STAT22000, Spring 2015 Homework 1

due Wednesday April 8

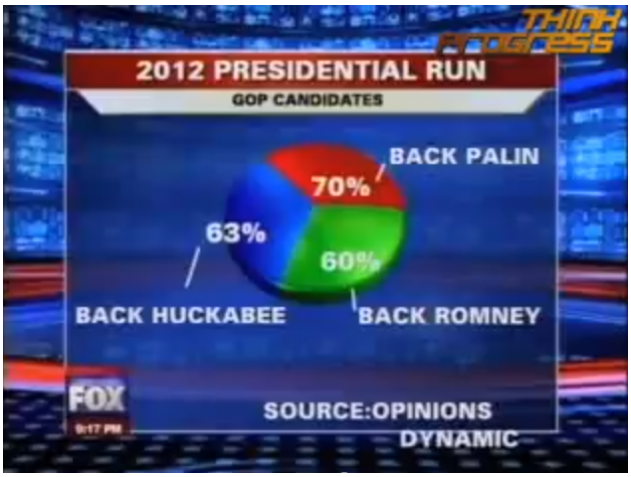
All page, section, and exercise numbers below are for the course text (Moore, McCabe and Craig, Introduction to the Practice of Statistics, 8th edition).

Reading: Section 1.1-1.4
Problems for Self-Study: (Do Not Turn In. Solutions to odd-numbered exercises are at the end of the textbook.)

1. Exercises 1.29, 1.35, 1.61, 1.63, 1.77, 1.93, 1.147, 1.171

Problems to Turn In:

1. The following pie chart comes from Fox News (Chicago) on November 23rd, 2009. (Here is the video `http://www.youtube.com/watch?v=-rbyhj8uTT8`)



   (a) What is wrong with the pie?

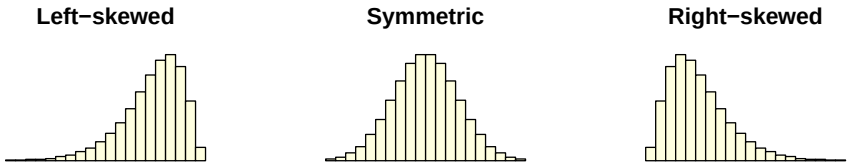
   (b) Assume the numbers on the pie chart are correct. Make a correct graph to display the numbers.

2. This website shows an interactive visualization of the past 200 years of immigration into the US:

   `http://insightfulinteraction.com/immigration200years.html`

   Take a few minutes to try it out. Hover your cursor over different countries of origin to see trends over time. Describe two things you could add or change to improve this data visualizer (to make it more clear, more informative, etc).

3. Suppose everyone in your class fills out a survey answering each of the following questions. For each question, do you expect the resulting distribution to be left-skewed, right-skewed, or symmetric? Explain why. (Some questions might have more than one acceptable answer.)



(a) How long is your first name (how many letters long)?

(b) How many classes have you taken at U of C?

(c) How many biology classes have you taken at U of C?

(d) How many minutes of the first class did you attend?

For the next problem, you will need to load data into R.
Instructions for how to do this are at the bottom of this HW.

4. New York City public schools have posted data on the performance of each school on a standardized math test given to students in grades 3–8.[1] We've reformatted the data and will look only at results from Manhattan schools in 2010. Download the file `NYC_math_test_data.txt` from Chalk. Use the R computing software to answer the following questions about the data.

(a) Find the mean, and the five-number summary of the total number of students tested (in each grade within each school) using the command

```
> summary(NumTested)
```

(b) Do the numerical summaries in part (a) suggest an approximately symmetric distribution or a skewed distribution for the number of students tested? Explain your conclusions.

(c) Explore the distribution of the numbers of students tested by creating 3 histograms of the values, each with a different bin widths

```
> hist(NumTested)
> hist(NumTested, breaks = 10)
```

etc. You can try different values of `breaks` to get different numbers of bins. You can also specify the range of the histogram and the size of class intervals. For example, the command below creates a histogram covering the range 0 to 4000 and the size of class intervals is 250.

```
> hist(NumTested, breaks = seq(0,500,by=10))
```

If you want to shade in the histogram, add a color using the `col` command:

```
> hist(NumTested, breaks = seq(0,500,by=10),col="gray")
```

Be sure to include axes labels (including units) and titles for your histograms. Check the R help file to see how the change the axes labels and titles using the command.

