Statistics foe Research Glossary

**Accuracy**

A term used in survey research to refer to the match between the target population and the sample.

**Adjusted R-Squared**

A measure of how well the independent, or predictor, variables predict the dependent, or outcome, variable. A higher adjusted R-square indicates a better model. Adjusted R-square is calculated based on the R-square, which denotes the percentage of variation in the dependent variable that can be explained by the independent variables. The adjusted R-squared adjusts the R-square for the sample size and the number of variables in the regression model. Therefore, the adjusted R-square is a better comparison between models with different numbers of variables and different sample sizes.

**Administrative Data**

Information about individual children, families, and/or providers of early care and education and other family benefits that are collected and maintained as part of the operation of government programs.

**Alpha Level**

The probability that a statistical test will find significant differences between groups (or find significant predictors of the dependent variable), when in fact there are none. This is also referred to as the probability of making a Type I error or as the significance level of a statistical test. A lower alpha level is better than a higher alpha level, with all else equal.

**Alternative Hypothesis**

The experimental hypothesis stating that there is some real difference between two or more groups. It is the alternative to the null hypothesis, which states that there is no difference between groups.

**Analysis of Covariance (ANCOVA)**

Same method as ANOVA, but analyzes differences between dependent variables.

**Analysis of Variance (ANOVA)**

A statistical test that determines whether the means of two or more groups are significantly different.

**Anonymity**  
An ethical safeguard against invasion of privacy whereby the researcher is unable to identify the respondents by their responses.

**Association**  
A relationship between objects or variables.

**Attrition**  
The rate at which participants drop out of a longitudinal study. If particular types of study participants drop out faster than other types of participants, it can introduce bias and threaten the internal validity of the study.

**Average**  
A single value (mean, median, mode) representing the typical, normal, or middle value of a set of data.

**Axiom**  
A statement widely accepted as truth.

**Bell-Shaped Curve**

A curve characteristic of a normal distribution, which is symmetrical about the mean and extends infinitely in both directions. The area under curve=1.0.

**Beta Level**

The probability of making an error when comparing groups and stating that differences between the groups are the result of the chance variations when in reality the differences are the result of the experimental manipulation or intervention. Also referred to as the probability of making a Type II error.

**Between-Group Variance**

A measure of the difference between the means of various groups.

**Between-Subject Design**

Experimental design in which a different group of subjects are used for each level of the variable under study.

**Bias**  
Influences that distort the results of a research study.

**Bimodal Distribution**

A distribution in which two scores are the most frequently occurring score. Interpretation of an average of biomodial distribution is problematic because the data represents non-normal distribution. Identifying biomodial distributions is done by examining frequency distribution or by looking at indices of skew or kutosis, which are frequently available with statistical software.

**Bootstrapping**  
A popular method for variance estimation in surveys. It consists of subsampling from the initial sample. Within each stratum in the sample, a simple random subsample is selected with replacement. This creates a finite number of new samples (or repetitions). The same parameter estimate is then calculated for each of the subsamples. The variance of the estimated parameter is then equal to the variance of the estimates from these subsamples.

**Case Study**

An intensive investigation of the current and past behaviors and experiences of a single person, family, group, or organization.

**Categorical Data**  
Variables with discrete, non-numeric or qualitative categories (e.g. gender or marital status). The categories can be given numerical codes, but they cannot be ranked, added, multiplied or measured against each other. Also referred to as nominal data.

**Causal Analysis**  
An analysis that seeks to establish the cause and effect relationships between variables.

**Ceiling**  
The highest limit of performance that can be assessed or measured by an instrument or process. Individuals who perform near to or above this upper limit are said to have reached the ceiling, and the assessment may not be providing a valid estimate of their performance levels.

**Census**  
The collection of data from all members, instead of a sample, of the target population.

**Central Limit Theorem**  
A mathematical theorem that is central to the use of statistics. It states that for a random sample of observations from any distribution with a finite mean and a finite variance, the mean of the observations will follow a normal distribution. This theorem is the main justification for the widespread use of statistical analyses based on the normal distribution.

**Central Tendency**  
A measure that describes the ¿typical¿ or average characteristic; the three main measures of central tendency are mean, median and mode.

**Chi Square**  
A statistic used when testing for associations between categorical, or non-numeric, variables. It is also used as a goodness-of-fit test to determine whether data from a sample come form a population with a specific distribution.

**Cluster Analysis**  
A type of multivariate analysis where the collected data are classified based on several characteristics in order to determine groups (or clusters) of cases that would be useful to explore further. This type of analysis can help one determine which groups of variables best predict an outcome.

**Cluster Sampling**  
A type of sample that is usually used when the target population is geographically disperse. First, clusters of potential respondents are randomly selected, and then respondents are selected at random from within the pre-identified clusters. For example, if it is prohibitively expensive to survey households that are spread out across the nation, a researcher may employ cluster sampling. The researcher would randomly select clusters of households, by randomly selecting several counties, and then the researcher would draw a random sample of households from within the selected counties. Clustered sampling designs necessitate the use of special variance estimation techniques.

**Codebook**  
Any information on the structure, content, and layout of a data set. The codebook typically provides background on the project, describes the data collection design, and gives detailed information on variable names and variable value codes.

**Codes**  
Values, typically numeric, that are assigned to different levels of variables to facilitate analysis of the variable. For example, codes such as strongly disagree=1, disagree=2, agree=3, and strongly agree=4 are often assigned.

**Coding**  
The process of assigning values, typically numeric values, to the different levels of a variable.

**Coefficient of Determination**  
A coefficient, ranging between 0 and 1, that indicates the goodness of fit of a regression model.

**Cohort**  
A group of people sharing a common demographic experience who are observed through time. For example, all the people born in the same year constitute a birth cohort. All the people married in the same year constitute a marriage cohort.

**Comparability**  
The quality of two or more objects that can be evaluated for their similarity and differences.

**Completion Rate**  
In survey research, this is the proportion of qualified respondents who complete the interview.

**Confidence Interval**  
A range of estimated values that is the best guess as to the true population’s value. Confidence intervals are usually calculated for the sample mean. In behavioral research, the acceptable level of confidence is usually 95%. Statistically, this means that if 100 random samples were drawn from a population and confidence intervals were calculated for the mean of each of the samples, 95 of the confidence intervals would contain the population’s mean. For example, a 95% confidence interval for IQ of 95 to 105, indicates with 95% certainty that the actual average IQ in the population lies between 95 and 105.

**Confidence Level**  
The percentage of times that a confidence interval will include the true population value. If the confidence level is .95 this means that if a researcher were to randomly sample a population 100 times, 95% of the time the estimated confidence interval for a value will contain the population’s true value. In other words, the researcher can be 95% confident that the confidence interval contains the true population value.

**Confidentiality**  
The protection of research subjects from being identified. A common standard in social science research is that records or information used for research should not allow participants to be identified and that researchers should not take any action that would affect the individual to whom the information pertains.

**Confounding Variable**  
A variable that is not of interest, but which distorts the results if the researcher does not control for it in the analysis. For example, if a researcher is interested in the effect of education on political views, the researcher must control for income. Income is a confounding variable because it affects political views and education is related to income.

**Consistency**  
The process in surveys whereby a question should be answered similarly to previous questions.

**Constant**  
A value that stays the same for all the units of an analysis. For instance, in a research study that explores fathers¿ involvement in their children¿s lives, gender would be constant, as all subjects (units of analysis) are male.

**Construct**  
A concept. A theoretical creation that cannot be directly observed.

**Construct Validity**  
The degree to which a variable, test, questionnaire or instrument measures the theoretical concept that the researcher hopes to measure. For example, if a researcher is interested in the theoretical concept of “marital satisfaction,” and the researcher uses a questionnaire to measure marital satisfaction, if the questionnaire has construct validity it is considered to be a good measure of marital satisfaction.

