Day 1 (Monday)


- Read Sections 2.4–2.5 in MRH4.

- **Problem ⋆7**: Load the builtin R dataset chickwts. (Use data(chickwts).)
  
  (a) How many individuals are in this dataset?
  
  (b) How many variables are in this dataset?
  
  (c) Classify the variables as quantitative or categorical.

- **Problem ⋆8**: Describe some situations where the mean or median is clearly a better measure of central tendency than the other.

- **Problem ⋆9**: A bowler normally bowls a series of three games. A child wanting to compute the average score over three games might misunderstand the concept of an average, first taking the average of Games 1 and 2, and then averaging that with Game 3. (That is, if \( \bar{x}_2 \) denotes the mean of the first two games and \( \bar{x}_3 \) denotes the mean of the three games, the child believes that \( \bar{x}_3 = (\bar{x}_2 + x_3)/2 \).)
  
  (a) Give a counterexample to this method of computing the average of three games.
  
  (b) Given \( \bar{x}_2 \) and \( x_3 \), how should \( \bar{x}_3 \) be computed?
  
  (c) Generalizing, given the mean \( \bar{x}_n \) of \( n \) observations and an additional observation \( x_{n+1} \), how should the mean \( \bar{x}_{n+1} \) of the \( n + 1 \) observations be computed?

- **Problem ⋆10**: Let \( SS(c) = \sum (x_i - c)^2 \). (SS stands for sum of squares.) Show that the smallest value of \( SS(c) \) occurs when \( c = \bar{x} \). This shows that the mean is a minimizer of SS. (Hint: use calculus.)

- **Problem ⋆11**: Sketch a boxplot of a distribution that is positively skewed.

- **Problem ⋆12**: Suppose that \( x_1, \ldots, x_n \) are the values of some variable and a new variable \( y \) is defined by adding a constant \( c \) to each \( x_i \). In other words, \( y_i = x_i + c \) for all \( i \).
(a) How does $\bar{y}$ compare to $\bar{x}$?
(b) How does $\text{Var}(y)$ compare to $\text{Var}(x)$?

- **Problem ★13**: Suppose that $x_1, \ldots, x_n$ are the values of some variable and a new variable $y$ is defined by multiplying each $x_i$ by a constant $c$ ($y_i = cx_i$ for all $i$).

(a) How does $\bar{y}$ compare to $\bar{x}$?
(b) How does $\text{Var}(y)$ compare to $\text{Var}(x)$?

- **Problem ★14**: Suppose that $x_1, \ldots, x_n$ are given and we define a new variable $z$ by

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

What are the mean and standard deviation of the variable $z$? This transformed variable is called the **standardization** of $x$. In R, the expression $z = \text{scale}(x)$ produces the standardization. The standard value $z_i$ of $x_i$ is also sometimes called the **z-score** of $x_i$.

**Day 2 (Wednesday)**

- Read Section 2.6.
- **Section 2.6**: Do Exercise 48 (labeled 2-48).

- **Problem ★15**: The dataset `singer` comes with the `lattice` package. Make sure that you have loaded the `lattice` package and then load that dataset. The dataset contains the heights of 235 singers in the New York Choral Society.

(a) Using a histogram of the heights of the singers, describe the distribution of heights.
(b) Using side-by-side boxplots, describe how the heights of singers vary according to the part that they sing.

- **Problem ★16**: The R dataset `barley` has the yield in bushels/acre of barley for various varieties of barley planted in 1931 and 1932. There are three categorical variables in play: the variety of barley planted, the year of the experiment, and the site at which the experiment was done (the site Grand Rapids is Minnesota, not Michigan). By examining each of these variables one at a time, make some qualitative statements about the way each variable affected yield. (e.g., did the year in which the experiment was done affect yield?)
• **Problem ★17**: You have witnessed a classroom illustration of Simpson’s paradox. Construct your own illustration to conform to the following story:

> Two surgeons each perform the same kind of heart surgery. The result of the surgery could be classified as “successful” or “unsuccessful.” They have each done exactly 200 surgeries. Surgeon A has a greater rate of success than Surgeon B. Now the surgical patient’s case can be classified as either “severe” or “moderate.” It turns out that when operating on severe cases, Surgeon B has a greater rate of success than Surgeon A. And when operating on moderate cases, Surgeon B also has a greater rate of success than Surgeon A.

By the way, who would you want to be your surgeon?

• **Problem ★18**: Consider the set of natural numbers $$P = \{1, 2, \ldots, 30\}$$ to be a population.

(a) How many prime numbers are there in the population?

(b) If a sample of size 10 is representative of the population, how many prime numbers would we expect to be in the sample? How many even numbers would we expect to be in the sample?

(c) Using R, choose 5 different samples of size 10 from the population $$P$$. Record how many prime numbers and how many even numbers are in each sample. Make any comments about the results that strike you as relevant.
**Useful R commands:**

<table>
<thead>
<tr>
<th>Commands</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>data.frame()</td>
<td>Create a data.frame</td>
</tr>
<tr>
<td>edit()</td>
<td>Edit the entries of a data.frame</td>
</tr>
<tr>
<td>as.numeric()</td>
<td>Coerces entries in a vector to be of numeric type</td>
</tr>
<tr>
<td>plot()</td>
<td>General 2-variable plotter (may be used for timeplots, scatterplots, etc.)</td>
</tr>
<tr>
<td>xyplot()</td>
<td>Lattice command for plotting 2 quantitative variables</td>
</tr>
<tr>
<td>boxplot()</td>
<td>Produces box plots</td>
</tr>
<tr>
<td>bwplot()</td>
<td>Lattice command for producing boxplots</td>
</tr>
<tr>
<td>sample()</td>
<td>Sample from a vector</td>
</tr>
<tr>
<td>replicate()</td>
<td>Repeat an operation a specified number of times</td>
</tr>
<tr>
<td>attach()</td>
<td>Copy columns from a data.frame into vectors</td>
</tr>
<tr>
<td>detach()</td>
<td>Destroy vectors made using attach()</td>
</tr>
<tr>
<td>rm()</td>
<td>Remove items from memory</td>
</tr>
<tr>
<td>cor()</td>
<td>Calculate the correlation</td>
</tr>
<tr>
<td>var()</td>
<td>Calculate the variance</td>
</tr>
<tr>
<td>sd()</td>
<td>Calculate the standard deviation</td>
</tr>
</tbody>
</table>

Examples (without output):

```r
> ss = read.table('http://www.calvin.edu/~scofield/data/tab/ssurv.txt',
                  sep='\t', header=T)
> names(ss)
> require(lattice)
> histogram(ss$cds[ss$cds < 251], breaks=seq(0,260,10), type='c',
           xlab='No. of CDs owned')
> histogram(~ cds | gender, data=ss, type='c', layout=c(1,2))

> table(ss$oncampus, ss$gender)
> xtabs(~ oncampus + gender, data=ss)

> boxplot(ss$hourssleep)
> boxplot(hourssleep ~ gender, data=ss)
> bwplot(hourssleep ~ gender, data=ss)

> wine = read.csv('http://www.calvin.edu/~scofield/data/comma/heartDiseaseAndWine.csv')
> xyplot(hddeaths ~ winealc, data=wine, type=c('p','r'))
```