Homework #3: Z-scores, the normal distribution, and distributions of the sample mean.
Due: In lab, Wednesday, Feb 13

DIRECTIONS: For this homework, please turn in hard copies in class on Wed, Feb 13. This homework counts double. Write answers out in a separate Microsoft Word document (note that some questions may not require an answer – so you can skip these or just say "completed" – but we'll try to put what the actual question to be answered, or what you need to do to get credit, by placing that in **bold** in the question). We encourage you to work in groups for R-related material, but make sure to do the work yourself too and *always* come up with independent answers. Homeworks that have identical answers will receive F's.

Question 1 [to do using R]: First, load the datasets that I have stored on an external server. To do that, use the command below:
load(url("http://www.matthewckeller.com/Stats3101/Stat3101.Datasets-2013.RData"))

For this question, we'll be working with the survey data you completed the first lab. To look at the first 6 rows of that data, type the following into your "Script Window" and run it:
head(survey2013)

We can also look at the dataset by making it the active dataset. Hit the button next to "Data set" and find "survey2013", then hit "OK". Finally, hit the button "View data set" to view it. Each row is a person in this class (I have randomly added or subtracted just a bit to height to make it impossible to identify anyone). The scroll bar at the bottom of the window lets you scroll to see additional variables to the right, whereas the scroll bar on the side lets you scroll to see additional people.

Let's concentrate on the variable *conscientiousness* – the next to last variable in the dataset (far right). Conscientiousness was measured on all students in this class attending the first lab. It was derived by taking the average of the 5 questions asked related to this construct (a common approach in psychometrics). It is defined as, "A tendency to show self discipline and aim for achievement; to do planned rather than spontaneous behavior." People with high scores are more like this and people with low scores less so.

A) First, let's take a look at the data. Create a histogram of the variable "conscientious" using the menu or scripting, as you did in the last homework (you're saving your homeworks, right? You'll want them for the lab final). **Paste the histogram into your homework.**

B) In words, describe the shape of the distributions of this variable.

C) Give the mean and standard deviation of this variable. In a single sentence, provide an interpretation of the standard deviation of this variable in language that grandma could understand.

D) Write down the general formula for a z-score. Below it, write down what every symbol means as applied to conscientiousness (e.g., " $\bar{X}$ is the mean of conscientiousness, etc.")

E) Use R to find the z-score of conscientiousness for each person in the dataset, and attach this new variable (called "z.consc") to your dataset, survey2013. To attach a new variable to your dataset, use the "$" like we did in the last homework:

```
survey2013$z.consc <-  (survey2013$conscientious -
mean(survey2013$conscientious))/sd(survey2013$conscientious)
```

[Note 1: I wouldn't copy R syntax into the Script Window because characters (such as minus signs -) can change (e.g., to dashes —) in PDF documents. Write it out instead.]

[Note 2: If you want to use the menu to do additional statistics on survey2013 and the new variable you just created, z.consc, you will need to re-load survey2013 in "Data set". To do this, go to the button next to "Data set", which should read "survey2013", hit it, choose any random other dataset, and hit "OK". Then, go back to the button next to "Data set", find "survey2013", hit it, and hit "OK". Whew!! Personally, I find it a lot easier just to use the syntax directly! Maybe you do too?]

Create a histogram of "z.consc". What is its mean, standard deviation, and shape? What changed and what remained the same compared to your answers about the shape, mean, and standard deviation in B & C above?

