Describing Distributions of Data

Contents

Last Updated: Wed Jul 27 8:31:31 PM

Introduction

Visualizing data is one of the most important things we can do to become familiar with the data. There are often features and patterns in the data that cannot be uncovered with summary statistics alone. There tends to be two forms in which data can be presented; Summary tables are used for comparing exact values between groups for example, and plots for conveying trends and patterns when exact numbers are not always necessary to convey a story.

Levels of care

There are three levels of visualizations that can be created, with examples shown in PMA6 Figure 4.1 a, b and c.

1. **For your eyes only** Made by the analyst, for the analyst, these plots are quick and easy to create, using the default options without any annotation or context. These graphs are meant to be looked at once or twice for exploratory analysis in order to better understand the data.
2. **For an internal report** Some chosen plots are then cleaned up to be shared with others, for example in a weekly team meeting or to be sent to co-investigators participating in the study. These plots need to be capable of standing on their own, but can be slightly less than perfect. Axis labels, titles, colors, annotations and other captions are provided as needed to put the graph in context.
3. **For publication or external report** These are meant to be shared with other stakeholders such as the public, your collaborator(s) or administration. Very few plots make it this far. These plots should have all the "bells and whistles" as they appear in formal reports, and are often saved to an external file of a specific size or file type, with high resolution. For publication in most printed journals and books, figures typically need to be in black and white (possibly grayscale).

Along with having the audience in mind, it is important to give thought to the purpose of the chart.

> The effectiveness of any visualization can be measured according to how well it fulfills the tasks it was designed for. (A. Cairo, 2018).

Why graph single variables?

Graphs are the first line of defense against data problems that won't show up with numerical summaries.

- Are there categories or entries that you don't expect?
  - Did you forget to set a code like -77 or 99 to missing?
- Are there major imbalances in categories? If 90% of your data is on non-smokers, then there is little to gain by analyzing a treatment affect against smoking status.
- Are your continuous variables greatly skewed? Or Zero inflated? This will mean standard analysis tests that require a more symmetric or normal distribution may not be valid

Additional Notes

Along with Chapter 4 of PMA6, we will use Chapter 2.3 of the Applied Stats notebook to examine the details of appropriate summary statistics and plot types to describe the distribution of a single variable.

Manage your expecations

There is no expectation for you to create publication quality graphs on your first try. Nor is it necessary for homework. Aim for "internal report" style graphs for your assignments.

There also is no expectation for "graph mastery" in this class. Homework 4 is your first try, it may look crude but that's okay. Poster prep stage I is a chance to revise and improve. Homework 5 is another chance to revise, so is poster prep stage II.

Always aim to improve, but don't expect perfection.

Back