### Table 1: Statistical Techniques, Descriptions, and Applications

| **Statistical Technique** | **Description** | **Data Requirements** | **Sample Application** |
| --- | --- | --- | --- |
| Descriptive Statistics (distribution, mean, median, mode) | These statistics describe variation and the central tendencies of variables.  The distribution is the number and percent of respondents in any given category of a variable.  The mean is the arithmetic average.  The median is the value that is the middle (50th) percentile of the distribution.  The mode is the most common category of the distribution.  Use descriptive statistics in exploratory research, to evaluate the feasibility of testing hypotheses, and to check assumptions of inferential statistics. | Any continuous (numeric) or categorical (e.g., gender, race, or yes/no) variables. Not all statistics are calculable for all types of variables. | What percent of program participants (distribution) learned about the program from a friend versus a family member?  What was the average (mean) number of visits by a social worker to a family?  What was the most common way (modal category) that tenants learned about a program? |
| Correlation | An examination of the strength and direction of a relationship between two continuous variables. This statistic shows whether or not the two are statistically related, but does not present any causal link. | Two continuous variables, such as age, income, years of education, or number of days in program. | Is greater program participation as measured by number of modules completed associated with years of education? |
| Chi-Square and Fisher’s exact test (and related tests of association within cross tabulations) | An examination of two categorical variables to assess whether the combinations of characteristics are due to chance. | Two categorical variables with two (or more) possible categories, such as gender, program participation (yes/no), or race/ethnicity. | Is there a difference between men and women in seeking mental health services (yes/no) in the program area? |
| T-Test (two group comparison) | A comparison of the means of two groups based on a single variable (that has only two values, such as yes/no) to see if observed differences in groups are larger than would be expected due to chance. | Variables that are continuous, such as test scores, indices from mental health instruments, income levels, or that can be divided by a single categorical variable with only two possible categories, such as gender or program participation. | Is program participation (yes/no) related to average score on an intake instrument? |
| ANOVA (and related techniques) | A comparison of means of an outcome variable is expected to vary based on one other variable (with 3 or more categories) or based on multiple other variables. It is used to determine if the means of the various groups of the target variable (based on a second variable) differ statistically. Related models such as MANOVA can use multiple dependent variables, while ANCOVA (and MANCOVA) use multiple independent covariates. | A categorical variable with at least three categories, such as types of program participated in, and a continuous variable, such as test scores, BMI, or income in dollars. Other model forms may include more variables. | Does participation in one of three different exercise programs affect young people’s BMI after completion of the programs? |
| Linear and Hierarchical Regression (and related regression techniques) | A model used to examine the relationship between one or more predictive (independent) variables and a continuous outcome (dependent) variable. This model allows for statistically accounting for the effects of multiple factors on a single outcome. | A continuous variable that is an outcome of interest, such as test scores, number of days in a program, or number of visits to a health care provider, and at least one continuous variable (and the model may include more continuous or categorical variables as well). | What factors, such as age, gender, or education level, predict test scores on a knowledge exam of health behaviors? |
| Logistic regression (and related techniques) | A model used to examine the relationship between independent variables and a dichotomous categorical dependent or dependent variable with limited categories. This model allows for statistically accounting for the effects of multiple factors on a single outcome. | A categorical variable that is an outcome of interest, such as completed program or not, and at least one continuous variable and/or categorical variable(s)). | Assess what factors, such as age, gender, or education level, are related to attending an afterschool program versus not attending. |
| Random effects models | A model that estimates program effects across a sub-level of variables (i.e., within-group; time points within-individuals). Using a regression model that predicts reading test scores, a researcher may estimate the time period before and after the treatment as a random effect. The model thus account for changes over time. The model assumes that the time period is uncorrelated with any of the other predictor variables as well as the outcome variable. | Multiple data points collected at the same level (e.g., person). Examples of outcomes (dependent variables) include test scores from students taken in three different years, blood cell counts taken multiple times across several months, or employment status documented every month for one year. | Was program participants’ change in reading over time significantly better than non-program participants? |
