Glossary

alternative hypothesis An hypothesis other than the null hypotheses. Also used to refer to hypotheses when there are several possible explanations for an observed pattern.

analysis of variance A statistical analysis that tests whether there is a significant difference among the means of different groups of numbers.

Anova An abbreviation for an analysis of variance.

bar graph A type of graph where values are represented with bars or columns; also called a bar chart or column chart.

categorical variable A variable that can can be placed into one of a few limited set of values or categories. For example, fur color in labrador retrievers can be categorized as black, yellow, or brown.

Chi-Square Test A statistical analysis that tests whether a set of observed values are statistically different from a set of expected values.

contingency table A table for organizing data that are characterized by two or more categorical variables. Contingency tables are often used for calculating expected values for comparison with observed values in a Chi-Square Test.

continuous variable A variable that is measured numerically and can have a wide range of values. For example, the pH of a solution can vary from 0 to 14 with many possible values along the pH scale.

correlation A method of analysis to explore the relationship between two variables. No cause and effect is hypothesized.

count A statistical parameter referring to the number of observations or data points in a group of numbers. Often synonymous with sample size. Count also refers to a categorical variable in which the numbers of individuals or observations that can be placed into different categories.

data transformation When a mathematical function is applied to data, such as taking the log or arcsine of all data in a data set. Usually this procedure is used so that the data conform more closely to a probability distribution such as the normal distribution.

degrees of freedom A statistical parameter that is related to sample size. In general, the greater the sample size, the greater the degrees of freedom and the greater the ability to detect statistical significance.

df An abbreviation for degrees of freedom.

dependent variable Usually the variable that the researcher is trying to explain. It is the variable that is affected by the independent variable. The dependent variable is sometimes called the response variable. Loosely speaking, it "depends on" the independent variable.

descriptive statistics Statistical parameters such as mean, maximum value, minimum value, etc. that are used to describe the location and spread of a group of numbers.

dummy variable In this manual, dummy variable refers to numbers that are entered into a spreadsheet in order to alter the appearance of a corresponding graph; they are often used to add labels or text to a graph. These numbers do not represent real variables that are part of a data set or statistical analysis.

equal variances Refers to the assumption that the variation within two or more groups of numbers is similar; an assumption for both the t-Test assuming equal variances and Anova.

error bars Vertical lines on a scatterplot or bar graph that represent the amount of variation in a group of numbers. They usually represent either the standard deviation or the standard error.

expected frequency The expected frequency of observations that should fall into a particular category. The sum of expected frequencies for a set of numbers should always equal 1.

expected value The expected number of observations that should fall into a particular category. The sum of the expected values for a set of numbers should always be the same as the sum of the observed values.

F critical In an analysis of variance or a regression analysis, the F critical corresponds to a 0.05 probability that random chance is causing the observed variation or pattern in a set of data. If the calculated value of F is $\geq$ F critical, then $p \leq 0.05$.

frequency distribution See histogram.

hidden data Data that are not represented in a scatterplot because they share identical x and y coordinates with other data points. In an MS Excel scatterplot, only one data point will be displayed in these situations, so hidden data should be represented by adding information to the scatterplot, such as numbers in parentheses near the data point indicating how many observations the point represents.

histogram Histograms (also called frequency distributions) show the frequency of different values in a set of data plotted as a bar graph. The height of the bars represents the frequency of the values on the x-axis. See Appendix V for examples and explanation.

independent Two events are independent if the probability of one occurring is not related to the probability of the other occurring.

independent variable A variable that affects the dependent variable. Also referred to as the predictor variable.

location of data Measures of location summarize where most of the data are found; examples include mean, median and mode.

mean In this manual, mean refers to arithmetic mean. The arithmetic mean is a measure of location that is calculated by taking the sum of all observations (in a group of numbers) and dividing by the number of observations.

mean square A measure of the spread in a group of numbers. In an Anova, the mean square for a particular category is calculated by dividing the sum of squares (SS) by the degrees of freedom.

median The value in a group of numbers that falls in the middle; half of the numbers fall below that value, the other half fall above.

mode The most common value in a group of numbers.

multiple regression A type of regression analysis that includes several independent variables.

negative relationship When there is a negative relationship between two continuous variables and the data are plotted using a scatterplot, a line drawn through the center of the scatter of points will slope downward from left to right (the slope of the line is negative).

non-linear regression A type of regression analysis that does not assume that the relationship between the dependent and independent variables is linear and therefore cannot be represented by a straight line.

non-parametric test A statistical test that does not assume that the data follow a particular probability distribution.

null hypothesis Null hypotheses are statements that any observed variability or pattern in the data is due to random chance. If $p > 0.05$ in the statistical test being performed there is "statistical significance" and the null hypotheses is accepted. If $p \leq 0.05$, there is "statistical significance" and the null hypothesis is rejected.

observation A term used to refer to an individual data point. A group of ten numbers has ten observations.

