Lecture 7: Exam 1 review

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

Topics

Will include:

- Toolkit

- EDA

- Describing distributions

- Comparing two variables

Will NOT include:

- Linear regression

Exam format

- **Format**: I will give you an Rmd file, a csv data file, and a codebook (README.md file). You will write your code and answer in the Rmd file, and submit the HTML file that is automatically generated on Canvas.

- **Questions**: The exam has four questions, most have multiple parts, and an extra credit question. Each is worth 10 points.

- **Materials**: The exam is "open book", which means you can see your assignments, the class materials, and look things up online. This simulates the environment one encounters when coding in the real world. In fact, if you do some data analysis later, you will have access to the materials from this class and the internet.

Some notes about Assignment 2

- If you're going to compare histograms on top of each other, make the limits of the x axis the same by using xlim=c(0,50) or whatever scale inside your hist code.

- If you change the color of your plots, try to make them color-blind friendly, or printer friendly. Basically try to make it as easy to see differences as possible.

- The age2 categories made it difficult to see whether the same was representative of the US population. If you can't tell whether an answer is "yes" or "no" from the data, you can say, I cannot tell with the information that is given. Honesty is probably the most important aspect of a careful data analysis.

Some notes about Assignment 2 (continued)

- If you drop missing values, please note that in your report.

- The variable age2 is actually balanced in terms of age! It's just really hard (maybe impossible) to tell from the way the categories are defined.

- The reader shouldn't have to go to a codebook to know what you mean (in the text or the plots).

- To make a barplot make sure you use a table as the argument *and* include a legend: e.g. `barplot(table(dat$gender), legend=TRUE)`.

- In general plots should have text that is legible, a title, labels on the axes, and labels on categories.