Analysis at Analyst Institute

Our experiments generally proceed from data acquisition and sample selection to randomization, implementation, and then analysis. This page is designed to be a quick overview of the important elements of that final analysis phase of an experiment.

In particular, this page will detail the analysis of an experiment measuring the effectiveness of a persuasion treatment (a treatment designed to cause the opinion change, which is usually measured through phone or online surveys).

**Data Format**

Experimental data may arrive from a variety of different sources. For example, the individuals in the sample may come from by a partner organization, the characteristics of the individual (covariates) from a data vendor, and outcome data from a survey firm. These files are usually share a common identifier that is unique to individuals.

**Data Cleaning**

Once you have data, you need to get it into a form suitable for analysis. Data cleaning may involve matching different data sets together, recoding covariates and outcomes, and removing duplicate observations of the same person (deduping). It’s particularly important to ensure that the data files match on a common id variable, that individuals are not unexpectedly dropped or duplicated, and that variables are re-coded correctly. Make sure that you are always looking at your data and running crosstabs, etc.

**Analysis**

We are generally estimating the average treatment effect in our sample. Often, we are also estimating conditional average treatment effects for various subpopulations in the sample, e.g., among women, or people under 35. When conducting analysis, it’s important to ensure that the treatment is being measured correctly, that covariates are being chosen and used sensibly, that an appropriate transform of the outcome data is being used (possibly the raw data), and that an appropriate statistical technique is being applied.

**Graphing**

We do most of our graphing in R. In general, we produce bar plots of average treatment effects including error bars. These graphs are included in reports sent to outside partners, so we expect the output of graphing in R to be polished and suitable for outside parties.

**Coding Style**

We generally follow Hadley Wickham’s [style guide](http://adv-r.had.co.nz/Style.html) for R code. In particular, variable names should be meaningful and consistent, and code should be well-documented. Basically, ask yourself, “if another analyst were reading my code, would s/he understand what is going on?”