All page, section, and exercise numbers below refer to the course text (\textit{OpenIntro Statistics}, 3rd edition, by Diez, Barr, and Cetinkaya-Rundel.).

**Reading**: Section 1.2, 1.6  
**Problems for Self-Study**: (Do Not Turn In)  
- Exercise 1.5, 1.39, 1.45, 1.47, 1.51, 1.53, 1.57  
- Answers can be found at the end of the book (p.405-407).  
- Self-study problems are as important as Turn-In problems. We don’t require submission because we think you can learn from those problems by doing them yourself and checking the answers, without grading feedbacks. If having questions about those problems, you are welcome to ask the instructor or TAs.

**Problems to Turn In**: due 10 am, Friday, October 6, on Canvas.

1. Exercise 1.8(c) on p. 57  
2. Exercise 1.44 on p. 66  
3. The data values below represent the prices per share of the 25 most actively traded stocks on the New York Stock Exchange (rounded to the nearest dollar) on September 27, 2017.

\[25, 24, 10, 53, 2, 4, 5, 12, 6, 15, 39, 14, 12, 7, 171, 10, 72, 5, 14, 4, 54, 10, 6, 14, 26.\]

Make a boxplot of the distribution of the prices for these 25 stocks. You can make the box plot in R with the following commands:

\[\text{x = c(25,24,10,53,2,4,5,12,6,15,39,14,12,7,171,10,72,5,14,4,54,10,6,14,26)}\]  
\[\text{bwplot(~x, xlab="Price Per Share ($)"')}\]

Please indicate what values the boundaries and the middle line of the box represent respectively. Please also show what values the two whiskers extend to, and the value(s) of the outlier(s) if any. Explain how these values are computed. Hint: The following R commands might be useful.

\[\text{summary(x)}\]  
\[\text{sort(x)}\]


4. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Explain your reasoning.

(a) Housing prices in a country where 25\% of the houses cost below $350,000, 50\% of the houses cost below $450,000, 75\% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

(b) Number of alcoholic drinks consumed by college students in a given week. Assume that more than half of these students don’t drink since they are under 21 years old, and only a few drink excessively.
(c) The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15.

5. The data file `quakedepth.txt` contains the focal depth in kilometers of 2178 earthquakes with body wave magnitude at least 5.8 on the Richter scale occurring between January 1964 and February 1986. Make a histogram of the focal depth of the earthquakes. Please try a number of different binwidths and identify the location of modes of the distribution. The modes of the distribution are the depths that earthquakes occur more often than elsewhere.

```r
quakes = read.table("quakedepth.txt", header=T)
library(mosaic)
histogram(~Depth, data=quakes, width=??)
```

As the histogram has a long right tail, which makes it hard to identify the modes on the left side of the histogram, we can take only those shallow quakes that occurred no more than 100 km deep and then make a histogram for those shallow quakes only.

```r
histogram(~Depth, data=subset(quakes, Depth<=100), width=??)
```

6. It has been hypothesized that allergies result from a lack of early childhood exposure to antigens. If this hypothesis were true, then we would expect allergies to be more common in very hygienic households with low levels of bacteria and other infectious agents. To test this theory, researchers at the University of Colorado sampled the houses of 61 children 9-24 months old and recorded two variables: (1) whether the child tested positive for allergies and (2) the concentration of bacterial endotoxin in the house dust (endotoxin units per ml, EU/ml)\(^1\). The data file is `allergy.txt`.

(a) Load the data set to R. Find the five-number summary of the bacterial endotoxin concentration for each of the two groups and make a side-by-side boxplot.

```r
allergy = read.table("allergy.txt", header=T)
library(mosaic)
favstats(Endotoxin ~ Allergic, data=allergy)
bwplot(Endotoxin ~ Allergic, data=allergy)
```

Compare the center and spread of the distributions of the endotoxin levels between the “sensitive” and “normal” groups.

(b) Find the five-number summary and make a side-by-side boxplot comparing the log bacterial endotoxin concentration of the two groups.

```r
favstats(log(Endotoxin) ~ Allergic, data=allergy)
bwplot(log(Endotoxin) ~ Allergic, data=allergy)
```

Compare the center and spread of the distributions of the logarithm of endotoxin levels between the two groups.