observed frequency The observed proportion or frequency at which the observations (or raw data) fall into particular categories. It is calculated by dividing the number of observations in a particular category by the total number of observations. The observed frequencies for a given set of data should always sum to a value of one.

observed value The number of observations (or counts) that fall into a particular category.

one-tailed test A type of t-Test where the direction of the difference between the means is predicted prior to analyzing the data; one mean is predicted to have a greater value than the other.

overlap The extent to which the minimum and maximum values of two groups of numbers are similar. If the two groups have identical maximum and minimum values, the two groups overlap entirely. If the maximum value of one group is less than the minimum value of another, there is no overlap.

paired test A type of t-Test that is performed when the observations in one group of numbers are paired with the observation in the second group. For example, if you are comparing the size of male and female birds of a particular species and you take all measurements on mating pairs, the data are naturally paired.

parametric test A type of statistical test that assumes that the data conform to a particular probability distribution.

pattern Any time the variation in a group of numbers is non-random (is related to another variable), there is pattern in the data. Often the goal of scientific research is to identify and explain pattern.

positive relationship When there is a positive relationship between two continuous variables and the data are plotted using a scatterplot, a line drawn through the center of the scatter of points will slope upward from left to right (the slope of the line is positive).

predictor variable An alternate term for the independent variable.

p-value The probability that the variation or observed pattern in the data is the result of random chance. The greater the p-value, the more likely the variation or observed pattern is the result of random chance and the less likely the independent variable is affecting the dependent variable. The lower the p-value, the more likely the independent variable is affecting the dependent variable. If $p \leq 0.05$, the result of a test is said to be statistically significant.

qualitative data Descriptive data that generally cannot be represented numerically and used in statistical analysis.

quantitative data Data that can be represented numerically and used in statistical analysis.

R^2 A statistical parameter in regression analysis that measures the amount of variation in the dependent variable that is explained by variation in the independent variable. In a scatterplot, the higher the R^2 , the more tightly the data will be clustered around the regression line. If all values fall on the line, $R^2 =1$. The value of R^2 can vary from 0 to 1.

random chance Random chance refers to the effect of random events on a data set. Every data set is subject to the effects of random chance. If the variability in a data set is all the result of random chance, then there is no meaningful influence of the independent variable on the dependent variable.

range The difference between the maximum and minimum values in a group of numbers.

regression A type of statistical analysis that tests for relationships between continuous independent and dependent variables. As used in this manual, the term regression is synonymous with simple linear regression.

regression line In a linear regression analysis, the estimated line that represents the relationship between the dependent and independent variables. The line is characterized by the two variables slope and intercept.

relationship In this manual, I use the term relationship to describe the possible pattern in the data in a regression analysis. If the independent variable accounts for at least some of the variation in the dependent variable, then there is a relationship between the two variables. A regression line with a positive slope indicates a positive relationship; a negative slope indicates a negative relationship.

research hypothesis In this manual, I use the term research hypothesis to distinguish from a null hypothesis. A research hypothesis is any statement that the researcher or scientist proposes to explain the effect of an independent variable on a dependent variable.

response variable A alternate term for the dependent variable.

scatterplot A graph in which data are represented as points in an x y coordinate system. The value of the independent variable is used for the x coordinate and the value for the dependent variable is used for the y coordinate.

spread of data Measures of spread summarize the variability in a group of numbers. If all numbers are near the mean (or median or mode), then there is little spread. If many numbers are far from the mean, there is a lot of spread. Measures of spread include the standard deviation and variance.

standard deviation A measure of spread in a group of numbers. It is equal to the square root of the variance.

standard error A measure of spread in a group of numbers. The standard error is smaller than the standard deviation so it is important to be clear about which measure is being reported.

statistical significance In a statistical analysis, if $p \leq 0.05$, the result is considered statistically significant. This means that the probability that random chance accounts for the variability or the observed pattern in the data is less than or equal to 0.05.

sum of squares (SS) An estimate of the spread or variation in a group of numbers. It is calculated as the sum of the square of the difference between each observation and the mean. This calculation is used in Anova.

t critical In a t-Test, t critical corresponds to a 0.05 probability that random chance is causing the difference in the means between the two groups. If the calculated value of t is $\geq$ t critical, then $p \leq 0.05$ and the difference between the means is considered statistically significant.

transformation See data transformation.

t-Test A statistical analysis designed to test whether the difference in the means of two groups of numbers is statistically significant.

two-tailed test A type of t-Test in which there is no prediction about which group of number has a greater mean value.

unequal variances Refers to the assumption that the variation within the two groups of numbers in a t-Test is not the same; an assumption for the t-Test assuming unequal variances.

unpaired test A type of t-Test in which there is no pairing between the observations in the two groups being compared.

variance A specific statistical parameter that is a measure of spread in a group of numbers. It is calculated by summing the square of the difference between each observation and the mean, then dividing by n-1 where n = the sample size.

variation A general term used to describe the spread in a group of numbers. In general, the greater the spread, the greater the variation.