BIOSTAT 702: Module 1

Selecting and Describing Study Participants; Part 2: Table 1

Dr. Marissa Ashner

Department of Biostatistics and Bioinformatics

Fall 2025

Duke University School of Medicine

Module Goals

- Be able to create a descriptive ‘Table 1’ from a randomized trial or observational study
- Understand why p-values should not be used to evaluate between-group differences in baseline prognostic factors in a randomized trial
- Understand why p-values should not be used to assess the presence of selection bias or confounding in a non-randomized study

Resources for this Module

Textbooks

- RMPH: Chapter 3

Websites

- Article: Guidelines for a Useful Table 1
- Article: Comparability of Randomised Groups
- Article: An Improper Use of Statistical Significance
- gtsummary Vignette
- tableone Vignette
- table1 Vignette
- SMD For Multi-Level Categorical Variables
- SMD with table1 package

Exploratory Data Analysis / Descriptive Statistics

- Before you do *any* sort of analysis with your data, the first thing that should be done is exploratory data analysis / descriptives
- *Visualize your data*: look at histograms / boxplots for continuous variables, bar charts for categorical, scatterplots if you are interested in any bivariate relationships, etc.
- *Check for Unusual Values*: Look at the minimum, maximum, spread of the data; do all the values align with what is reasonable?
- *Check for Missing Data:* Is any data missing? You need to decide what to do if it is
    - *Note:* In this class, we will simply remove all observations with missing data after describing the amount of missingness

The Famous ‘Table 1’

- In most published articles, the standard way to display quantitative descriptive statistics is with “Table 1”,
  - which is named as such because it is usually the first table in the results section
- Typically,
  - Continuous Variables have the mean and standard deviation displayed
  - Categorical Variables have the frequency and percent/proportion displayed
  - The proportion of missing values for each variable is also displayed
- It is typical to have the variables take up the rows of the table, while there is a column for the “Overall” stats of the analytic cohort and sometimes columns stratified on a predictor/outcome

Coding Table 1

- Do not do this “by hand”
- There are several packages in R to do this:
  - `gtsummary`
  - `tableone` (best for including SMD)
  - `table1` (my personal package of choice)

What are the p-values typically seen in Table 1?

- If you have columns stratifying by a predictor/outcome, the p-values on the right hand side of the table typically display the results from the hypothesis test where the null is that the mean (or proportion) is equal in all groups of that stratification variable
- *Why include them?*: Some people include them to provide unadjusted tests of association between the stratification variable and all the other variables
    - This can serve as an attempt to quantify potential confounders or check for imbalance across groups
    - Or as an attempt to check for selection bias (if you stratify on inclusion in the analysis dataset)

To include or not to include p-values?

- **Randomized Trials**
  - Do not include!!!
  - You are testing whether differences across treatments in certain characteristics are significant when you designed the study such that this should not be true – nonsensical
  - p-values do not indicate whether or not any group imbalance will actually affect the analysis results
- **Observational Studies**
  - Do not include!!!
  - In terms of confounding, you are not testing what is actually important in terms of choosing a confounder

For both, results are very dependent on sample size!!

Standardized Mean Differences (SMD)

→ bias

- defined as the mean difference between two groups divided by the average variability among the groups
- helpful for looking at balance of various factors between groups
- better than simply comparing point estimates between groups because the standardized difference takes into account the amount of variability in the sample
- In simplistic terms, it is a measure that compares the signal (the mean difference) to the noise in the data (the average variability among the groups).
- In general, a standardized difference of 0.2 or higher indicates a potential concern for bias

Calculating SMD

- For continuous variables this is defined as:

$$\boxed{d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}}$$

where $\bar{x}_1$, $\bar{x}_2$ and s_1^2, s_2^2 are the means and variances, respectively, of the two groups being compared.

- To compare proportions between groups use the following:

$$\boxed{d = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)}{2}}}}$$

- The standardized difference for a multi-level categorical variable is more complex (example on resources page)