F) We might expect that this personality dimension differs in some way between males and females. Do college-aged women tend to be more or less conscientious than college-aged males? Is there similar spread in the two genders? Let's use our sample (this class) to get insight into these questions. You can either use the menu system (by going to Statistics -> Summaries -> Numerical Summaries.. choose the variable "z.consc" in the "Variables" box, then make sure that you choose "summarize by groups" and choose "gender"). Or you can use the syntax below – it's up to you:

```
mean(survey2013$z.consc[survey2013$gender=='f'],na.rm=TRUE)
mean(survey2013$z.consc[survey2013$gender=='m'],na.rm=TRUE)

sd(survey2013$z.consc[survey2013$gender=='f'],na.rm=TRUE)
sd(survey2013$z.consc[survey2013$gender=='m'],na.rm=TRUE)
```

If you're into this kind of thing, you might take some time to see if you can figure out *how* the syntax above works! Just break it apart bit by bit.

Report the means and standard deviations for males vs. females on both "z.consc" (as above) and on "conscientious" (which you'll need to figure out how to find using R yourself)**.**

G) How does the mean of "conscientious" differ between males vs. females? Interpret this difference (note that the scale is somewhat arbitrary).

H) How does the mean of "z.consc" differ between males and females? Interpret this difference (note that the scale is no longer arbitrary). [Note for tie-in of future material: When we interpret mean differences in units of standard deviation, we call this "Cohen's d" – a common metric for knowing *how* far apart two means are].

I) Which mean difference (in G or in H) do you find easier to interpret? Why?

J) What is your opinion on whether the mean conscientiousness difference between males vs. females is "real" (reflective of a true population difference) or "just due to chance" (just a difference we happened to observe in this sample but probably wouldn't observe in other samples)? What factors might affect your opinion?

<u>Question 2</u> [to do by hand]: The following questions refer to these scores on a test given to my honors stats class: 65 89 92 94 70 75 83 82 90 78 73 88 94 92 44
Do the problems below by HAND, and show your work:

A) By hand, find the mean, median, 1st quartile, 3rd quartile, interquartile range, and standard deviation of the data above.

B) Find the z-scores for each of the scores above.

C) Draw a histogram of the original data. The x-axis should be the original scores.

D) Under the original scores on the x-axis, write the z-scores for each bin. In other words, transform the x-axis to a "standardized scale," and put this new standardized scale under the original.

E) Draw a boxplot of the data above

<u>Question 3</u> [to be done using R – please use only scripting language for this problem, as it is too difficult to be done using the Menu system. In other words, just type the commands below into "Script Window" and run them.] NOTE: It is important that you read and think about what is written in HW question #3 thoroughly, as I'm trying to teach you about sampling distributions not just through lecture, but also through what I write in this assignment. You will not understand the concepts I'm trying to convey just by racing through this question in order to put down all the correct answers! The goal for this question is to teach you the concept of a sampling distribution of means – so take some time and think about this as you do it.

For the rest of this homework, we're going to work with GPA in major, which is defined as the self-reported major GPA for people in this class. It is the variable "gpa.major" in survey2013.

We're going to use this variable to demonstrate the concept of a sampling distribution of means. We use sampling distributions throughout the rest of the class in order to perform inferential

statistics. Normally, sampling distributions of means are *conceptual* constructs, and we don't actually *see* them, rather, we just know three important things about them (their shape, their mean, and their standard deviation).

However, for this example, you're going to actually build a sampling distribution and *see* it (something we're doing for didactic purposes, but which is rarely done in conducting statistical tests). We'll then test whether the sampling distribution you built conforms to our theoretical expectations.

HYPOTHESIS: Each day, the people who sit in my front row of class are slightly different. My question is, do people who sit in the front row of class tend to have higher GPAs (perhaps because they tend to be more reponsible students)? Formally, we'd say that the ("alternative") hypothesis is that front row sitters tend to have higher GPAs.

THE NULL HYPOTHESIS: The people who sit in the front row of this class are a random sample of students in this class (i.e., a random sample from the 'population') with respect to their GPAs.

OUR POPULATION here is the entire class (i.e., we're not interested in generalizing to any other classes than this one).

OUR SAMPLE is the 9 people sitting on the front row of class on a given day.