**Content Analysis**  
A procedure for organizing narrative, qualitative data into themes and concepts.

**Content Validity**  
Similar to face validity except that the researcher deliberately targets individuals acknowledged to be experts in the topic area to give their opinions on the validity of the measure.

**Context Effects**  
The change in the dependent variable which is resulted from the influence of the research environment. This influence is external to the experiment itself.

**Continuous Variable**  
A variable that, in theory, can take on any value within a range. The opposite of continuous is discrete. For example, a person’s height could be 5 feet 1 inch, 5 feet 1.1 inches, 5 feet 1.11 inches, and so one, thus it is continuous. One’s gender is either “male” or “female”, thus it is discrete.

**Control**  
The processes of making research conditions uniform or constant, so as to isolate the effect of the experimental condition. When it is not possible to control research conditions, statistical controls often will be implemented in the analysis.

**Control Group**  
In an experiment, the control group does not receive the intervention or treatment under investigation. This group may also be referred to as the comparison group.

**Control Variable**  
A variable that is not of interest to the researcher, but which interferes with the statistical analysis. In statistical analyses, control variables are held constant or their impact is removed to better analyze the relationship between the outcome variable and other variables of interest. For example, if one wanted to examine the impact of education on political views, a researcher would control income in the statistical analysis. This removes the impact of income on political views from the analysis.

**Controlled Experiment**  
A form of scientific investigation in which one variable, termed the independent variable, is manipulated to reveal the effect on another variable, termed the dependent or responding variable, while all other variables in the system are held fixed.

**Convenience Sampling**  
A sampling strategy that uses the most easily accessible people (or objects) to participate in a study. This is not a random sample, and the results cannot be generalized to individuals who did not participate in the research.

**Cooperation Rate**  
In survey research, this is the ratio of completed interviews to all contacted cases capable of being interviewed.

**Correlation**  
The degree to which two variables are associated. Variables are positively correlated if they both tend to increase at the same time. For example, height and weight are positively correlated because as height increases weight also tends to increases. Variables are negatively correlated if as one increases the other decreases. For example, number of police officers in a community and crime rates are negatively correlated because as the number of police officers increases the crime rate tends to decrease.

**Correlation Coefficient**  
A measure of the degree to which two variables are related. A correlation coefficient in always between -1 and +1. If the correlation coefficient is between 0 and +1 then the variables are positively correlated. If the correlation coefficient is between 0 and -1 then the variables are negatively correlated.

**Coverage**  
In survey research, this is the process of selecting a sample of individuals that reflect the larger population that the researchers wish to describe.

**Cross-Sectional Data**  
Data collected about individuals at only one point in time. This is contrasted with longitudinal data, which is collected from the same individuals at more than one point in time.

**Cross-Tabulation**  
A method to display the relationship between two categorical variables. A table is created with the values of one variable across the top and the values of the second variable down the side. The number of observations that correspond to each cell of the table are indicated in each of the table cells.

**Curvilinear**  
A statistical relationship between two variables that is not linear when plotted on a graph, but rather forms a curve.

**Data**  
Information collected through surveys, interviews, or observations. Statistics are produced from data, and data must be processed to be of practical use.

**Data Analysis**  
The process by which data are organized to better understand patterns of behavior within the target population. Data analysis is an umbrella term that refers to many particular forms of analysis such as content analysis, cost-benefit analysis, network analysis, path analysis, regression analysis, etc.

**Data Collection**  
The observation, measurement, and recording of information in a research study.

**Data Imputation**  
A method used to fill in missing values (due to nonresponse) in surveys. The method is based on careful analysis of patterns of missing data. Types of data imputation include mean imputation, multiple imputation, hot deck and cold deck imputation. Data imputation is done to allow for statistical analysis of surveys that were only partially completed.

**Deduction**  
The process of reasoning from the more general to the more specific.

**Deductive Method**  
A method of study that begins with a theory and the generation of a hypothesis that can be tested through the collection of data, and ultimately lead to the confirmation (or lack thereof) of the original theory.

**Degrees of Freedom**  
The number of independent units of information in a sample used in the estimation of a parameter or calculation of a statistic. The degrees of freedom limits the number variables that can be included in a statistical model. Models with similar explanatory power, but more degrees of freedom are generally prefered because they offer a simpler explanation.

**Dependent Variable**  
The outcome variable. In experimental research, this variable is expected to depend on a predictor (or independent) variable.

**Descriptive Statistics**  
Basic statistics used to describe and summarize data. Descriptive statistics generally include measures of the average values of variables (mean, median, and mode) and measures of the dispersion of variables (variance, standard deviation, or range).

**Dichotomous Variables**  
Variables that have only two categories, such as gender (male and female).

**Direct Effect**  
The effect of one variable on another variable, without any intervening variables.

**Direct Observation**  
A method of gathering data primarily through close visual inspection of a natural setting. Direct observation does not involve actively engaging members of a setting in conversations or interviews. Rather, the direct observer strives to be unobtrusive and detached from the setting.

**Discomfirming Evidence**  
A procedure whereby, during an open-ended interview, a researcher actively seeks accounts from other respondents that differs from the main or consensus accounts in critical ways

**Discrete Variables**  
A variable that can assume only a finite number of values; it consists of separate, indivisible categories. The opposite of discrete is continuous. For example, one’s gender is either “male” or “female”, thus gender is discrete. A person’s height could be 5 feet 1 inch, 5 feet 1.1 inches, 5 feet 1.11 inches, and so on, thus it is continuous.

**Discrimant Analysis**  
A grouping method that identifies characteristics that distinguish between groups. For example, a researcher could use discriminant analysis to determine which characteristics identify families that seek child care subsidies and which identify families that do not.

**Dispersion**  
The spread of a variable’s values. Techniques that describe dispersion include range, variance, standard deviation, and skew.

**Distribution**  
The frequency with which values of a variable occur in a sample or a population. To graph a distribution, first the values of the variables are listed across the bottom of the graph. The number of times the value occurs are listed up the side of the graph. A bar is drawn that corresponds to how many times each value occurred in the data. For example, a graph of the distribution of women’s heights from a random sample of the population would be shaped like a bell. Most women’s height are around 5’4″ This value would occur most frequently, so it would have the highest bar. Heights that are close to 5’4″, such as 5’3″ and 5’5″ would have slightly shorter bars. More extreme heights, such as 4’7″ and 6’1″ would have very short bars.

**Double Barreled Question**  
A survey question whereby two separate ideas are erroneously presented together in one question.

**Double Blind Experiment**  
A research design where both the experimenter and the subjects are unaware of which is the treatment group and which is the control.

**Dummy Coding**  
A coding strategy where each value of a categorical variable is turned into its own dichotomous variable. The dichotomous variable is coded as either 0 or 1. Dummy coding is used in regression analysis to measure the effect of a categorical variable on the outcome when the categorical variable has more than 2 values.

**Dummy Variables**  
Categorical variables that are assigned a value of 0 or 1 for use in a statistical analyses (see Dummy Coding).

**Duration Models**  
A group of statistical models used to measure the length of a status or process.

**Ecological Fallacy**  
False conclusions made by assuming that one can infer something about an individual from data collected about groups.

**Econometrics**  
A field of economics that applies mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory.

**Effect Size**  
A measure of the strength of the effect of the predictor (or independent) variable on the outcome (or dependent) variable.

