticsbyjim.com/glossary/effect/) at the population level.

If the [p-value](https://statisticsbyjim.com/glossary/p-value/) for a variable is less than your [significance level](https://statisticsbyjim.com/glossary/significance-level/), your sample data provide enough evidence to reject the null hypothesis for the entire population. Your data favor the hypothesis that there *is* a non-zero correlation. Changes in the independent variable *are* associated with changes in the dependent variable at the population level. This variable is statistically significant and probably a worthwhile addition to your regression model.

On the other hand, when a p value in regression is greater than the significance level, it indicates there is insufficient evidence in your sample to conclude that a non-zero correlation exists.

The regression output example below shows that the South and North predictor variables are statistically significant because their p-values equal 0.000. On the other hand, East is not statistically significant because its p-value (0.092) is greater than the usual significance level of 0.05.



It is standard practice to use the coefficient p-values to decide whether to include variables in the final model. For the results above, we would consider removing East. Keeping variables that are not statistically significant can reduce the model’s precision.

The p-value is a crucial statistical measure in regression analysis, and it serves several important purposes:

1. **Hypothesis Testing:** The primary role of the p-value in regression is to help with hypothesis testing. In the context of regression, the p-value assesses the statistical significance of individual coefficients (predictor variables) in the model. Specifically, it answers the question: "Is there a statistically significant relationship between a particular predictor variable and the response variable?"
2. **Coefficient Significance:** Each coefficient in a regression model represents the change in the response variable associated with a one-unit change in the corresponding predictor variable, while holding all other variables constant. The p-value associated with a coefficient helps you determine whether that variable has a statistically significant impact on the response variable. A small p-value (typically below a chosen significance level, like 0.05) suggests that the variable is likely to have a significant effect.
3. **Model Selection:** The p-value can aid in model selection by helping you decide which predictor variables to include in your model. Variables with high p-values (indicating low significance) may be candidates for removal from the model, as they may not contribute significantly to explaining the variation in the response variable.
4. **Interpretation:** When presenting regression results, the p-value provides important information about the reliability of the coefficient estimates. A low p-value indicates that you can have more confidence in the significance of the relationship between a predictor and the response variable.
5. **Control of Type I Error:** The p-value is closely related to Type I error, which is the probability of incorrectly concluding that there is a significant effect when there is none (i.e., a false positive). By setting a significance level (e.g., 0.05), you control the risk of making Type I errors. If the p-value is below this threshold, you reject the null hypothesis and conclude that the predictor variable is statistically significant.
6. **Scientific and Practical Significance:** While a predictor variable may have a statistically significant p-value, it is also important to consider whether the observed effect is practically or scientifically significant. Even if a relationship is statistically significant, it may not be meaningful in real-world terms. The p-value helps identify statistical significance, but it is not a measure of the size of the effect.

It is important to note that while the p-value is a valuable tool in regression analysis, it should not be used in isolation. It should be considered alongside other factors, such as the magnitude and direction of coefficients, domain knowledge, and practical implications. Additionally, the p-value has limitations and should not be used as the sole determinant of the importance or validity of a relationship in regression.

***8. Sampling and Data Sampling in Probability and Regression (Sampling in Data Science):***

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

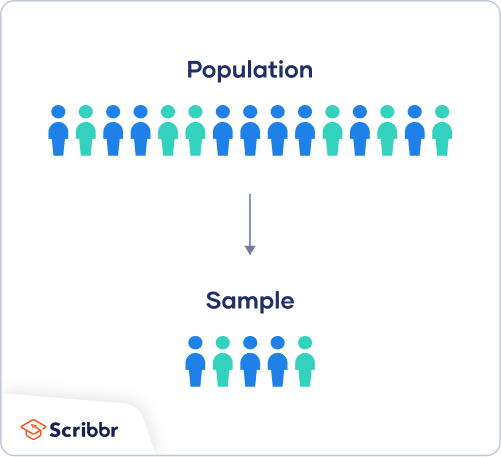
**Data sampling** is a statistical analysis technique used to select, process, and analyze a representative subset of a population. It is also used to identify patterns and extrapolate trends in an overall population.

## *Population vs. sample*

First, you need to understand the difference between [a population and a sample](https://www.scribbr.com/methodology/population-vs-sample/), and identify the target population of your research.

* The **population** is the entire group that you want to draw conclusions about.
* The **sample** is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, or many other characteristics.



It can be very broad or quite narrow: maybe you want to make inferences about the completely adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project. If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample. A lack of a representative sample affects the [validity](https://www.scribbr.com/methodology/types-of-validity/) of your results, and can lead to several [research biases](https://www.scribbr.com/faq-category/research-bias/), particularly [sampling bias](https://www.scribbr.com/research-bias/sampling-bias/).

Sampling is a fundamental concept in both probability theory and data science, and it plays a significant role in various aspects, including probability theory, statistical inference, and regression analysis. Here is how sampling is relevant in these domains:

**Sampling in Probability Theory:**

1. **Simple Random Sampling:** In probability theory, sampling often refers to selecting a random subset (sample) of elements from a larger population. Simple random sampling is a common method where each element in the population has an equal chance of being included in the sample. Probability distributions, such as the binomial distribution and the hypergeometric distribution, are used to model and analyze random sampling processes.
2. **Sampling Distribution:** Probability theory explores the concept of a sampling distribution, which describes the distribution of a statistic (e.g., sample mean or sample variance) calculated from different random samples drawn from the same population. The Central Limit Theorem is a fundamental concept in this context, stating that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population distribution.

