**Agenda**

1. Center, Shape, and Spread

**Warmup: Lurking Variables** 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.

1. The amount of ice cream sold in New England and the number of deaths by drowning

2. The salary of U.S. ministers and the price of vodka

3. The number of doctors in a region and the number of crimes committed in that region

4. The number of storks sighted and the population of Oldenburg, Germany, over a six-year period

5. 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 variable:
  - Histogram [`geom_histogram()`]
  - Density plot [`geom_density()`]
  - Box (and whisker) plot [`geom_boxplot()`]
- Numerical Techniques for summarizing the *center* and *spread* of the distribution of one 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? Is it normal? 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?

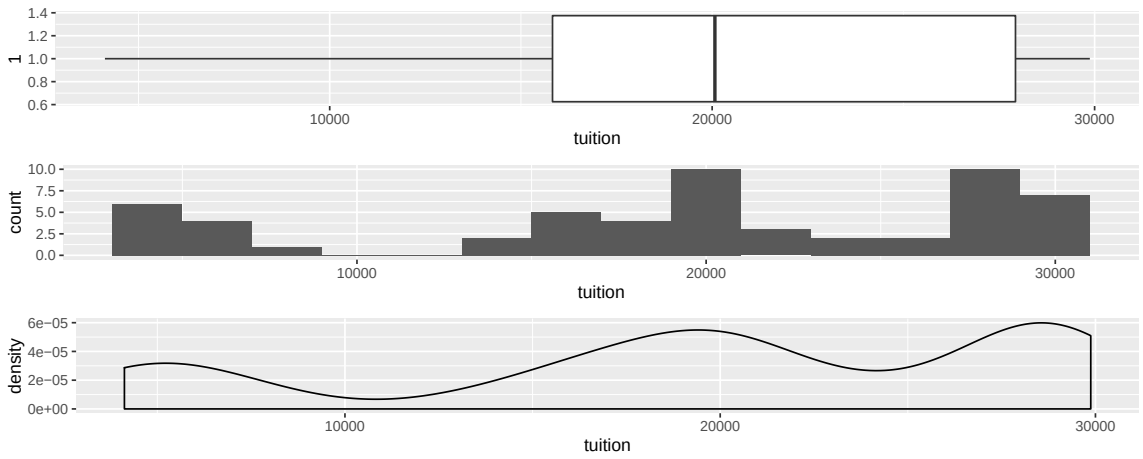
**College Tuition** The data set shows the tuitions and fees charged by the 56 four-year colleges in Massachusetts in the late 1990's.

```
require(mosaic)
Tuition <- read.csv("http://www.math.smith.edu/ips6eR/ch01/ex01_061.csv")
favstats(~ tuition, data = Tuition)

##   min      Q1 median       Q3   max     mean       sd  n missing
##  4123 15825.5  20072 27930.75 29875 19544.96 8476.124 56       0
```

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

```
gridExtra::grid.arrange(
  qplot(data = Tuition, y = tuition, geom = "boxplot", x = 1) + coord_flip(),
  qplot(data = Tuition, x = tuition, geom = "histogram", binwidth = 2000),
  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?