Lecture 1 Handout: Data and EDA

Statistics 139 Team

September 07, 2023

Topics

- Concept Checks: Stat 110/111 Review
- Handling data in R
- Numerical summaries: mean, SD, median, IQR
- Graphical summaries: barplots, boxplots, histograms, scatterplots

The material in this lab corresponds to the Lecture 1 Notes.

Concept Checks (Stat 110/111 Review):

a) What is the distinction between $\bar{X}$, $\bar{x}$, and μ?

b) What is the sampling distribution of $\bar{X}$? When is this exact? When is this an approximation?

c) What is the sampling distribution of the sample variance, S^2? When is this exact? When is this an approximation?

The General Social Survey

The General Social Survey (GSS) is a biennial, nationally representative survey conducted by the National Opinion Research Center at the University of Chicago. The GSS is 'second only to the U.S. Census as the most cited social science dataset in the country.' Even though the data are collected via a complex sampling design, the data can reasonably be analyzed validly as if it were a simple random sample from the US population (we will get into handling other forms of samples more carefully via survey weights later in the course), see this blog by Andrew Gelman, a world-class statistician.

The following questions will be explored in this lab with the GSS 2018 data:

1. At what age do Americans no longer further their education?
2. How is marital status linked to education and low income status?
3. How does income relate to being a government vs. private sector employee?

The full GSS 2018 data are available in the data file 'gss18.csv'. For convenience, descriptions of the variables used in this lab exercise are included below. To view the complete list of study variables and their descriptions, access the GSS documentation code book by clicking here.

- `age`: age of respondent, in years. Respondents 89 years or older were recorded as 89 years of age.
- `educ`: years of education of respondent (beyond kindergarten). Respondents with 20 or more years of education were recorded as 20 years.
- `rincom16`: the respondent's income, categorized into many groups. See: https://gssdataexplorer.norc.org/variables/6168/vshow
- `hrs1`: number of hours respondent worked the previous week.
- `marital`: marital status, with categories 1 = married, 2 = widowed, 3 = divorced, 4 = separated, and 5 = never married.
- `wrkgovt`: does respondent work for the government? 1 = government job, 2 = private sector job.

Note: there are many, many more variables in the data set.

Question 1.

a) Using numerical and graphical summaries, describe the distribution of ages of the respondents.

b) Calculate the median and interquartile range of the variable `hrs1`. Write a sentence explaining the median in the context of these data.

c) Use the following code to draw a random sample of 500 participants from the entire dataset. Using the random sample, `gss18.samp`, visually investigate the relationship between age and education. Based on this visual, at what age do respondents appear to no longer further their education? Use this smaller sample only for this part of the problem.

```r
# draw a random sample
set.seed(139)
row.num = sample(1:nrow(gss18), 500, replace = FALSE)
gss18.samp = gss18[row.num, ]

# create a visual
```

d) Compare the distribution of `educ` across each group in `marital` among adults (defined as individuals 25 years of age or older). Describe any trends or interesting observations.

Question 2.

a) Create a dummy/indicator/binary variable `lowincome` to indicate those individuals that make less than $15,000. Construct a two-way table, with `marital` as the row variable and `lowincome` as the column variable. Which group is at lowest risk of being low income? Highest risk?

b) Relative risk can measure how two categorical variables are related (really, two binary variables). Here, we are interested in measuring the relative risk as the ratio of: the proportion of respondents who are low income among those who are divorced to the proportion of respondents who are low income among those who are married. Calculate this relative risk for these respondents. From these calculations, is it possible to conclude that getting divorced reduces or raises one's chance of being low income?

Question 3.

a) Describe the distribution of income of the respondents. Estimate the median income, and provide a rough estimate for both the mean income and standard deviation of incomes.

b) The following code creates a new variable, `estimated_income`, within `gss18` that records the rough median of each income group from the variable `rincom16` for each respondent (in thousands of dollars). Use this variable to substantiate your 3 estimates in the previous part. Which of the 3 estimates will be biased?

```
medians = c(0.5, 2, 3.5, 4.5, 5.5, 6.5, 7.5, 9, 11.25, 13.75, 16.25, 18.75, 21.25,
            23.75, 27.5, 32.5, 37.5, 45, 55, 67.5, 82.5, 100, 120, 140, 160, 200)
gss18$estimated_income = medians[gss18$rincom16]
```

c) Which of the 3 estimates will be biased in the previous part? Will it be over or under estimated? Why?

d) Propose a better way to create the estimated income for each respondent that will end up with a less biased estimate than in part (b)? You do not need to implement this.

e) Compare the distribution of `estimated_income` for government employees vs. private industry employees. Describe what you see in a few sentences.