**Endogeneity**  
A threat to the assumption that the independent (exogenous) variable actually causes the dependent (or endogenous) variable. Endogeneity occurs when the dependent variable may actually be a cause of the independent variable. Sometimes this is referred to as reverse causality. For example, a researcher may note that states with the death penalty also have high murder rates. The researcher may conclude that the death penalty causes an increase in the murder rate; however, it could be that states that experience a high murder rate are more likely to institute the death penalty. Endogeneity is the opposite of exogeneity.

**Epistemology**  
A way of understanding and explaining how we know what we know. Each research methodology is underpinned by an epistemology that serves as a guiding philosophy and provides a concrete process of research steps.

**Error**  
The difference between the actual observed data value and the predicted or estimated data value. Predicted or estimated data values are calculated in statistical analyses, such as regression analysis.

**Error Term**  
The part of a statistical equation that indicates what remains unexplained by the independent variables. The residuals in regression models.

**Estimated Sampling Error**  
The predictable and built-in level of error that accompanies all samples of a given size.

**Estimation**  
The process by which data from a sample are used to indicate the value of an unknown quantity in a population.

**Ethnographic Decision Models**  
A qualitative method for examining behavior under specific circumstances. An EDM is often referred to as a decision tree or flow chart and comprises a series of nested ¿if-then¿ statements that link criteria (and combinations of criteria) to the behavior of interest.

**Ethnographic Interviewing**  
A research method in which face-to-face interviews with respondents are conducted using open-ended questions to explore topics in great depth. Questions are often customized for each interview, and topics are generally probed extensively with follow-up questions.

**Ethnography**  
Literally meaning ¿folk¿ or ¿people¿ ¿writing,¿ ethnography is a field method focused on recording the details of social life occurring in a society. A primary objective is to gain a rich, ¿thick¿ understanding of a setting and of the members within a society. Ethnographers seek to learn the language, thoughts, and practices of a society by participating in the rituals and observing the everyday routines of the community. Ethnography is primarily based upon participant observation, direct observation, and in-depth interviewing

**Evaluation Research**  
The use of scientific research methods to plan intervention programs, to monitor the implementation of new programs and the operation of existing programs, and to determine how effectively programs or clinical practices achieve their goals.

**Exogeneity**  
The condition of being external to the process under study. For example, a researcher may study the effect of parental characteristics on their children’s behaviors. A parent’s religious upbringing is exogenous to their children’s behaviors because it is impossible for children’s current behavior to impact parent’s upbringing, which occurred prior to the birth of the child. The opposite of exogeneity is endogeneity.

**Experimental Control**  
Processes used to hold the conditions uniform or constant under which an investigation is carried out.

**Experimental Design**  
A research design used to establish cause-and-effect relationships between the independent and dependent variables by means of manipulation of variables, control and randomization. A true experiment involves the random allocation of participants to experimental and control groups, manipulation of the independent variable, and the introduction of a control group for comparison purposes. Participants are assessed after the manipulation of the independent variable in order to assess its effect on the dependent variable (the outcome).

**Experimental Group**  
In experimental research, the group of subjects who receive the experimental treatment or intervention under investigation.

**Explanatory Analysis**  
A method of inquiry that focuses on the formulating and testing of hypotheses.

**Exploratory Study**  
A study that aims to identify relationships between variables when there are no predetermined expectations as to the nature of those relations. Many variables are often taken into account and compared, using a variety of techniques in the search for patterns.

**External Validity**  
The degree to which the results of a study can be generalized beyond the study sample to a larger population.

**Extraneous Variable**  
A variable that interferes with the relationship between the independent and dependent variables and which therefore needs to be controlled for in some way.

**Extrapolation**  
Predicting the value of unknown data points by projecting beyond the range of known data points.

**Face Validity**  
The extent to which a survey or a test appears to actually measure what the researcher claims it measures. For example, a researcher may create survey questions that s/he claims measure gender role attitudes. To have face validity, other researchers who read the survey questions must also agree that the questions do appear to measure gender role attitudes.

**Factor Analysis**  
An exploratory form of multivariate analysis that takes a large number of variables or objects and aims to identify a small number of factors that explain the interrelations among the variables or objects.

**Field Notes**  
A text document that detail behaviors, conversations, or setting characteristics as recorded by a qualitative researcher. Field notes are the principle form of data gathered from direct observation and participant observation.

**Field Research**  
Research conducted where research subjects live or where the activities of interest take place.

**Field Work**  
Observing human behavior or interviewing individuals within their own communities. Field work is generally used in collecting qualitative data. It generally involves the researchers long-term relocation to the community under study. Data collection generally takes place over an extended period of time.

**Fixed Effects Regression**  
Regression techniques that can be used to eliminate biases associated with the omission of unmeasured characteristics. Biases are eliminated by including an individual-specific intercept term for all cases.

**Floor**  
The lowest limit of performance that can be assessed or measured by an instrument or process. Individuals who perform near to or below this lower limit are said to have reached the floor, and the assessment may not be providing a valid estimate of their performance levels.

**Focus Group**  
An interview conducted with a small group of people, all at one time, to explore ideas on a particular topic. The goal of a focus group is to uncover additional information through participants’ exchange of ideas.

**Forecasting**  
The prediction of the size of a future quantity (e.g., unemployment rate next year).

**Frequency Distribution**  
The frequency with which values of a variable occur in a sample or a population. To graph a distribution, first the values of the variables are listed across the bottom of the graph. The number of times the value occurs are listed up the side of the graph. A bar is drawn that corresponds to how many times each value occurred in the data. For example, a graph of the distribution of women’s heights from a random sample of the population would be shaped like a bell. Most women’s height are around 5’4″ This value would occur most frequently, so it would have the highest bar. Heights that are close to 5’4″, such as 5’3″ and 5’5″ would have slightly shorter bars. More extreme heights, such as 4’7″ and 6’1″ would have very short bars.

**GIS (Geographical Information Systems)**  
A computer system that enables one to assemble, store, manipulate, and display geographically referenced information.

**Generalizability**  
The extent to which conclusions from analysis of data from a sample can be applied to the population as a whole.

**Gini Coefficient**  
A measure of inequality or dispersion in a group of values (e.g.; racial inequality in a population). The larger the coefficient the greater the dispersion.

**Grounded Theory**  
The development of social science theory from the inductive analysis of data. This approach is generally used in qualitative research. The specific and detailed observations in the data are studied and understood to such an extent that a theory of more general patterns of behavior can be generated.

**Heterogeneity**  
The degree of dissimilarity among cases with respect to a particular characteristic.

**Heteroskedastic**  
A distribution characterized by a changing (non-constant) variance or standard deviation. Heteroskedasticity is problematic in statistical models because estimated standard errors will be inefficient and biased. Consequently, traditional significance test will not be valid.

**Hierarchical Linear Modeling (HLM)**  
A multi-level modeling procedure that works well for nested circumstances (e.g., estimating the effects of children nested within classrooms nested within schools). HLM enables a researcher to estimate effects within individual units, formulate hypotheses about cross level effects and partition the variance and covariance components among levels.

**Histogram**  
A visual presentation of data that shows the frequencies with which each value of a variable occurs. Each value of a variable typically is displayed along the bottom of a histogram, and a bar is drawn for each value. The height of the bar corresponds to the frequency with which that value occurs.

**Hypothesis**  
A statement that predicts the relationship between the independent (causal) and dependent (outcome) variables.

**Hypothesis Testing**  
Statistical tests to determine whether a hypothesis is accepted or rejected. In hypothesis testing, two hypotheses are used: the null hypothesis and the alternative hypothesis. The alternative hypothesis is the hypothesis of interest; it generally states that there is a relationship between two variables. The null hypothesis states the opposite, that there is no relationship between two variables.

