**Name:**

**Classwork 4**

**Exploring Variation in Heads (Lecture)**

Today we’re going to look at our Head Sketch Survey data in the DataCamp Sandbox: <http://bit.ly/datacampsandbox>

1. First, we need to put this google spreadsheet into the DataCamp Sandbox. Then we can use our fancy R coding skills on this data. Make some notes here so that you will be able to do this yourself in lab.

General code for importing csv file (more instructions on Page 2.6):

DataFrameName <- read.csv("https://url.com", header=TRUE)

1. We’ve got some choices here. What variable are you interested in exploring? What research question(s) do you have about that variable?
2. As a class, let’s explore variation in self-measured head height (head length?).
3. Predict what that histogram might look like. Roughly sketch it below.

(b) Now let’s actually make the histogram in R. Write the code and roughly sketch it below. Make some observations (think: shape, center, spread, weird things).

1. Let’s get the 5 number summary and label your histogram sketch above. Why aren’t the quartiles equally sized?
2. The weird things that we see – should we filter the data in some way? Why or why not?
3. Now let’s make a histogram of partner-measured head height (write code and sketch it here). How is it similar to the self-measured head heights? How is it different?
4. If a different class of college students measured their head heights, would histogram of their data be similar? How would it be similar? How would it be different?
5. If a different class of 1st grade students measured their head heights, would the histogram of their data be similar? How would it be similar? How would it be different?

**Explaining Variation**

We start with variation in general. Part of our job as statisticians will be to “explain” the variation.

1. We talked about this a bit in the previous class, what does it mean to “explain variation”?
2. What are some variables that you might want to collect data on in the future if your goal was to “explain variation” in head height?

We won’t be able to explain all of the variation, so some of it will be unexplained variation, sometimes called “random variation,” or “error.”

1. The **random variation** will include a few different kinds of variation such as the types listed below.
   1. Not-yet-explained Variation:
   2. Measurement error:
   3. Sampling variation (think back to the many samples we drew for the Martin v. Westvaco simulation...)

**Different Types of Variables (Lab)**

1. Take a moment to import the csv file for the Head Sketch data into DataCamp Sandbox.
2. Which of these variables are categorical and which are quantitative? Which variables can we make a histogram for? Which ones should we make a bar chart for?
3. Should we tell R that some of these variables are categorical? Which ones? How would we do that?
4. Create a new variable for whether someone was correct or incorrect about where their eyes would fall relative to the midline. Write the code here.
5. How many people were correct about where their eyes would be? What percentage of people in our data were correct?
6. If we knew about the location of their eyes (that is, whether someone’s eyes were above, at, or below midline), do you think we could make a better prediction about their head height? Why or why not?
7. How would we explore our data to see if the location of the eyes helps us make a better prediction about their head size?