[1] The original data is at:
`https://data.cityofnewyork.us/Education/NYS-Math-Test-Results-By-Grade-2006-2011-School-Le/jufi-gzgp`

```
> ?hist
```

Use these plots to comment on the shape of the distribution and any values that you consider to be potential outliers. Also hand in a printout of one of the histograms with axes labels (including units) and a title.

(d) Calculate the proportion of students (within each grade and each class) whose test results are Level 1:

```
> ProportionLevel1 = Level1/NumTested
```

Are there any potential outliers in these values according to the 1.5 × IQR rule?

5. The 10 tallest skyscrapers in the world have the following heights, in meters:

`828,632,601,541.3,509,492,484,452,452,450`

These values have mean 544.13 meters and SD 117.83 meters.

(a) Chicago's Willis tower is the 11th tallest, with a height of 442 meters. Suppose that for each of the 10 tallest buildings, we record how much taller it is, than the Willis tower. What would be the mean and SD of this list? (Do this without calculating the 10 new numbers.)

(b) The Kingdom Tower, currently under construction in Saudi Arabia, has a planned height of 1000 meters. For each of the current 10 tallest buildings, we record how much shorter it is than the Kingdom Tower. What would be the mean and SD of this list? (Do this without calculating the 10 new numbers.)

(c) Suppose that for each of the 10 tallest buildings, we record how much taller it is, than the Willis tower, <u>in feet</u> (1 meter = 3.28 feet). What would be the mean and SD of this list? (Do this without calculating the 10 new numbers.)

6. 1.144 in the textbook (in edition 7 it's problem 1.142), parts (a) and (c) only.

7. Temperature in Chicago:

(a) The following temperatures were observed in Chicago last June, in degrees Fahrenheit:
`66 72 80 81 61 57`
Show how to calculate the mean and SD of these observations.

(b) Now suppose that the temperature in Chicago in June comes from a N(70°,9°) distribution. Calculate the z-scores of the first three observations 66°, 72°, 80°.

(c) What percent of the time will the temperature be between 60° and 65° in Chicago in June?

(d) In Chicago in June, 20% of the time, the temperature will be above _____.

(e) In Chicago in June, 20% of the time, the temperature will be between _____ and 72°.

Instructions for using R for the `NYC_math_test_data` data set:

Downloading R: Go to `http://www.r-project.org` to download & install R.
Opening R: there will usually be an icon for R on your desktop / in your dashboard.

Once you see this symbol `>`, you can start typing commands.
See `http://www.stat.uchicago.edu/~yibi/R/Rtutorial.html` for some examples to get familiar with R.

To work on the HW problem, you'll need to load the `NYC_math_test_data` data into R. First, download the file `NYC_math_test_data.txt` to your computer from Chalk. (If clicking on the file just shows you the data instead of downloading the file, try a right-click (Windows) or ctrl-click (Mac) to access the option of saving the file.)
Next, in R, change the working directory to whichever directory contains the file. To do this,

- In Windows: go to the File menu, select Change Working Directory, and select the appropriate folder/directory.
- In Macs: go to the Misc menu, select Change Working Directory, and select the appropriate folder/directory).

Now, when you ask R to read in the data, it looks inside the correct directory. Read the data using the command:

```
> NYC_math_test_data = read.table("NYC_math_test_data.txt", header=TRUE)
```

Here, `NYC_math_test_data` is a data frame containing several variables: `School`, `NumTested`, `Level1`, etc. For example try

```
> mean(NYC_math_test_data$NumTested)
```

This computes the mean of the `NumTested` variable in the `NYC_math_test_data` data frame. For a shortcut, if you want to be able to type just `NumTested` instead of `NYC_math_test_data$NumTested`, use the command

```
> attach(NYC_math_test_data)
```

Now, every time you type `NumTested`, R knows to look for this variable inside of the `NYC_math_test_data` data frame, so you can type things like

```
> mean(NumTested)
```

When you're done using the data, to undo the `attach` command,

```
> detach(NYC_math_test_data)
```