**Imputed Response**  
A missing survey response that is filled in by the data analyst. The method to fill in the missing response is based on careful analysis of patterns of missing data. Imputation is done to allow for statistical analysis of surveys that were only partially completed.

**In-depth Interviewing**  
A research method in which face-to-face interviews with respondents are conducted using open-ended questions to explore topics in great depth. Questions are often customized for each interview, and topics are generally probed extensively with follow-up questions.

**Independence**  
The lack of a relationship between two or more variables. For example, annual snow fall and the Yankee’s season record are independent, but annual snow fall and coat sales are not independent.

**Independent Variable**  
The variables that the researcher expects to be the cause of an outcome of interest. For example, if a researcher wants to examine the effect of gender on income, gender is the independent variable. Sometimes this variable is referred to as the treatment variable or the causal variable.

**Independent and Identically Distributed (IID)**  
A collection of two or more random variables {X1, X2, . . . , } is independent and identically distributed if the variables are independent and also have the same probability distribution.

**Index**  
A type of composite measure that summarizes several specific observations and represents a more general dimension.

**Index Variable**  
A variable that is a summed composite of other variables that are assumed to reflect the same underlying construct.

**Indicator**  
An observation assumed to be evidence of the attributes or properties of some phenomenon. Indicators allow assessment of progress toward the achievement of intended outputs, outcomes, goals, and objectives.

**Indicator Variable**  
A variable that has two values, which are typically coded 0 and 1. Also referred to as a dummy variable.

**Indirect Effect**  
A condition where one variable affects another indirectly through an intervening variable. For example, gender may have an indirect effect on income if gender affects wage rates.

**Inductive Method**  
A method of study that begins with specific observations and measures, from which patterns and regularities are detected. These patterns lead to the formulation of tentative hypotheses, and ultimately to the construction of general conclusions or theories

**Informed Consent**  
The agreement between concerned parties about the data-gathering process and/or the disclosure, reporting, and/or use of data, information, and/or results from a research experiment.

**Instrument Error**  
A type of non-sampling error caused by the survey instrument (or questionnaire) itself, such as unclear wording, asking respondents for information they are unable to supply or the instrument being changed in some way during the course of the research.

**Inter-Rater Reliability**  
A measure of the consistency between the ratings or values assigned to a behavior that is being rated or observed; usually expressed as a percentage of agreement between two raters/observers, or as a coefficient of agreement which can be stated as a probability.

**Interaction Effect**  
A situation where the effect of the independent variable on the dependent variable varies depending on the value of another, additional variable. For example, teaching style and student’s gender would have an interactive effect if boys learned more in a lecture style classroom, while girls learned more in a discussion style classroom. In other words, the effect of teaching style on learning varies depending on student’s gender.

**Intercept**  
The expected value of a dependent variable when all the independent variables are equal to zero.

**Internal Validity**  
The extent to which researchers provide compelling evidence that the causal (independent) variable causes changes in the outcome (dependent) variable. To do this, researchers must rule other potential explanations for the changes in the outcome variable.

**Interval Scale**  
A scale of measurement where the distance between any two adjacent units of measurement is the same but the zero point is arbitrary. Scores on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided.

**Interval Variable**  
A variable wherein the distance between units is the same but the zero point is arbitrary.

**Intervention**  
The situation or variable introduced to the dependent variable; manipulations of the subject or the subject¿s environment that are performed for research purposes.

**Interviewer Error**  
A type of non-sampling error caused by mistakes made by the interviewer. These may include influencing the respondent in some way, asking questions in the wrong order, or using slightly different phrasing (or tone of voice) than other interviewers. It can include intentional errors such as cheating and fraudulent data entry.

**Jackknife Technique**  
A (usually) computer-intensive method to estimate parameters, and/or to gauge uncertainty in these estimates. The name is derived from the method that each observation is removed (i.e. cut with the knife) one at a time (or two at a time for the second-order Jackknife, and so on) in order to get a feeling for the spread of data.

**Kurtosis**  
A statistical equation that measures how peaked a distribution is. The kurtosis of a normal distribution is 0. If kurtosis is different than 0, then the distribution is either flatter or more peaked than normal.

**Least Squares**  
A commonly used method for calculating a regression equation. This method minimizes the difference between the observed data points and the data points that are estimated by the regression equation.

**Level of Significance**  
See significance level.

**Likert Scale**  
A scale that on which survey respondents can indicate their level of agreement or disagreement with a series of statements. The responses are often scaled and summed to give a composite measure of attitudes about a topic.

**Linear Regression**  
A statistical technique used to find a linear relationship between one or more (multiple) continuous or categorical predictor (or independent) variables and a continuous outcome (or dependent) variable.

**Literature Review**  
A comprehensive survey of the research literature on a topic. Generally the literature review is presented at the beginning of a research paper and explains how the researcher arrived at his or her research questions.

**Logistic Regression**  
A special form of regression used to analyze the relationship between predictor variables and a dichotomous outcome variable. A dichotomous variable is a variable with only two possible values, e.g. gender (male/female). Same as logit.

**Logit Model**  
A special form of regression used to analyze the relationship between predictor variables and a categorical outcome variable.

**MANOVA (Multivariate Analysis of Variance)**  
A statistical test that measures that varying group effects on many dependent variables.

**Main Effect**  
The effect of a predictor (or independent) variable on an outcome (or dependent) variable.

**Matched Samples**  
Two samples in which the members are paired or matched explicitly by the researcher on specific attributes, such as IQ or income. Also refers to samples in which the same attribute or variable is measured twice on each subject under different circumstances; also referred to as repeated measures.

**Maxima**  
The maxima are points where the value of a function is greater than other surrounding points.

**Mean**  
A descriptive statistic used as a measure of central tendency. To calculate the mean, all the values of a variable are added and then the sum is divided by the number of values. For example, if the age of the respondents in a sample were 21, 35, 40, 46, and 76, the mean age of the sample would be (21+35+40+46+76)/5 = 43.6

**Measurement Error**  
The difference between the value measured in a survey or on a test and the ¿true¿ value, if the difference is due to factors beyond the control of the respondent. Some factors that contribute to measurement error include the environment in which a survey or test is administered (e.g., administering a math test in a noisy classroom could lead students to do poorly even though they understand the material), poor measurement tools (e.g., using a ruler that is only marked in feet to measure height would lead to inaccurate measurement), rater effects (e.g., if a police man in uniform conducted interviews with individuals about drug use, they might not feel comfortable revealing their drug use.) There are many more such factors that can contribute to measurement error.

**Measures of Association**  
Statistics that measure the strength and nature of the relationship between variables. For example, correlation is a measure of association

**Median**  
A descriptive statistic used to measure central tendency. The median is the value that is the middle value of a set of values. 50% of the values lie above the median, and 50% lie below the median. For example, if a sample of individuals are ages 21, 34, 46, 55, and 76 the median age is 46.

**Member Checking**  
During open-ended interviews, the practice of a researcher restating, summarizing, or paraphrasing the information received from a respondent to ensure that what was heard or written down is in fact correct.

**Meta-Analysis**  
A statistical technique that combines and analyzes data across multiple studies on a topic.

**Methodology**  
The principles, procedures, and strategies of research used in a study for gathering information, analyzing data, and drawing conclusions. There are broad categories of methodology such as qualitative methods or quantitative methods; and there are particular types of methodologies such as survey research, case study, and participant observation, among many others.

**Metropolitan Statistical Area (MSA)**  
A term used by the U.S. Census Bureau to designate an area of adjacent counties (except in New England where they are defined by adjacent cities). Metropolitan Statistical Areas (MSAs) are often used to geographically understand labor markets because individuals often look for work outside of the city or county in which they live.

