STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Answers

1. (a) True
2. (a)
3. (b)
4. (d)
5. (c)
6. (b) False
7. (b)
8. (a)
9. (c)

Subjective Type

10. **Normal Distribution:** 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".

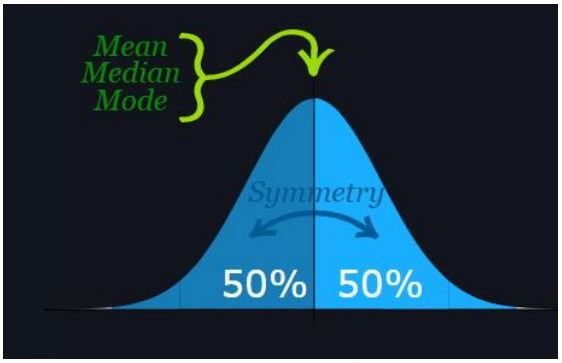

Figure 1Graphical representation of Normal distribution

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

Parameters of the Normal Distribution: -

As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two parameters, the mean and standard deviation. The Gaussian distribution does not have just one form. Instead, the shape changes based on the parameter values, as shown in the graphs below.

- Mean

The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. On a graph, changing the mean shifts the entire curve left or right on the X-axis.

- Standard deviation

The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average. On a graph, changing the standard deviation either tightens or spreads out the width of the distribution along the X-axis. Larger standard deviations produce wider distributions.

The Formula for the Normal Distribution: -

The normal distribution is produced by the normal density function,

$$P(x) = [e^{-\{(x-\mu)^2\}/(2\sigma^2)}]/ \sigma\sqrt{(2\pi)}$$, where,

- x = value of the variable or data being examined and P(x) the probability function
- μ = the mean
- σ = the standard deviation

Properties of the Normal Distribution: -

The normal distribution has several key features and properties that define it.

i. First, its mean (average), median (midpoint), and mode (most frequent observation) are all equal to one another. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

ii. The Empirical Rule
For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations. This fact is sometimes referred to as the "empirical rule," a heuristic that describes where most of the data in a normal distribution will appear. This means that data falling outside of three standard deviations ("3-sigma") would signify rare occurrences.

iii. Skewness
Skewness measures the degree of symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set instead has a skewness less than zero, or negative skewness (left-skewness), then the left tail of the distribution is longer than the right tail; positive skewness (right-skewness) implies that the right tail of the distribution is longer than the left.

iv. Kurtosis
Kurtosis measures the thickness of the tail ends of a distribution in relation to the tails of a distribution. The normal distribution has a kurtosis equal to 3.0. Distributions with larger kurtosis greater than 3.0 exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). This excess kurtosis is known in statistics as leptokurtic, but is more colloquially known as "fat tails." The occurrence of fat tails in financial markets describes what is known as tail risk. Distributions with low kurtosis less than 3.0 (platykurtic) exhibit tails that are generally less extreme ("skinnier") than the tails of the normal distribution.

Standard Normal Distribution and Standard Scores: -

The normal distribution has many different shapes depending on the parameter values. However, the standard normal distribution is a special case of the normal distribution where the mean is zero and the standard deviation is 1. This distribution is also known as the Z-distribution.

A value on the standard normal distribution is known as a standard score or a Z-score. A standard score represents the number of standard deviations above or below the mean that a specific observation falls. For example, a standard score of 1.5 indicates that the observation is 1.5 standard deviations above the mean. On the other hand, a negative score represents a value below the average. The mean has a Z-score of 0.

Standard scores are a great way to understand where a specific observation falls relative to the entire normal distribution. They also allow us to take observations drawn from normally distributed populations that have different means and standard deviations and place them on a standard scale. This standard scale enables us to compare observations that would otherwise be difficult.

This process is called standardization, and it allows us to compare observations and calculate probabilities across different populations. To standardize our data, we need to convert the raw measurements into Z-scores.

To calculate the standard score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is the following:

$Z=(X-\mu) /\sigma$, X represents the raw value of the measurement of interest. μ and σ represent the parameters for the population from which the observation was drawn.

Importance of normal distribution: -

1) It has one of the important properties called central theorem. Central theorem means relationship between shape of population distribution and shape of sampling distribution of mean. This means that the sampling distribution of the mean approaches normal as the sample size increase.
2) In case the sample size is large the normal distribution serves as a good approximation.
3) Due to its mathematical properties, it is more popular and easier to calculate.
4) It is used in statistical quality control in setting up of control limits.
5) The whole theory of sample tests t, f, and chi-square test is based on the normal distribution.