**Remark:** When two distributions have unequal spread, one cannot compare them simply by comparing their centers. After log transformation, as the two groups have about equal spread, we can

simply compare the centers of the two distribution. As the median of the log bacterial endotoxin concentration of the two group differ about 6.91 \(- 5.97 = 0.94\), we can say, the median bacterial endotoxin concentration in the “normal” group is about \(e^{0.94} \approx 2.56\) times higher than that in the “sensitive” group since \(\log(X) - \log(Y) = d\) means \(X/Y = e^d\).

7. Why do some mammals have larger brains for their size? The table below are five rows of the data of the average values of brain weight, body weight, gestation lengths (length of pregnancy), and litter size for 96 species of mammals. The data file is `mammals.txt`.

<table>
<thead>
<tr>
<th>Species</th>
<th>Brain Weight (g)</th>
<th>Body Weight (kg)</th>
<th>Gestation Length (days)</th>
<th>Litter Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aardvark</td>
<td>9.60</td>
<td>2.20</td>
<td>31</td>
<td>5.0</td>
</tr>
<tr>
<td>Acouchis</td>
<td>9.90</td>
<td>0.78</td>
<td>98</td>
<td>1.2</td>
</tr>
<tr>
<td>African elephant</td>
<td>4480.00</td>
<td>2800.00</td>
<td>655</td>
<td>1.0</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>360.00</td>
<td>45.00</td>
<td>230</td>
<td>1.0</td>
</tr>
<tr>
<td>Human being</td>
<td>1300.00</td>
<td>65.00</td>
<td>270</td>
<td>1.0</td>
</tr>
</tbody>
</table>

You can refer to Lab #2: [http://www.stat.uchicago.edu/~yibi/s220/labs/lab02.html](http://www.stat.uchicago.edu/~yibi/s220/labs/lab02.html) to find the R commands you need to make the plots required below.

Please label all the plots properly.

(a) Make a scatterplot with body weight on the horizontal axis and brain weight on the vertical axis. Can you observe anything from the scatterplot about the relationship between brain weight versus body weight of mammals?

```r
mammals = read.table("mammals.txt", header=T)
library(mosaic)
qplot(Body, Brain, data=mammals)
qplot(Body, Brain, data=mammals, xlab="Body Weight (kg)", ylab="Brain Weight (g)")
```

The last command add labels to the scatterplot.

(b) As mammals in the data set differ in size by several orders of magnitude (from as tiny as a deer mouse which weighs 0.017 kg to as gigantic as an African elephant which weighs 2800 kg), please redraw the scatterplot in (a) after brain weight and body weight are transformed to their logarithms. What does the new scatterplot reveal about relationship between brain weight versus body weight of mammals?

```r
qplot(log(Body), log(Brain), data=mammals)
```

---

(c) Is brain weight associated the length of pregnancy? Make a scatterplot with gestation length on the horizontal axis and brain weight on the vertical axis. Describe the relationship between brain size and gestation length based on the scatterplot. Try transforming one or both variables and make a new scatterplot. Does the new scatterplot reveals a clear or simpler relationship?

(d) Make scatterplots to explore the relation between body weight and gestation length. You might want to transform one or both variables.

(e) One possible explanation for the positive linear association between (the log of) brain weight and (the log of) gestation length observed in part (c) is that they are both positive linearly associated with (the log of) body weight. Does gestation length have its own effect on brain weight, even after accounting for body weight? In other words, for mammals with similar body weights, do those with longer gestation period tend to have larger brains? To investigate this, we can inspect the relationship between the three variable simultaneously using a coded scatterplot in which the third variable is represented by the size of points.

```r
qplot(log(Body), log(Brain), size=Gestation, data=mammals)
```

When using the size of points to represent a variable, note that sometimes large points may mask other smaller points. So it is better to use hollow points rather than solid points.

```r
qplot(log(Body), log(Brain), size=Gestation, shape=I(1), data=mammals)
```

From the coded scatterplot, for mammals with similar body weights, do those with longer gestation period tend to have larger brains?

(f) Make a coded scatterplot similar to the one in part (e) to investigate whether litter size is associated with brain weight, after accounting for body weight. What do you conclude? In other words, for mammals with similar body weights, do those with smaller litter size tend to have larger brains?