**Minima**  
The minima are points where the value of a function is less than other surrounding points.

**Missing Completely at Random (MCAR)**  
The term implies that all respondents are equally likely/unlikely to respond to the item and that the estimate is approximately unbiased. To ignore the missing data and restrict analyses to those records with reported values for the variables in the analysis, implicitly invokes the assumption that the missing cases are a random subsample of the full sample, that is, they are missing completely at random (MCAR). This is a strong assumption.

**Missing Data**  
Values in a data set values that were not recorded. Missing values can have many causes including a respondent’s refusal to answer survey questions, an interviewer incorrectly coding a response, or questions that do not apply to a respondent. The more missing data there are in a data set, the greater the likelihood of bias. There are several coding strategies that can “fill in” missing data for statistical analyses. These strategies are called imputation (see Data Imputation).

**Missing Data Imputation**  
A method used to fill in missing values (due to nonresponse) in surveys. The method is based on careful analysis of patterns of missing data. Types of data imputation include mean imputation, multiple imputation, hot deck and cold deck imputation. Data imputation is done to allow for statistical analysis of surveys that were only partially completed.

**Misspecification**  
Misspecification occurs when the predictor (independent) variables in a statistical model are incorrect. The most common cause of model misspecification is that important predictor (independent) variables are left out of the model. Misspecification often leads to incorrect estimates of the effects of the predictor (independent) variables that are included in the model on the outcome (dependent) variable.

**Mode**  
A descriptive statistic that is a measure of central tendency. It is the value that occurs most frequently in the data. For example, if survey respondents are ages 21, 33, 33, 45, and 76, the modal age is 33.

**Moving Average**  
A form of average which has been adjusted (or ¿smoothed¿) to allow for seasonal or cyclical components of a time series.

**Multicollinearity**  
A situation in which two or more predictor (independent) variables in a sample are highly related to each other. When using regression analysis, this can lead to incorrect estimates of their individual effects on the outcome (dependent) variable. Multicollinearity violates an underlying assumption of regression that each predictor (independent) variable has an independent impact on the outcome (dependent) variable.

**Multilevel Modeling**  
A model involving variables measured at more than one level of a hierarchy. An obvious hierarchy consists of children nested in classes, and classes nested in schools. Measurements can be obtained for child characteristics, class and teacher characteristics, or school characteristics. Multilevel models are also known as hierarchical linear models or random coefficient models. Multilevel are use to solve the statistical problems caused by dealing with hierarchically nested data.

**Multinomial Distribution**  
A distribution that arises when a response variable is categorical in nature. For example, if a researcher recorded the type of child care a child used, then the distribution of the counts in these categories would be multinomial. The multinomial distribution is a generalization of the binomial distribution to more than two categories. If the categories for the response variable can be ordered, then the distribution of that variable is referred to as ordinal multinomial.

**Multinomial Logit Model**  
A special form of regression used to analyze the relationship between predictor variables and a categorical outcome variable. The multinomial logit is used when the categorical outcome variable has more than two values, e.g., marital status could be never married, married, or divorced.

**Multiple (Linear) Regression**  
A statistical technique used to find the linear relationship between an outcome (dependent) variable and several predictor (independent) variables.

**Multivariate Analysis**  
Any of several statistical methods for examining more than one predictor (independent) variable or more than one outcome (dependent) variable or both. Allows researchers to examine the relation between two variables while simultaneously controlling for the influence of other variables.

**Multivariate Probit Model**  
The multivariate probit model is a generalization of the bivariate probit, which includes several distinct indicators as right hand side variables.

**Mutually Exclusive**  
Said of variables, events or conditions that can be placed into one category and no other. If there is no overlapping part between two events, we say they are mutually exclusive. However, mutually exclusive doesn¿t mean the two events are independent.

**Nominal Data**  
See categorical data.

**Nominal Scale**  
A scale that allows for the classification of elements into mutually exclusive categories based on defined features but without numeric value.

**Nonresponse Error**  
A type of error that is caused when a portion of the sample with particular characteristics do not respond to a survey. For example, individuals who are trying to dodge bill collectors might be less likely to answer their telephone and therefore may be less likely to respond to a telephone survey. This could lead to biased statistical results because individuals who do not pay their bills would be less likely to answer the survey. Researchers try to correct for this problem by determining the characteristics of those who were less likely to answer the survey and controlling for those characteristics in the analysis or by imputing missing data.

**Nonresponse Rate Bias**  
A source of bias that occurs when non-respondents differ in important ways from respondents.

**Nonsampling Error**  
Errors that can occur at any phase of the sampling process. Nonsampling error can result from nonresponse to surveys or from mismeasurement of survey responses.

**Nonsignificant Result**  
The result of a statistical test that indicates that there is not sufficient evidence to conclude that the predictor (independent) variable had an impact on the outcome (dependent) variable.

**Normal Curve**  
The bell-shaped curve that is formed when data with a normal distribution are plotted.

**Normal Distribution**  
This distribution describes a frequency distribution of data points that resembles a bell shape. (To graph a distribution, first the values of the variables are listed across the bottom of the graph. The number of times the value occurs are listed up the side of the graph. A bar is drawn that corresponds to how many times each value occurred in the data. See Frequency Distribution) In a normal distribution, the mean data point is the most likely data point to occur, data points that are equally higher or lower than the mean have an equal chance of occurring, and the farther a data point is from the mean the less likely it is to occur. The normal distribution exhibits important mathematical properties that are necessary for performing most statistical tests.

**Null Hypothesis**  
This hypothesis states that there is no difference between groups. The alternative hypothesis states that there is some real difference between two or more groups.

**Observation Unit**  
The actual unit observed during a study.

**Odds Ratio**  
A way to express a probability; the ratio of the odds of having a response or experience to the odds of not having it.

**Omitted Variable Bias**  
A form of bias in research resulting from the absense of key variables into the research design that would influence the results. When there is omitted variable bias, the results of the study could be due to alternative expalnations that are not addressed in the study.

**One-Way ANOVA**  
A test of whether the mean for more than two groups are different. For example, to test whether the mean income is different for individuals who live in France, England, or Sweden, one would use a one-way ANOVA.

**Open-Ended Data**  
Data derived from open-ended inquiries, such as interview questions, to which responses are not predetermined, such as would be the case with multiple choice or true/false questions.

**Ordinal Data**  
Data that is discrete categories, but that can also be ranked. For example, if a survey ask individuals whether they “strongly agree”, “agree”, “disagree”, or “strongly disagree” with a statement, the responses would be ordinal because they are in categories, but they can also be ranked.

**Ordinal Scale**  
A scale that allows for classification and labeling into mutually exclusive categories based on features that are ranked or ordered with respect to one another, although equal differences between numbers do not reflect an equal magnitude of difference.

**Ordinary Least Squares Estimation**  
A commonly used method for calculating a regression equation. This method minimizes the difference between the observed data points and the data points that are estimated by the regression equation.

**Outcomes**  
Measured behaviors; the behaviors that experimental research seeks to explain.

**Outlier**  
An observation in a data set that is much different than the other observations in the data set. The data point is unusually larger or an unusually smaller compared to the other data points.

**Oversampling**  
A sampling procedure in which a large proportion of subjects with a particular characteristic are sampled. Oversampling is used to ensure that researchers have enough data from groups with particular characteristics to yield good estimates for that group. For example, researchers often over sample African-Americans because just 12% of the population is African-American. This ensures that enough African-Americans are in the sample to yield good models and estimates for African-Americans.

**P-Value**  
The probability that the results of a statistical test were due to chance. A p-value greater than .05 is usually interpreted to mean that the results were not statistically significant. Sometimes researchers use a p-value of .01 or a p-value of .10 to indicate whether a result is statistically significant. The lower the p-value the more rigorous the criteria for concluding significance.