These are the importance or uses or benefits of normal distribution.

11. **Handle missing data & imputation techniques:** Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such for research purposes. However, until recently, most researchers have drawn conclusions based on the assumption of a complete data set. Missing data represents various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions.

    Techniques for Handling the Missing Data: -

    Listwise or case deletion

    By far the most common approach to the missing data is to simply omit those cases with the missing data and analyze the remaining data. This approach is known as the complete case (or available case) analysis or listwise deletion. Listwise deletion is the most frequently used method in handling missing data and thus has become the default option for analysis in most statistical software packages. Some researchers insist that it may introduce bias in the estimation of the parameters. However, if the assumption of MCAR is satisfied, listwise deletion is known to produce unbiased estimates and conservative results. When the data do not fulfill the assumption of MCAR, listwise deletion may cause bias in the estimates of the parameters.
    If there is a large enough sample, where power is not an issue, and the assumption of MCAR is satisfied, the listwise deletion may be a reasonable strategy. However, when there is not a large sample or the assumption of MCAR is not satisfied, the listwise deletion is not the optimal strategy.

    Pairwise deletion

    Pairwise deletion eliminates information only when the particular data point needed to test a particular assumption is missing. If there is missing data elsewhere in the data set, the existing values are used in the statistical testing. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion, which may delete the case with any missing data. This approach presents the following problems: 1) the parameters of the model will stand on different sets of data with different statistics, such as the sample size and standard errors; and 2) it can produce an intercorrelation matrix that is not positive definite, which is likely to prevent further analysis. Pairwise deletion is known to be less biased for the MCAR or MAR data, and the appropriate mechanisms are included as covariates. However, if there are many missing observations, the analysis will be deficient.

    Mean substitution

    In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This allows the researchers to utilize the collected data in an incomplete dataset. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. However, with missing values that are not strictly random, especially in the presence of great inequality in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. Furthermore, this approach adds no new information but only increases the sample size and leads to an underestimate of the errors. Thus, mean substitution is not generally accepted.

Regression imputation

Imputation is the process of replacing the missing data with estimated values. Instead of deleting any case that has any missing value, this approach preserves all cases by replacing the missing data with a probable value estimated by other available information. After all missing values have been replaced by this approach, the data set is analyzed using the standard techniques for complete data.

In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as an actually obtained value. This approach has a number of advantages because the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.

Maximum likelihood

There are a number of strategies using the maximum likelihood method to handle the missing data. In these, the assumption that the observed data are a sample drawn from a multivariate normal distribution is relatively easy to understand. After the parameters are estimated using the available data, the missing data are estimated based on the parameters which have just been estimated.

When there are missing but relatively complete data, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That is, the missing data may be estimated by using the conditional distribution of the other variables.

Expectation-Maximization

Expectation-Maximization (EM) is a type of the maximum likelihood method that can be used to create a new data set, in which all missing values are imputed with values estimated by the maximum likelihood methods. This approach begins with the expectation step, during which the parameters (e.g., variances, covariances, and means) are estimated, perhaps using the listwise deletion. Those estimates are then used to create a regression equation to predict the missing data. The maximization step uses those equations to fill in the missing data. The expectation step is then repeated with the new parameters, where the new regression equations are determined to "fill in" the missing data. The expectation and maximization steps are repeated until the system stabilizes when the covariance matrix for the subsequent iteration is virtually the same as that for the preceding iteration.

An important characteristic of the expectation-maximization imputation is that when the new data set with no missing values is generated, a random disturbance term for each imputed value is incorporated in order to reflect the uncertainty associated with the imputation. However, the expectation-maximization imputation has some disadvantages. This approach can take a long time to converge, especially when there is a large fraction of missing data, and it is too complex to be acceptable by some exceptional statisticians. This approach can lead to biased parameter estimates and can underestimate the standard error.

For the expectation-maximization imputation method, a predicted value based on the variables that are available for each case is substituted for the missing data. Because a single imputation omits the possible differences among the multiple imputations, a single imputation will tend to underestimate the standard errors and thus overestimate the level of precision. Thus, a single imputation gives the researcher more apparent power than the data in reality.

Multiple imputations

Multiple imputations are another useful strategy for handling the missing data. In multiple imputations, instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.

