Glossary: Statistics Fundamentals using Excel

Welcome! This alphabetized glossary contains many terms used in this course. Understanding these terms is essential when working in the industry, participating in user groups, and participating in other certificate programs.

Estimated reading time: 7 minutes

Term	Definition
Alpha value	An average of significance.
Average	The sum of all values divided by the number of values.
Bar chart	A pictorial representation of data in the form of vertical or horizontal rectangular bars. In bar charts, the length of bars is proportional to the measure of data.
Big data	A vast and diverse collection of data points that increases in volume, velocity, and variety.
Binary variable	Variable which has only two variables, that is, one dependent and one independent variable.
Boxplot	A graphical depiction of numerical data that helps us to comprehend and analyze its distribution readily.
Bubble chart	A graph that uses circles to represent the data.
Business analyst	Data professionals who bridge the gap between business objectives and technical solutions. They identify and document business requirements, conduct feasibility studies, and propose improvements to business processes.
Categorical variable	A variable that assigns each unit of observation to a group based on a qualitative property.
Counterplot	Refers to the number of times each observation appears in a categorical variable. It provides a visual representation of the distribution of observations in the variable.
Cross-sectional data	A data observation from multiple subjects, such as individuals, firms, countries, or regions, at a single point in time.
Cumulative distribution function	The probability that X will take a value less than or equal to x of a real-valued random variable X. This function helps to describe the probability distribution of random variables in a table.
Data science	A field to study data to extract meaningful insights for businesses.
Delphi method	A structured communication technique developed as an interactive forecasting method that relies on a panel of experts.
Dependent variable	A variable that changes as a result of the manipulation of the independent variable.
Descriptive statistics	A statistic that summarizes the characteristics of a sample through quantitative or visual means.
Dichotomous variable	Variables that only take on two possible values.
Dispersion	A means of describing the extent of distribution of data around a central value or point.
Endogeneity	The correlation between the independent variable and unexplained variation (or "error") in the dependent variable.
F-statistics	A value is presented for the F distribution. Various statistical tests generate an F value. It determines whether the test is statistically significant.
Histogram	A graph displays a frequency distribution with continuous classes grouped.
Homoscedacity	A condition in which the variance of the error term in a regression model is not constant.
Hypothesis	A formal statement that explains the relationship between two or more variables of the specified population.
Hypothesis test	A statistical inference that concludes a population parameter or probability distribution using data from a sample.
Independent variable	A variable that is not affected by the other variables that need to be measured.
Linear regression	A data analysis method to predict the value of unknown data by using another related and known data value.
Linearity	The relationship between two variables when the change in one variable changes the other variable proportionally.
Logistic regression	A statistical method that estimates the probability of an event, such as voted or didn't vote, based on a given data set of independent variables.
Mean absolute error (MAE)	An average absolute difference between the fitted values within the model (one-step ahead in-sample forecast) and the observed historical data.
Median	A value of separating the higher value of a data sample, a population, or a probability distribution from the lower value. It is the middle value of a data set.
Mode	A value that appears most frequently in the data set.
Multicollinearity	A regression analysis where several independent variables in a model are highly correlated leads to inaccurate and unreliable results when analyzing the relationship between the dependent and independent variables.
Multiple linear regression	A statistical technique that uses multiple independent variables to predict a dependent variable.
Multivariant data set	Data sets with two or more variables.
Nominal variable	A type of categorical variable that can have two or more categories.
Normal distribution	A symmetric probability distribution about the mean indicates that data near the mean occurs more frequently than data far from the mean.
Normal distribution curve	A symmetric frequency distribution around the mean indicates a higher occurrence of data near the mean than far from it.
Null hypothesis	A type of statistical hypothesis proposing that there is no significant difference between the observed data and the expected results.
Overfitting	A concept in data science, which occurs when a statistical model fits exactly against its training data. Unfortunately, when this happens, the algorithm cannot perform accurately against unseen data, defeating its purpose.
Pie chart	A circular statistical presentation of data.
Polynomial regression	The relationship between the dependent and independent variables Y and X modeled as the n^{th} degree of the polynomial.
Probability	A number reflecting the chance or likelihood for a particular event to occur. It measures values between zero and one for the possibility that something or some event might happen.
P-value	The probability of observing results equally or more extreme than the observed results, assuming that the null hypothesis is true.
Qualitative forecasting	The prediction of finances using expert opinion.
Quantitative forecasting	The prediction of future demands based on historical data.
Quantitative relationship	Relationship between variables that are expressed in numerical values.
Regression analysis	The prediction of a variable's value based on another variable's value.
Regression model	A model that determines the relationship between the dependent and independent variables.
R-squared	A statistical measure used in regression models to determine the amount of variation in the dependent variable that the independent variable can explain.
Scatter plot	Graphs are used to represent the relationship between two variables in the data sets.
Simple linear regression	A statistical technique that estimates the relationship between one independent variable and one dependent variable, and its graph is a straight line.
Skewness	A measure of the asymmetry of a statistical data distribution from a normal distribution.

Term	Definition
Standard deviation	A measure of the amount of variation or dispersion of a set of values.
Standard error	The amount of the value of a test statistic varies from sample to sample. It is a measure of the uncertainty of the test statistic.
Statistical analysis	A process of collecting and analyzing large amounts of data to identify trends and patterns.
Statistical parameter	A summary statistic is a parameter that describes an aspect of a statistical population, such as its mean or standard deviation.
Statistical tools	Tools by which statistical methods are applied.
Statistics	A branch of study that helps to analyze, interpret, and present data.
Symmetrical distribution	The data set has the same mean, median, and mode.
T distribution	The estimated standard deviation of a population is when only the sample mean is known and the observations come from a normally distributed population.
Underfitting	A scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately generating a high error rate on both the training set and unseen data.
Univariant data set	Data sets with one variable.
Variance inflation factors (VIF)	The amount of multicollinearity in regression analysis.
Z score	A statistical measurement describing how far a value is from the mean of a group.

Author(s)

- Bhavika Chhatbar