**Paired T-Test**  
This test is usually used to determine whether an intervention brought about a change in some characteristic of respondents (e.g., respondents’ math knowledge). To perform a paired t-test, respondents’ math knowledge would be measured prior to the intervention, then the intervention would be performed (e.g., teaching a class on math), then respondent’s math knowledge would be measured after the intervention. The change from before to after the intervention is used to assess whether the intervention was successful.

**Panel Study**  
A longitudinal study in which a group of individuals (a panel) is interviewed on several occasions over time.

**Parameter**  
A characteristic of a population.

**Participant Observation**  
A field research method whereby the researcher develops knowledge of the composition of a particular setting or society by taking part in the everyday routines and rituals alongside its members. A principle goal of participant observation is develop an understanding of a setting from a member¿s perspective, which may be accomplished through both informal observations and conversations as well as in-depth interviews.

**Participant-As-Observer**  
The investigator takes part in the group activity that the researcher plans to study. The researcher also reveals to the group that s/he is studying the group’s activities.

**Path Analysis**  
A special use of multiple regression to help understand and parcel out the sources of variance. Path analysis is a form of analysis that looks explicitly at cause.

**Pearson’s Correlational Coefficient**  
Usually denoted by r, this is a measure of the degree to which two variables are associated. Pearson’s correlational coefficient is used when the two variables are continuous. The coefficient can range from -1 to +1. If the coefficient is between 0 and +1, the variables are positively correlated, which means they both tend to increase at the same time. For example, height and weight are positively correlated because as height increases weight also tends to increases. If the coefficient is between 0 and -1, the variables are negatively correlated, which means as one increases the other decreases. For example, number of police officers in a community and crime rates are negatively correlated because as the number of police officers increase the crime rate tends to decrease. The closer the coefficient is to either -1 or +1, the stronger the association between the two variables. This is also called a Product Moment Correlation

**Percentage**  
A proportion times 100.

**Percentile**  
The percent of observations in a sample that have a value below a given score.

**Pile Sorting**  
A task used to elicit judgments of similarity among items in a specific domain. The technique uses a set of index cards on which the name or short description of a domain item is written; the respondent is asked to sort them into piles according to their similarity.

**Pilot Studies**  
A small scale research study that is conducted prior to the larger, final study. The pilot study gives researchers a chance to identify any problems with their proposed sampling scheme, methodology, or data collection process. These studies are very useful in accessing strengths and weakness of a potential study.

**Point Estimate**  
A statistic calculated from a sample that is an estimate of some single characteristic of the population. For example, the sample mean is the point estimate of the population mean.

**Poisson Distribution**  
A distribution that describes the number of events that occur in a certain time interval or spatial area. For example, the number of child care arrangements during a given period of time.

**Population**  
A clearly defined group of people or objects. Samples are drawn from the population and statistical results that are derived from random samples can be generalized to the whole population.

**Power**  
The degree to which a statistical test will detect significant differences between groups in a sample, when the differences do in fact exist. Sometimes statistical tests are not “powerful” enough to detect significant differences between groups in a sample that actually do exist in the population. The primary reason that a statistical test is not powerful is a small sample.

**Predictive Validity**  
A measure of whether a test assesses what is intended that is based on the correlation between the test score and some external criterion The higher the predictive validity, the more useful the test.

**Predictor Variable**  
The variable whose effect on an outcome variable is being modeled. A predictor variable is also called an “independent” variable.

**Pretesting**  
Measure taken at the outset of research, before the experimental manipulation or condition is applied or takes place.

**Primary Sampling Units**  
The pieces into which an area frame sampling divides land. It is these pieces, typically called PSUs, out of which a set of representative samples is taken.

**Probability**  
A description of the likely occurrence of a particular event. Probability is conventionally expressed on a scale from 0 to 1; a rare event has a probability close to 0, a very common event has a probability close to 1.

**Probability Sampling**  
A random sample of a population, which ensures that each member of the population has a chance of being selected for the sample.

**Probability of Selection**  
In probability samples, the probability of selection is the probability that a member of the population will be selected to participate in the study sample.

**Product Moment Correlation Coefficient**  
See Pearson’s Correlation Coefficient.

**Program Evaluation**  
Research that is conducted in order to determine the effectiveness of an intervention program.

**Projection**  
Estimates of the future size and other demographic characteristics of a population, based on an assessment of past trends and assumptions about the future course of demographic behavior.

**Proxy Variable**  
A variable used to ¿stand in¿ for another variable. Proxy variables are used when the variable of interest is not available in the data, either because it was not collected in the data or because it was too difficult to measure in a survey or interview.

**Purposive Sampling**  
A sampling strategy in which the researcher selects participants who are considered to be typical of the wider population. Since the sample is not randomly selected, the degree to which they actually represent the population being studied is unknown.

**Qualitative Research**  
A field of social research that is carried out in naturalistic settings and generates data largely through observations and interviews. Compared to quantitative research, which is principally concerned with making inferences from randomly selected samples to a larger population, qualitative research is primarily focused on describing small samples in non-statistical ways.

**Quartiles**  
A set of three values that divide the total frequency into four equal parts

**Quasi-Experimental Research**  
Research in which individuals cannot be assigned randomly to two groups, but some environmental factor influences who belongs to each group. For example, if researchers want to look at the effects of smoking on health, they cannot ethically assign individuals to a group that smokes and a group that does not smoke. Researchers might rely on some environmental factor, for example an ad campaign that discourages smoking, to examine changes in health following the campaign. The theory behind quasi-experimental designs is that following an environmental intervention, individuals’ characteristics play a smaller role in determining whether they smoke or do not smoke, and thus membership in these groups is closer to random assignment.

**Questionnaire**  
A survey document with questions that are used to gather information from individuals to be used in research.

**Quota Sampling**  
A sampling method in which interviewers are each given a quota of subjects of specified type to attempt to recruit. Widely used in opinion polling and market research.

**R-Squared**  
A measure of how well the independent, or predictor, variables predict the dependent, or outcome, variable. A higher R-square indicates a better model. The R-square denotes the percentage of variation in the dependent variable that can be explained by the independent variables. An Adjusted R-squared is a better comparison between models that have with different numbers of variables and different sample sizes than is the R-Squared. Please see Adjusted R-squared for more information.

**Random Coefficient**  
A variable that varies in ways the researcher does not control. For instance, if research subjects sign up for a study after seeing a posting asking for people between the ages of 20 and 24, age would not be a random coefficient, but factors such as gender and race would be.

**Random Error**  
An error that affects data measurements in a non-systematic way because of random chance.

**Random Sampling**  
A sampling technique in which individuals are selected from a population at random. Each individual has a chance of being chosen, and each individual is selected entirely by chance.

**Random Selection**  
A technique used to choose subjects at random so as to get a representative sample of the population. In random selection, each individual in the eligible population has a fixed and determinate probability of selection into the sample.

**Random Variable**  
A variable that numerically measures some characteristic of a sample, or population (e.g., height). The value of the variable will differ depending on which individual is measured (i.e., people are of different heights). The variable is said to be random because the variation in the value of the variable is due, at least in part, to chance (i.e., some people are just taller than other people).

**Randomization**  
Assigning individuals in a sample to either an experimental group or a control group at random.

**Range**  
A measure of dispersion of data. The range is calculated by subtracting the value of the lowest data point from the value of the highest data point.

**Rank Order**  
A scale of objects presented to research subjects,. whereby they are asked to rank the objects according to a specific criterion.

**Rating Scale**  
A rating scale is a measuring instrument for which judgments are made in order to rate a subject or case at a specified scale level with respect to an identified characteristic or characteristics.