A) First, let's look at the statistics of our population (the GPAs of everyone in the class). There are three missing values in here (NA's). So we first need to remove those. Let's create a new variable called "gpa" that has the GPAs of all people after removing the NA's:
gpa <- as.vector(na.omit(survey2013$gpa.major))

What is the population mean of gpa (i.e., the mean gpa in this class)? What is the standard deviation? Use this to find these:
mean(gpa)
sd(gpa)

B) Actually, because we're considering this class as a *population*, we shouldn't be using R's sd() function, which divides the sums of squared deviations by n-1 (the *df*). Instead, let's create a new function called "pop.sd" that finds the population standard deviation (dividing by n instead):

pop.sd <- function(x) {sqrt(sum((x-mean(x))^2)/length(x))}

Now report the *population* standard deviation of gpa:
pop.sd(gpa)

C) Create a histogram of gpa using the code below and paste it into your homework.
Describe the shape of this distribution. Is it normally distributed?
hist(gpa,breaks=10,col='pink')

D) Let's say that I ask the 9 people sitting in the front row what each of their GPAs is. These GPAs are: 2.6, 4.0, 2.8, 3.4, 3.8, 3.4, 3.6, 3.7, and 2.8. Create a variable in R called "gpa.frontrow" containing these 9 values using the c() function in R like so:
gpa.frontrow <- c(2.6, 4.0, 2.8, 3.4, 3.8, 3.4, 3.6, 3.7, 2.8)

What is the mean of gpa.frontrow? What is its standard deviation? What is its shape (from looking at a histogram)?

E) Our hypothesis is in terms of the *mean* of ten individuals: is the mean of the 9 people sitting in the front row *different* than what would be expected "by chance." In other words, is it different than what we'd expect if a random group of 9 people from this class sat in the front row?

As a first step to answering this, choose 9 random people in this class and figure their mean hours/week studying. To do this, use the function sample(x=?,size=?,replace=?), where x= is the argument for what we're sampling (here, the gpa variable), size= is the argument for how many to take, and replace= is the argument of whether to sample with replacement (don't worry about why sampling with replacement is relevant; it's a minor issue). So run this:
sample1 <- sample(x=gpa,size=9,replace=TRUE)

This is a *random* sample of individuals. If there is no relationship between sitting in the front row and GPA, then a group of 9 people's GPAs might look like "sample1". Look at your sample in R by typing sample1 and running it:
sample1

What is the mean of sample1? Is it above or below the mean of the actual sample of people who sit in the front row, gpa.frontrow?

F) Comparing the mean of just a single random sample of GPAs to the mean of gpa.frontrow is not a great way to test whether the mean of gpa.frontrow is an unusual mean – it's just a single mean from a single random sample.

Take a second random sample of 9 individuals each from this class:
sample2 <- sample(x=gpa,size=9,replace=TRUE)

What is the mean of sample2? Is it above or below the mean of the actual sample of people who sit in the front row, gpa.frontrow?

G) Find the means of three additional random samples, sample3, sample4, and sample5, and report these three means. Of the five means you have collected, how many are greater than the mean of gpa.frontrow? Given these five means, drawn randomly from the class, compared to the mean of gpa.frontrow, do you think that gpa.frontrow is a random selection of 9 students, or not?

H) Now create a vector, called "random.5sample.means" that contains the means of each of the five random samples you just created. E.g.:
random.5sample.means <- c(mean(sample1),mean(sample2),mean(sample3),mean(sample4),mean(sample5))

These five means are called a "sampling distribution of means." Rather than each score in this distribution being an individual GPA, each score in this distribution is itself a *mean* of 9 GPA scores! THIS IS A SAMPLING DISTRIBUTION OF MEANS (albeit a small one of only 5 means rather than a ~ infinite number we'll be dealing with shortly). THE SAMPLING DISTRIBUTION OF MEANS IS THE MOST IMPORTANT CONCEPT IN THIS CLASS!!