| Fixed effects models | A form of regression analysis that holds constant factors which do not change over time, allowing for analysis of factors that do change over time. This model allows for statistically accounting for the effects of multiple factors on a single outcome. For more information see: <http://harrisschool.uchicago.edu/Blogs/EITM/wp-content/uploads/2011/06/Fixed-Effects-Lecture1.pdf> | Data collected from the same subjects at least twice (and preferably multiple times). Examples of outcomes (dependent variables) include test scores from students taken in three different years, blood cell counts taken multiple times across several months, or employment status documented every month for one year. Any time-invariant factors are removed from the analysis algebraically, and do not need to be measured. | What factors were associated with a net increase in income levels over time for a group of unstably employed individuals, once any time-invariant factors such as race and gender are accounted for? |
| Mixed Models | Models that include both fixed and random effects. This model will include fixed effects which eliminates omitted variable bias (via fixed effects) by accounting for unobserved variables via fixed effect, and random effects which assumes that unobserved differences between individuals are the same over time. For more information see: <http://www.stat.cmu.edu/~hseltman/309/Book/chapter15.pdf> | Data collected from the same subjects at least twice (and preferably multiple times). Examples of outcomes (dependent variables) include test scores from students taken in three different years, blood cell counts taken multiple times across several months, or employment status documented every month for one year, and comparison group data. Time invariant factors are measured and included in the dataset. | Was program participants’ change in reading over time significantly better than non-program participants? What effect do time varying and invariant factors have on the outcomes? |
| Difference-in-differences models | An analysis technique where outcomes are observed for two groups over two time periods. One of the groups is exposed to some desired treatment in the second time period, but not in the first time period. The other group is not exposed to the treatment in either time period. The average gain in the second (control) group is subtracted from the average gain in the first (treatment) group. For more information, see: http://econ.lse.ac.uk/~amanning/courses/ec406/ec406\_DinDPanel.pdf. | Data collected from the same subjects at least twice (and preferably multiple times). Examples of outcomes (dependent variables) include test scores from students taken in three different years, blood cell counts taken multiple times across several months, or employment status documented every month for one year, AND comparison group data. | Was program participants’ change in reading over time significantly better than non-program participants? |
| Hierarchical Linear Models (HLM) | A form of regression that takes into account the unique contribution of variables within and across hierarchical levels to predict a single outcome. This model allows for statistically accounting for the effects of multiple factors on a single outcome. | Independent variables within a level that are nested within one another, such as students within classrooms within schools or neighborhoods. The dependent variable can be continuous or categorical. | How well did students do on a test within an entire school district, taking into account program participation, different classroom experiences, and different school level programs? |
| Path Analysis | A series of regression models formulated into a path diagram that is used to describe both the size and directionality of the relationship between predictor variables and an outcome variable. It is used to test theories of causal relationships between variables. This model allows for statistically accounting for the effects of multiple factors on a single outcome. A model that is used to describe both the size and directionality of the relationship between predictor variables and an outcome variable. This model allows for statistically accounting for the effects of multiple factors on a single outcome. | Continuous variables that are outcomes of interest, such as test scores, number of days in a program, or number of visits to a health care provider, and at least one continuous variable (and the models may include more continuous or categorical variables as well). | Given that family stability, education level, and previous level of service utilization all contribute to predicting a score on a psychometric instrument, which relationships are causally most important? |
| Structural Equation Modeling | A modeling technique that takes into account observed (or measured) variables and latent (or unmeasured) variables by specifically modeling measurement error. It is used to examine causal relationships between variables, taking into account the influence of unmeasured variables. | Continuous variables that are outcomes of interest, such as test scores, number of days in a program, or number of visits to a health care provider, and at least one continuous variable (and the model may include more continuous or categorical variables as well). | Can the outcome of increased college matriculation be assessed, taking into account both known characteristics, such as test scores and parental education, as well as unobserved measures, such as motivation and ambition? |