**Ratio**  
The quotient of two values.

**Ratio Scale**  
A scale in which the difference between the values on the scale are equivalent and the scale has a fixed zero point; values on the scale can be meaningfully measured against each other.

**Raw Score**  
A score obtained from a test, assessment, observation, or survey that has not been converted to another type of score such as a standard score, percentile, ranking, or grade. By itself, a raw score provides little useful information about a subject.

**Refusal Rate**  
The percentage of contacted people who decline to cooperate with the research study. This is the opposite of the Response Rate.

**Regression Analysis**  
A statistical technique that measure the relationship between a dependent (outcome) variable and one or more independent (predictor) variables (see linear, logistic and multiple regression).

**Regression Coefficient**  
A coefficient that is calculated for each independent (predictor) variable. The regression coefficient indicates how much the dependent (outcome) variable will change, on average, with each unit change in the independent variables.

**Regression Equation**  
An mathematical equation that indicates the relationship between a dependent (outcome) variable and one or more independent (predictor) variables. The equation indicates the extent to which the dependent variables can be predicted by knowing the value of the independent variables.

**Reliability**  
The consistency and dependability of a survey question or set of questions to gather data. Reliability indicates the degree to which survey questions will provide the same result over time for the same person, across similar groups, and irrespective of who collects the survey data. A reliable set of questions will always give the same result on different occasions, assuming that what is being measured has not changed during the intervening period.

**Replicability**  
The degree to which a scientific investigation can be easily repeated to see if its findings and outcomes can be tested again or by others. Replicability is an ideal in social science research, and is related to the reliability of study findings.

**Representativeness**  
The idea that research subjects in a sample, as a group, represent the population from which the sample was selected.

**Research Method**  
Specific procedures used to gather and analyze data.

**Research Question**  
A clear statement in the form of a question of the specific issue that a researcher wishes to analyze.

**Respondent**  
The person who responds to a survey questionnaire and provides information for analysis.

**Response Categories**  
The valid values on a variable.

**Response Rate**  
The number of individuals who completed interviews divided by the number individuals who were originally asked or selected to be interviewed.

**Robustness**  
The state whereby a statistic remains useful even when one or more of its assumptions are violated.

**Sample**  
A group that is selected from a larger group (the population). By studying the sample the researcher tries to draw valid conclusions about the population.

**Sample Size**  
The number of subjects in a study. Larger samples are preferable to smaller samples, all else being equal.

**Sampling**  
The process of selecting a subgroup of a population that will be used to represent the entire population.

**Sampling Bias**  
Distortions that occur when some members of a population are systematically excluded from the sample selection process. For example, if interviews are conducted over the phone, only individuals with telephones will be in the sample. This could produce bias if the researcher intends to draw conclusions about the entire population, including those with a phone and those without a phone.

**Sampling Design**  
The part of the research plan that specifies how and how many respondents will be selected for a study.

**Sampling Distribution**  
The frequency with which data values appear in the sample. The sampling distribution can be characterized by the mean and the variance of the sample.

**Sampling Error**  
Fluctuation in the value of a statistic that is calculated from different samples that are drawn from the same population. For example, if several different samples of 5 people are drawn at random from the U.S. population, the average income of the 5 people in those samples will vary. (In one sample, Bill Gates may have been selected at random from the population, which would lead to a very high mean income for that sample.) It is not incorrect to have sampling error, and in fact statistical techniques take into account that sampling error will occur.

**Sampling Frame**  
A list of the entire population eligible to be included within the specific parameters of a research study.

**Scale**  
A group of survey questions that measures the same concept. For example, a researcher may be interested in individuals’ gender role attitudes, and use several questions to their attitudes. This group of questions make up a gender role attitude scale.

**Scaled Score**  
A mathematical transformation of a raw score so that scores can be compared across individuals and over time.

**Scatter Plot**  
A display of the relationship between two quantitative or numeric variables. A scatter plot shows the value of one variable plotted against the value of another variable.

**Selection Bias**  
Error due to systematic differences in the characteristics of those who are selected for a study and those who are not. For example, if a survey about health insurance is administered by randomly selecting patients who are waiting in doctors’ offices, only individuals who go to the doctor will be included in the sample. This will exclude individuals who do not go to the doctor and, therefore, introduce selection bias. Selection bias is a very serious problem in research, and it can negate research findings if the researcher does not carefully address the issue within the research study.

**Selective Observation**  
The act of only attending to observations that correspond to current belief.

**Semantic Differential Scale**  
A type of categorical, non-comparative scale with two opposing adjectives separated by a sequence of unlabelled categories.

**Semi-Structured Interview**  
A method of data collection in which the interviewer uses a pre-determined list of topics or questions to gather information from a respondent. The interviewer, however, may stray from the list to follow-up on things the respondent says during the interview.

**Significance Level**  
The probability that a relationship observed in statistical analyses were actually due to chance. The significance level is established before the statistical analysis is undertaken. If the statistical tests indicate that the chances of finding the observed results are higher than the set significance level, the results are “not significant.” Significance levels are usually set at .05, which means that significant results may actually be due to chance 5 out of 100 times.

**Simple Linear Regression**  
A statistical technique that measure the relationship between a dependent (outcome) variable and one independent (predictor) variable.

**Simple Random Sampling**  
The basic sampling technique where a group of subjects (a sample) for study is selected from a larger group (a population). Each individual is chosen entirely by chance and each member of the population has an equal chance of being included in the sample.

**Simulation**  
A process whereby a researcher uses either a table or a computer program to produce random digits to be used in studying random phenomena.

**Skewness**  
The tendency of a distribution to depart from symmetry or balance.

**Slope**  
The coefficient of the independent variable indicating the change in dependent variable per unit change in the independent variable.

**Snowball Sampling**  
A strategy used to gather a sample for a research study, in which study participants give the researcher referrals to other individuals who fit the study criteria. Snowball samples cannot be generalized to the population because they are not selected randomly. Snowball samples are usually used to investigate groups that have some unique, rare, or unusual quality and groups in which members know each other through an organization or common experience. For example, snowball samples might be used to identify marathon runners or cancer survivors who attend support groups.

**Social Desirability Bias**  
The tendency for respondents to give answers that are socially desirable or acceptable, that may not be accurate.

**Sociogram**  
A display of networks of relationships among variables, designed to enable researchers to identify the nature of relationships that would otherwise be too complex to conceptualize.

**Spurious Relationship**  
A statistical association between two variables is produced by a third variable rather than by a causal link between the two original variables. For example, children start school at the same time of year that the leaves begin to fall from the trees. This does not mean that leaves falling from trees affects when children start school or vice versa, instead both leaves falling from trees and children starting school occur during autumn.

**Standard Deviation**  
A measure of variability or dispersion of a set of data. The standard deviation (SD) is the square root of the variance. It is calculated based on the difference between each individual observation and the mean observation.

**Standard Error**  
A measure of the extent to which the sample mean fluctuates. The standard error is the standard deviation (SD) of the sample means. Conceptually, the standard error of the mean would be calculated by selecting multiple samples at random from a population, calculating the mean for each of the samples, then calculating the standard deviation of these sample means. Because only one sample is generally drawn from a population for a research study, the standard error is calculated by dividing the sample deviation by the number of the observations in the sample. Generally speaking, the larger the sample, the smaller the standard error.

**Standard Score**  
An indicator of the relative standing of a score within a normal distribution of scores, defined by its mean and standard deviation. By transforming raw scores to standard scores, the user can interpret intervals between any two score points in relation to a reference population.

**Standardization**  
A scale transformation procedure that involves manipulating data from different types of scales so that they can then be compared. It consists of subtracting the sample mean from each score and dividing by the standard deviation.