This approach begins with a prediction of the missing data using the existing data from other variables. The missing values are then replaced with the predicted values, and a full data set called the imputed data set is created. This process iterates the repeatability and makes multiple imputed data sets (hence the term "multiple imputations"). Each multiple imputed data set produced is then analyzed using the standard statistical analysis procedures for complete data, and gives multiple analysis results. Subsequently, by combining these analysis results, a single overall analysis result is produced.

The benefit of the multiple imputations is that in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in a valid statistical inference. Restoring the natural variability of the missing data can be achieved by replacing the missing data with the imputed values which are predicted using the variables correlated with the missing data. Incorporating uncertainty is made by producing different versions of the missing data and observing the variability between the imputed data sets.

Multiple imputations have been shown to produce valid statistical inference that reflects the uncertainty associated with the estimation of the missing data. Furthermore, multiple imputations turn out to be robust to the violation of the normality assumptions and produce appropriate results even in the presence of a small sample size or a high number of missing data.

With the development of novel statistical software, although the statistical principles of multiple imputation may be difficult to understand, the approach may be utilized easily.

Sensitivity analysis
Sensitivity analysis is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs.
When analyzing the missing data, additional assumptions on the reasons for the missing data are made, and these assumptions are often applicable to the primary analysis. However, the assumptions cannot be definitively validated for correctness.

12. **A/B testing:** A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Statistically, it is a Two-sample hypothesis testing methodology for making decisions that estimate population parameters based on sample statistics. Two-sample hypothesis testing is a method of determining whether the differences between the two samples are statistically significant or not.

    Before conducting A/B testing, we have to state our null hypothesis and alternative hypothesis:
    The null hypothesis is one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant groups.
    The alternative hypothesis is one that states that sample observations are influenced by some non-random cause. From an A/B test perspective, the alternative hypothesis states that there is a difference between the control and variant groups.
    When developing our null and alternative hypotheses, it's recommended that we should follow a PICOT format.
    Picot stands for:
    - ✓ Population: the group of people that participate in the experiment
    - ✓ Intervention: refers to the new variant in the study
    - ✓ Comparison: refers to what we plan on using as a reference group to compare against our intervention
    - ✓ Outcome: represents what result we plan on measuring
    - ✓ Time: refers to the duration of the experience (when and how long the data is collected)

13. **Whether mean imputation of missing data is acceptable practice:** The process of replacing null values in a data collection with the data mean is known as mean imputation. It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. But that doesn't make it a good solution, and it may not help us find relationships with strong parameter estimates. Even if they exist in the population. The following problems explain the many reasons not to use mean imputation: -

    - ✓ Problem 1: Mean imputation does not preserve the relationships among variables. If all we are doing is estimating means, and if the data are missing completely at random, mean imputation will not bias our parameter estimate. It will still bias our standard error. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

    - ✓ Problem 2: Mean Imputation Leads to An Underestimate of Standard Errors. A second reason applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

In other words, yes, we get the same mean from mean-imputed data that we would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. But our statistical software doesn't know that. It treats it as real data. Ultimately, because our standard errors are too low, so are our p-values. Now we're making Type I errors without realizing it.

In simple words, mean imputation is typically considered terrible practice since it ignores feature correlation. Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. **Linear Regression in Statistics:** Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable. The variable we are using to predict the other variable's value is called the independent variable.
This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. We then estimate the value of X (dependent variable) from Y (independent variable).

Importance of linear regression: -
Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.
It can be found that linear regression is used in everything from biological, behavioural, environmental, and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear regression models are well understood and can be trained very quickly.

Reliable model to predict the future: -
Business and organizational leaders can make better decisions by using linear regression techniques. Organizations collect masses of data, and linear regression helps them use that data to better manage reality, instead of relying on experience and intuition. We can take large amounts of raw data and transform it into actionable information.
Linear regression can be used to provide better insights by uncovering patterns and relationships that the data have previously seen and thought they already understood. For example, performing an analysis of sales and purchase data can help us uncover specific purchasing patterns on particular days or at certain times. Insights gathered from regression analysis can help business leaders anticipate times when their company's products will be in high demand.

Key assumptions of effective linear regression: -
Assumptions to be considered for success with linear-regression analysis:

- For each variable: Consider the number of valid cases, mean and standard deviation.
- 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 observations.
- Normality: The data follows a normal distribution.
- Plots: Consider scatterplots, partial plots, histograms, and normal probability plots.
- Data: Dependent and independent variables should be quantitative. Categorical variables need to be recoded to binary (dummy) variables or other types of contrast variables.
- Other assumptions: For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear and all observations should be independent.