If the null hypothesis is true, and the GPA's of people in the front row is just a random collection of people [i.e., there is no association between sitting in the front row and GPA], then the mean of the 9 front row sitters should look kind of like one of these five random means, right?

How does the mean of gpa.frontrow compare to these five means from random samples? What does this tell you about how likely or unlikely it is that the 9 people in the front row are just a random sample of people in the class with respect to GPA?

I) Why stop at just five random samples? Let's take a look at 8,000! I could easily say 20,000 or 100,000 or a million. Indeed, in theoretical sampling distributions of means, which is what we'll be doing from here on out, it's theoretically an infinite number of sample means, but for time's sake, let's stick with just getting a representative sample of 8,000 means. Each of these means is what we might observe in a random collection of 9 GPAs in this class. Here is the code to create a vector, random.8000sample.means, of length 8000, each entry being a mean of a random sample of 9 individuals. You do NOT need to know this code for a test, but it's kind of cool understanding loops in R if you're so inclined…

```
random.8000sample.means <- vector(length=8000)

for (i in 1:8000){
        sample.this.iteration <- sample(x=gpa,size=9,replace=TRUE) #over-written each loop
        random.8000sample.means[i] <- mean(sample.this.iteration)}
```

The above may take a while on slow computers. When done, look at random.8000sample.means:

```
sort(random.8000sample.means)  # a lot of numbers here!!
```

random.8000sample.means is just like random.5sample.means, except that we have 8000 means taken from random samples of size 9 rather than just 5 means of random samples of size 9. Thurs, random.8000sample.means is a much more accurate sampling distribution of means.

What is the mean of random.8000sample.means? How does it compare to the population mean (i.e., the mean of gpa)?

J) What is the standard deviation of random.8000sample.means? How does it compare to the population standard deviation (i.e., the pop.sd of gpa)?

K) What does the sampling distribution of means look like? I.e., make a histogram of random.8000sample.means and paste it into your homework (make it color blue to differentiate it from the pink distribution of the original gpa scores). How does its shape compare to the original distribution of gpa? Is it more or less normally distributed?

This is the central limit theorem in action! Distributions of sample means tend to become normal as *n* increases. Here, *n* is 9. So if we took random samples of size 30 instead, this distribution would be even more normally distributed.

L) Theoretically, we expect the mean of the sampling distribution to equal the population mean, and we expect the standard deviation to be equal to the standard deviation in the population, divided by the square root of how big the sample is. The standard deviation of a sampling distribution of means has a special name, the "standard error of the mean", or SEM. The samples are of size *n*=9 here, so:

$$SEM = \frac{SD}{\sqrt{n}} = \frac{.518}{\sqrt{9}} = .173$$

How does the standard deviation of random.8000sample.means compare to the theoretically expected standard deviation (or "SEM") of a sampling distribution of means drawn from this population? It should be very close, within +/- .005 of .173 (IS THAT NOT FREAKING COOL?? NO. SERIOUSLY. IF YOU DON'T FIND THIS COOL, DROP MY CLASS.)

M) Where does the mean of the original sample gpa.frontrow fall on this sampling distribution? Just by eye-balling the histogram of random.8000sample.means and seeing where the mean of gpa.frontrow would fall on that distribution of means, **is it likely that the 9 people in the front row are a random collection of students with respect to GPA? Why or why not?**

N) Out of the 8000 samples of random means, what proportion of them are as high or higher than the mean of gpa.frontrow? You can do this by using sort(random.8000sample.means) and counting the number of times a mean of a random sample of 9 is over the mean of gpa.frontrow, and then dividing this number by 8000. But an easier way is to use this code:
sum(random.8000sample.means > mean(gpa.frontrow))/8000

Call this proportion a "p-value" (in fact, it's a "one-tailed" p-value, but we'll get to that later). **What is your p-value? Describe in words what this "p-value" is telling you.**

PS – welcome to inferential statistics!!