**Standardized Test**  
A form of measurement whose scores have been based on its administration within a specific population. Standardization is obtained by administering the test to a given population and then calculating means, standard deviations, standardized scores, and percentiles. Comparisons can then be made between an individual score and the group’s performance.

**Standardized Variables**  
The values of variables from different types of scales are transformed so that they can be compared with one another.

**Statistic**  
A measure of the characteristics of a sample (e.g., the mean is a statistic that measures the average of a sample). It gives an estimate of the same value for the population from which the sample was selected.

**Statistical Analysis**  
The principle of gathering data from a sample of individuals and using those data to make inferences about the wider population from which the sample was drawn.

**Statistical Control**  
A variable that is not of interest to the researcher, but which interferes with the statistical analysis. In statistical analyses, control variables are held constant or their impact is removed to better analyze the relationship between the outcome variable and other variables of interest. For example, if one wanted to examine the impact of education on political views, a researcher would control income in the statistical analysis. This removes the impact of income on political views from the analysis.

**Statistical Significance**  
If there is a very small probability that a relationship observed in statistical analyses is due to chance, the results are said to reach statistical significance. This means that the researcher concludes that there is a real relationship between the observed variables or a real difference between two groups. See Significance Level for additional information.

**Stratification**  
Grouping the study population into subgroups by their homogenous characteristics before sampling so as to improve the representativeness of a sample.

**Stratified Sampling**  
A statistical method for testing different theorized models, including the “structures” of relationships among the observed indicators and their underlying concepts.

**Stratified sampling**  
A type of probability sample where the units in a population of interest are divided into mutually exclusive and collectively exhaustive strata and a (proportionate or disproportionate) random sample is drawn from each stratum.

**Structural Model**  
A theoretical model assumed to underlie the data that expresses the relationship between the dependent variable and the independent variables.

**Structured Interview**  
The interviewer asks respondents the same questions using a predetermined series of interview questions. Deviations from the predetermined series of questions are not allowed in the interview process.

**Subjectivity**  
A reflection of the person¿s mind, or thoughts, which is the result of his/her experiences, moods or attitudes.

**Subjects**  
Those who participate in research and from whom data are collected.

**Subsample**  
A sample selected from a sampling frame, that is itself a sample of the population. Often the original sample is broken down into smaller groups (subsamples) to be sampled at different rates.

**Survey Research**  
A research approach designed to systematically collect data about a group of individuals. Data are obtained through direct questioning, using either a written or oral surveys, of a sample of respondents.

**Systematic Bias**  
When the recorded data from a sample is systematically higher or lower than the true data values within the population. Systematic bias can occur as a result of sampling bias or measurement bias. Sampling bias is an error in sampling when some subgroup of the target population is unintentionally left out of the sampling process. Measurement bias is an error in data collection when some occurrence distorts the responses in the same way (e.g., a test is administered in a noisy classroom). Bias is a serious error in data collection and should be handled through a researcher’s careful attention to sources of bias.

**T Distribution**  
A symmetrical bell-shaped distribution that is used for testing samples smaller than 30 or where the variance is unknown.

**T-Test**  
A statistical test that is used to compare the means of two samples or the mean of one sample with some fixed value. The test is appropriate for small sample sizes (less than 30).

**Target Population**  
The population to which the researcher would like to generalize her or his results based on analysis of a sample. The sample is selected from a target population.

**Test-Retest Reliability**  
The degree to which a measure produces consistent results over several administrations.

**Theoretical Sampling**  
The selection of individuals within a naturalistic research study based on emerging findings as the study progresses to ensure that key issues are adequately represented.

**Theory**  
General statement that describes a hypothesized relationship between different phenomena or characteristics. Theories should be specific enough to be testable with a well-designed research study.

**Time Series**  
A sequence of observations which are ordered in time or space.

**Treatment Effect**  
Change in the outcome variable that is due to some intervention. Sometimes used to describe the change in an outcome variable that is due to changes in the independent (predictor) variable, even if the independent variable is not an intervention.

**Triangulation**  
Using of a variety of research methods to gain a diverse source of data pertaining to a specific research problem or question. This process helps to enhance the validity of data since it does not overly rely on any particular method.

**Two-Tailed Test**  
A type of test that is used when a researcher is unsure of whether the independent (predictor) variable has a positive or negative effect on the dependent (outcome) variable.

**Two-Way ANOVA**  
A statistical test to study the effect of two categorical independent variables on a continuous outcome variable. Two-way ANOVAs analyze the direct effect of the independent variables on the outcome, as well as the interaction of the independent variables on the outcome.

**Type I Error**  
An error that occurs when a researcher concludes that a statistically significant relationship between two variables exists (based on the analysis of the sample), when in fact it the relationship does not exist in the population from which the sample was selected. The probability of making a type I error is decided at the outset of the statistical analysis. This probability is also called a significance level.

**Type II Error**  
An error that occurs when a researcher concludes that no significant relationship between two variables (based on analysis of sample data) when in fact the relationship does exist in the population from which the sample was drawn. The probability of not making a type II error is also called the power of a statistical test.

**Typology**  
A descriptive, categorical scheme that may serve as the foundation for data coding or data analysis in a qualitative study. The types or categories comprising a typology should be mutually exclusive ¿ separate and distinct with no overlap ¿ and exhaustive ¿ all the data must fall into some category.

**Unbalanced Scale**  
A scale where the number of favorable and unfavorable categories is not the same.

**Unbiased**  
A statistic that is free of systematic bias. Systematic bias occurs when the recorded data from a sample is systematically higher or lower than the true data values within the population. Systematic bias can occur as a result of sampling bias or measurement bias. Sampling bias is an error in sampling when some subgroup of the target population is unintentionally left out of the sampling process. Measurement bias is an error in data collection when some occurrence distorts the responses in the same way (e.g., a test is administered in a noisy classroom). Bias is a serious error in data collection and should be handled through a researcher’s careful attention to sources of bias.

**Unit of Analysis**  
The person, group of people, or object that is under study. For example, if an analysis examines children’s well-being, children are the unit of analysis. If an analysis examines family income, families are the unit of analysis.

**Univariate Analysis**  
Examination of the properties of one variable only and not the relationship between variables. Generally univariate analysis is performed by examining the mean and standard deviation of a variable.

**Unstructured Interview**  
An interview in which the researcher asks open-ended questions. The researcher aims to give respondents the latitude to talk freely on a topic and to influence the direction of the interview. There is no predetermined plan about the specific information to be gathered from these types of interviews.

**Validity**  
The degree to which data and results are accurate reflections of reality. Validity refers to the concepts that are investigated, the people or objects that are studied; the methods by which data are collected; and the findings that are produced.

**Variable**  
An attribute or characteristics of a person or an object that varies within the population under investigation (e.g. age, weight, IQ, child care type).

**Variance**  
A commonly used measure of dispersion for variables. The variance is calculated by squaring the standard deviation. The variance is based on the square of the difference between the values for each observation and the mean value.

**Weighted Score**  
A score adjusted by such factors as the importance of the attribute assessed or the reliability and validity of the assessment from which the score was derived, or a combination of such factors.

**Weighting**  
A process used to ensure that statistics produced from a sample are representative of the population from which the sample was drawn. Most large surveys include weights and descriptive statistics produced from such surveys should be based on weighted data.

**Z Score**  
A score that is produced by subtracting the mean value from an individual data value and dividing by the standard deviation. This standardizes data values and allows for individual data values from different distributions (distributions with different means and standard deviations) to be compared.

**Z Test**  
A statistical test that is used to compare the means of two samples or the mean of one sample with some fixed value. The test is appropriate for larger samples (over 30) and for smaller samples in which the variance of the population is known.