Linear Models: -
Despite the name, linear regression can model curved relationships. In this context, the term "linear" describes the form of the regression equation. A regression equation is linear when all its terms are one of the following:
i. Constant.
ii. Parameter multiplying an independent variable.

Additionally, a linear regression equation can only add terms together, producing one general form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Or,

$$y = \beta_0 + \beta_1 X + \epsilon$$

Where,

i. β_0 is the intercept, the predicted value of y when the x is 0.
ii. β_i's are the regression coefficient – how much we expect y to change as x increases.
iii. x is the independent variable (the variable we expect is influencing y).
iv. ε is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

Advantages: -

- ✓ Linear Regression is simple to implement and easier to interpret the output coefficients.
- ✓ When it is known that the relationship between the independent and dependent variable has a linear relationship, this algorithm is the best to use because of its less complexity to compared to other algorithms.
- ✓ Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques, and cross-validation.

Disadvantages: -

- ✓ Linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
- ✓ Linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
- ✓ Linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

15. **Various branches of Statistics: -** The root of statistics is driven by variables. A variable is a data set that can be counted that marks a characteristic or attribute of an item. Statistics is the branch of mathematics that deals with data. Data (technically a plural word; the singular is 'datum') is a collection of values. A collection of data is often referred to as a data set or set of data¸ but other words such as a list or simply collection are also often used.

    There are three real branches of statistics: data collection, descriptive statistics, and inferential statistics.

i. *DATA COLLECTION:* - Data collection is all about how the actual data is collected. There are issues in the collection of the data; It should be made sure that the data has been collected fairly before going on to deal with it, and try to present it and make conclusions. The population is the entire set of data, and a sample is a representative or subset of the population – so just some of the data values. Why would we need a sample? Because, it's usually unrealistic to get data from the entire population, incredibly expensive and time-consuming, and also the population may be changing as the data is collected, so often we need to simply take a sample. We can consider the following techniques in this scenario:

Statistics Sampling Techniques: - To gather statistical information, it would often not be possible to gather data from every data point within a population. Instead, statistics relies on different sampling techniques to create a representative subset of the population that is easier to analyze. In statistics, there are several primary types of sampling.

- Simple random sampling calls for every member within the population to have an equal chance of being selected for analysis. The entire population is used as the basis for sampling, and any random generator based on chance can select the sample items. For example, 100 individuals are lined up and 10 are chosen at random.
- Systematic sampling calls for a random sample as well. However, its technique is slightly modified to make it easier to conduct. A single random number is generated, and individuals are then selected at a specified regular interval until the sample size is complete. For example, 100 individuals are lined up and numbered. The 7th individual is selected for the sample followed by every subsequent 9th individual until 10 sample items have been selected.
- Stratified sampling calls for more control over your sample. The population is divided into subgroups based on similar characteristics. Then, you calculate how many people from each subgroup would represent the entire population. For example, 100 individuals are grouped by gender and race. Then, a sample from each subgroup will be taken in the proportion of how representative that subgroup is of the population.
- Cluster sampling calls for subgroups as well. However, each subgroup should be representative of the population. Instead of randomly selecting individuals within a subgroup, the entire subgroup is randomly selected.

ii. *DESCRIPTIVE STATISTICS:* - Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median, and mode. Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation.

The distribution refers to the overall "shape" of the data, which can be depicted on a chart such as a histogram or a dot plot, and includes properties such as the probability distribution function, skewness, and kurtosis. Descriptive statistics can also describe differences between observed characteristics of the elements of a data set. Descriptive statistics help us understand the collective properties of the elements of a data sample and form the basis for testing hypotheses and making predictions using inferential statistics.

iii. INFERENTIAL STATISTICS: - Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.

Regression analysis is a widely used technique of statistical inference used to determine the strength and nature of the relationship (i.e., the correlation) between a dependent variable and one or more explanatory (independent) variables. The output of a regression model is often analyzed for statistical significance, which refers to the claim that a result from findings generated by testing or experimentation is not likely to have occurred randomly or by chance but is likely to be attributable to a specific cause elucidated by the data.

Having statistical significance is important for academic disciplines or practitioners that rely heavily on analyzing data and research.