Agenda

1. WiDS Conference March 5th
2. Homework 2 is due on Wed in class
3. Center, Shape, and Spread

Recap of Lab Session

- Lab is due Monday just before midnight
- If there is a question, make sure it is answered in the text
- Procedure for submitting HTML files to Moodle (see Resources tab)
- R study guide
- Adding a newline after the R chunk helps prevent text next to figure.
- Polleverywhere.com

Warmup: Experiments and Confounding Variables

1. A study showed that women who work in the production of computer chips have abnormally high numbers of miscarriages. The union claimed that exposure to chemicals used in production caused the miscarriages. Another possible explanation is that these workers spend most of their work time standing up. Illustrate these relationships in a diagram.

2. A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Use a diagram to explain the association.

3. Students sign up to be subjects in a psychology experiment. When they arrive, they are told that interviews are running late and are taken to a waiting room. The experimenters then stage a theft of a valuable object left in the waiting room. Some subjects are alone with the thief, and others are in pairs – these are the treatments being compared. Will the subject report the theft? Afterwards, a consent form is given and the true nature of the experiment is explained to them. Do you think this study is ethically OK?

4. For each of the following pairs of variables, a statistically signficant positive relationship has been observed. Identify a potential lurking variable that might cause the spurious correlation.

   (a) The amount of ice cream sold in New England and the number of deaths by drowning

   (b) The salary of U.S. ministers and the price of vodka

   (c) The number of doctors in a region and the number of crimes committed in that region

   (d) The amount of coffee consumed and the prevalence of lung cancer

Thinking about Distributions Shape, Center, and Spread

- Graphical techniques for summarizing the *shape* of the distribution of one numerical variable:
  - Histogram [`geom = "histogram"`]
  - Density plot [`geom = "density"`]
  - Box (and whisker) plot [`geom = "boxplot"`]
- Statistics for summarizing the *center* and *spread* of the distribution of one numerical variable:
  - Center: mean [`mean()`], median [`median()`]
  - Spread: standard deviation [`sd()`], variance [`var()`], range [`range()`], IQR [`IQR()`]

Thought Experiment Consider the following two variables:

- The `height` of all adults in the United States
- The annual `income` of all working adults in the United States

Think about the distribution of each variable, and discuss the following questions with a neighbor.

1. Sketch a density plot for the distribution. What features does it have? Is it symmetric? It is unimodal?

2. Label the axes on your density plot. What is the range of each variable?

3. How would you summarize each distribution numerically? Which measures are most appropriate?

4. Suppose that the government issued a tax rebate in the amount of $2000 to each American taxpayer. How would the distribution of `income` change? What would happen to your measures of center and spread?

College Tuition The data set shows the tuitions and fees charged by the 50 colleges in Massachusetts from 2016-2017.

```
library(mosaic)
library(rvest)
library(readr)

url <- "http://www.collegecalc.org/colleges/new-england/"

Tuition <- read_html(url) %>%
  html_nodes("table") %>%
  html_table(fill=TRUE)

Tuition <- Tuition[[1]] %>%
  mutate(tuition = parse_number(Tuition)) %>%
  select(-Tuition) %>%
  arrange(desc(tuition))

head(Tuition, 7)

##                School                       Location tuition
## 1    Tufts University       Medford, Massachusetts   51304
## 2      Boston College Chestnut Hill, Massachusetts   50480
## 3    Brown University     Providence, Rhode Island   50224
## 4   Dartmouth College        Hanover, New Hampshire   49998
## 5 Brandeis University       Waltham, Massachusetts   49586
## 6     Yale University       New Haven, Connecticut   49480
## 7   Boston University        Boston, Massachusetts   49176
```

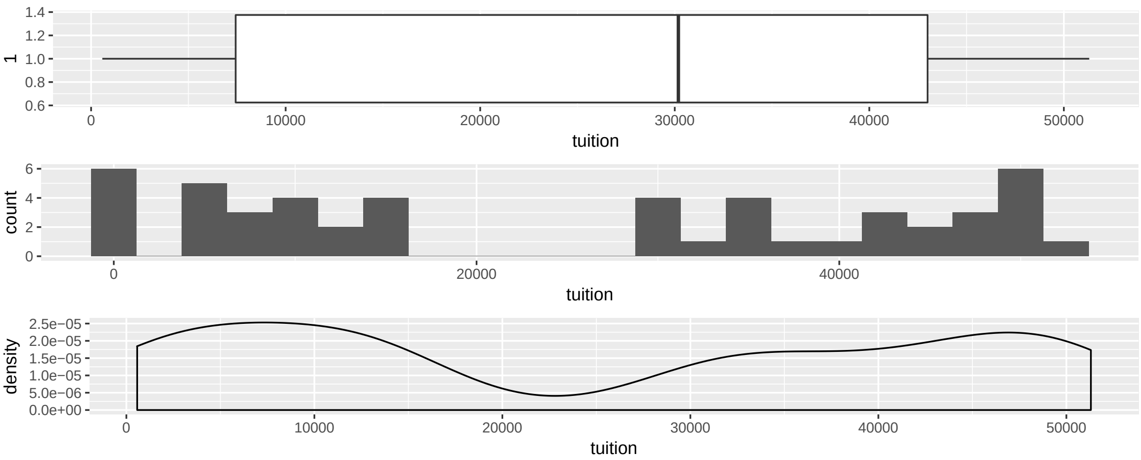
Next, we can calculate some measures of center and spread for tuition.

```
favstats(~ tuition, data = Tuition)

##  min   Q1 median    Q3   max     mean       sd  n missing
##  576 7431  30186 42996 51304 25217.86 18344.91 50       0
```

A box plot, histogram, and density plot reveal different features of the distribution.

```
library(gridExtra)
grid.arrange(
  qplot(data = Tuition, y = tuition, geom = "boxplot", x = 1) + coord_flip(),
  qplot(data = Tuition, x = tuition, geom = "histogram", binwidth = 2500),
  qplot(data = Tuition, x = tuition, geom = "density", adjust = 0.6))
```



1. What information can you glean from the histogram or density plot that is not revealed by the numerical table or the box plot?

2. What do you know about college tuition that might explain the features of this distribution?