Data Analysis Basics – Part II

Hello! I'm Judy Savageau from the Center for Health Policy and Research at UMass Medical School following up on yesterday's post with Part II of basic data analyses. A number of posts outlining statistical/analytic details are in AEA365's archives. For example, there are some great posts on "Readings for Numbers People (or Those Who Wish They Were)", "Starting a Statistics [Book] Club", and "Explaining Statistical Significance". These posts discuss multivariate modeling, longitudinal data analysis, propensity score matching, factor analysis, structural equation modeling, and more. But what defines multivariate analyses and how do they differ from bivariate analyses?

Hot Tip:

Decisions about bivariate statistics (i.e., assessing the relationship between 2 variables; e.g., gender and school performance) are made based on the 'type' of data (e.g., categorical vs continuous; see yesterday's Part I post). There are many reputable resources that show simple tables for determining which statistic to use (see Rad Resources below), including:

- Chi-square test: 2 categorical variables (e.g., program participation: yes/no and job type)
- T-test: 1 categorical variable with 2 levels (e.g., gender: male/female) and 1 continuous variable (e.g., IQ, SAT scores)
- ANOVA – Analysis of Variance: 1 categorical variable with 3 or more levels (e.g., program performance: low / moderate / high) and 1 continuous variable (e.g., years of education)
- Correlation coefficient: 2 continuous variables (e.g., years of employment and number of correct responses to knowledge about job-related standards)

Hot Tip:

Finally, use multivariate analyses when you want to look at a large number of variables and their relationship (collectively) to one outcome. The most appropriate multivariate statistic depends, in large part, on the categorical or continuous nature of the outcome variable. For example, in one federally-funded study assessing the multiple factors related to return to work after a work-related injury (e.g., severity of injury, years until anticipated retirement, pre-injury job satisfaction, employer assessment of re-injury potential, etc.), our outcome variable was 'return to work' measured in multiple ways:

- Categorical measure: return to work – Yes/No. To determine which factors are most predictive of whether or not a person with a work-related injury will come back to work might best be explored using logistic regression.
- Continuous measure: how quickly (in weeks) might a person return to work following a work-related injury might best be explored using linear regression.

There are many decisions to be made when developing a data analysis plan. I'm hoping that this 2-part introduction to the basics of statistical analyses gets you started in thinking about the best way to explore and analyze your quantitative data. Of course, having a statistician/data analyst sitting 'at the table' with the team as early as possible will ensure that you collect data in the best format to answer your research questions.

Rad Resources:

Here are just a couple of web pages that help with some decision-making about when it's most appropriate to choose one statistical test over another – depending on the type of data you have.

- http://www.biostathandbook.com/testchoice.html
- https://stats.idre.ucla.edu/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/