**Sampling in Data Science:**

1. **Data Collection:** In data science, obtaining a sample from a larger dataset is often the initial step. Collecting data from an entire population can be time-consuming and costly, so a well-designed sample can represent the population accurately while reducing data collection efforts.
2. **Statistical Inference:** In statistical analysis and regression modeling, samples are used to make inferences about the entire population. The properties of the sample, such as sample means and variances, are used to estimate population parameters and test hypotheses about the population.
3. **Model Validation and Testing:** When developing regression models, it's common to split the available data into a training set and a testing (or validation) set. The training set is used to build the model, while the testing set is used to assess the model's performance and generalization to unseen data.
4. **Bootstrapping:** Bootstrapping is a resampling technique used in data science to estimate the sampling distribution of a statistic by repeatedly sampling with replacement from the observed data. It's particularly useful for constructing confidence intervals and assessing model stability.
5. **Stratified Sampling:** In data science, stratified sampling is often used to ensure that the sample represents subgroups (strata) within the population. For example, in regression analysis, you may want to ensure that your sample includes a proportional representation of different demographic groups.
6. **Imbalanced Data:** In classification and regression tasks, imbalanced datasets can be problematic. Sampling techniques such as oversampling the minority class or undersampling the majority class can be used to address class imbalance issues.

**Sampling in Regression:**

In regression analysis, sampling is crucial for several reasons:

1. **Sample Selection:** You need to carefully choose a sample that represents the population of interest. Biased or unrepresentative samples can lead to inaccurate regression models and misleading conclusions.
2. **Training and Testing:** As mentioned earlier, splitting the data into training and testing sets is essential for model evaluation. Proper sampling ensures that the model's performance is assessed on unseen data.
3. **Residual Analysis:** Residuals (the differences between observed and predicted values) are crucial in regression. A well-chosen sample allows for meaningful analysis of residuals to check for model assumptions, such as homoscedasticity and normality.
4. **Cross-Validation:** Techniques like k-fold cross-validation involve splitting the data into multiple subsets (folds) to assess a model's performance. Proper sampling of the data ensures the reliability of cross-validation results.

**In summary**, sampling is a fundamental concept in probability theory and data science, with implications for statistical inference, model development, and evaluation, including regression analysis. Careful consideration of sampling methods and techniques is essential to ensure the validity and reliability of statistical and data science analyses.

Data sampling is a fundamental concept in both probability theory and regression analysis, and it plays a crucial role in various aspects of these fields. Here's how data sampling is relevant in probability and regression:

**Data Sampling in Probability:**

1. **Probability Distributions:** Sampling is closely linked to probability distributions. In probability theory, various distributions (e.g., normal, binomial, Poisson) describe the likelihood of observing different values of a random variable. Sampling is used to estimate parameters of these distributions or to generate random samples from them.
2. **Law of Large Numbers:** The Law of Large Numbers in probability theory states that as the sample size increases, the sample mean (average) converges to the population mean. Sampling plays a critical role in illustrating this concept, showing how the behavior of sample statistics approaches population parameters with increasing sample size.
3. **Sampling from a Probability Distribution:** In probability, you might be interested in drawing random samples from a specific probability distribution to simulate outcomes or conduct Monte Carlo simulations. Sampling methods, such as random number generators, are used to achieve this.

**Data Sampling in Regression:**

1. **Sample Selection:** In regression analysis, you typically start by collecting a sample of data from a larger population. The sample should be carefully selected to be representative of the population of interest. Proper sampling is crucial for the validity of regression results.
2. **Model Estimation:** Once you have a sample, you use regression techniques to estimate the relationships between variables in your model. The estimated coefficients represent the relationships based on the sample data.
3. **Hypothesis Testing:** In regression, you often use hypothesis tests to assess the statistical significance of coefficients. These tests rely on sampling distributions to evaluate whether the coefficients are significantly different from zero.
4. **Residual Analysis:** Residuals, which are the differences between observed and predicted values, play a central role in regression analysis. The distribution of residuals should approximate normality for the validity of regression assumptions, and sampling is involved in residual analysis.
5. **Cross-Validation:** When developing regression models, you may use cross-validation techniques, such as k-fold cross-validation, to assess how well the model generalizes to unseen data. These techniques involve partitioning the data into multiple subsets for training and testing.
6. **Bootstrap Resampling:** Bootstrap resampling is a technique used to estimate the sampling distribution of a statistic by repeatedly sampling with replacement from the observed data. It's particularly useful for constructing confidence intervals and assessing model stability in regression.
7. **Out-of-Sample Testing:** It's common practice to evaluate the performance of a regression model on a separate testing dataset, which is sampled from the same population but not used in model estimation. This helps assess how well the model predicts new, unseen data.

**Summary**:

Data sampling is integral to both probability theory and regression analysis. In probability, it's used to understand and simulate random processes and distributions. In regression, it's the foundation for model development, hypothesis testing, and model validation, ensuring that regression models provide meaningful insights into real-world data. Proper sample selection and analysis are essential for robust and reliable regression results.