Homework #2: Central Tendency and Variability.
Due: In lab, week of Feb 4 – Feb 8

DIRECTIONS: Please turn in hard copies of homeworks in lab the week after assigned. 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.

Part 1: To be done in R:

Question 1: Open up R and then load the R commander library, just like you did in the last homework (make sure you're saving your homeworks, as each homework will assume knowledge from the last ones). Then load some datasets that I have stored on a server. To do that, use the command below:

```
load(url("http://www.matthewckeller.com/Stats3101/Stat3101.Datasets-2013.RData"))
```

This command should be place in the Script Window and then run by placing the cursor on it and hitting the "Submit" button (or by hitting "control – r"). It will take up to half a minute or so for you to load your data if many people are doing so at once. Once the datasets have loaded, you can see the names of each dataset by hitting the button next to "Data set", or you can run the command below in your Script Window:

```
ls()
```

A *dataset* is a very common thing that we use in analyzing data in science. Usually (but not always), each row represents a single unit that we're interested in (a person, a classroom, a city, a state, etc.), and each column represents that unit's score on a variable. For example, the **fcq** dataset has information has information on the Faculty Course Questionnaire for 1016 classes previously taught at CU Boulder. In this case, each unit is one class and is represented as one row. Information for each class is contained in each column. To look at the first 6 rows of the **fcq** dataset, run the command below in your Script Window:

```
head(fcq)
```

Let's load the **fcq** dataset so that it is the "active" dataset. Do this by hitting the button next to Data set:" and then by choosing "fcq". You can now look at the **fcq** dataset by hitting the "View data set" button. (Note that R-commander kindly shows you the correct syntax for running this command in the Script Window).

Let's look at some basic descriptive statistics of all the variables in **fcq**. Do this by going to "Statistics" Menu -> "Summaries" -> "Active data set" (yes, you do want to look at all the variables). Note that "NAs:" means "Not Applicable" and is R-speak for missing data. R will tell you how many "missing" datapoints there are for each variable. Data missingness is a part of analyzing real data. Data can be missing because subjects didn't answer a question, data was lost, etc.

You should be able to figure out what the variables mean by looking at the names of the variables and the values of the variables, but let's look at the variable (i.e., column) "**course**". This variable is the average course rating (from 0-4) from all the students in a particular course. **Write a brief description of this variable (course) – in words and backed up with numbers – including a description of its central tendency (quantified using two statistics), its spread (quantified using two statistics), and its shape (include a pasted histogram, and describe it).** You should be able to figure out how to get all these by playing around with the pull down menus in R-commander, but as a hint, go to "Statistics" -> "Summaries" -> "Numerical summaries" if you want to get a variable's mean and standard deviation; go to "Graphs" -> "Histogram" to make a histogram. I'll leave the rest up to you to figure out by hunting around the pull down menus/checkboxes.

Question 2: Let's say that we're interested in understanding class sizes at CU Boulder – i.e. the typical class size, how much differences there are in class sizes, and what the shape of the distribution of class sizes is. To investigate this, look at the **nGrade** variable in the same ways you looked at **course** above. This variable (**nGrade**) measures the official size of a random selection of classes during the time period (2001-2004) covered.

Note that if you want to look at just the variable **nGrade** in R, you can do that using this syntax:

```
fcq$nGrade
```

If you run the command above, you get to look at every class size in the dataset, in the order in which they appear in the dataset. The "$" after the name of the dataset says "grab the **nGrade** variable in the fcq dataset". So if we want the mean or standard deviation of nGrade, we'd do this in syntax:

```
mean(fcq$nGrade, na.rm=TRUE)  #finds the mean
sd(fcq$nGrade,na.rm=TRUE)   #finds the standard deviation
hist(fcq$nGrade)    #makes a histogram
```

Note the use of the argument "na.rm" in the function mean() and sd(). Functions in R often have arguments that allow us to shape how the function behaves. Here, we're telling the functions to ignore missing values (NA's). Of course, you can accomplish the same thing (using a different function with different arguments and defaults) by fishing around in the Menus in R commander, but often it is easier and more efficient to write the syntax directly. It is up to you whether you want to rely more on syntax in this class or more on the menu, although there are times when we can only accomplish what we want using syntax.

Write a paragraph describing what you found about class sizes at CU. Begin your paragraph with a sentence suggesting what the purpose of our study is. In your next sentence, write how we measured this variable (in this case, I'm looking for something like, "To study this, we looked at the enrollment size of a random selection of 1016 classes at the University of Colorado between 2001 and 2004"). In your next sentence(s), write what you found; usually, this can be done in a single sentence. When you write this sentence, put the statistics (e.g., mean; standard deviation) in

parentheses. End your paragraph with a no-number sentence stating some conclusion or overall summary.

Question 3: In this class, we will often ask you for a "four sentence summary of results." That's basically what you just did above. We will be doing a LOT of four sentence summaries in this class; the four sentence summary will be on every test you take and every homework you do. It is an important skill to master in scientific writing; the most important part of any scientific paper is the "abstract", which briefly summarizes the motivation, methods, findings, and conclusion of a scientific study. The four sentence summary is practice in writing an abstract. It mirrors the four sections of an APA formatted paper, and is made up of the following:

1st sentence: The Introduction. State the problem, or what you are interested in looking at.
2nd sentence: The Method. How did you go about solving this problem?
3rd sentence: The Results. What did you find?
4th sentence: The Discussion. What is your conclusion?

For example, let's say we're interested in whether female undergraduates check their email more frequently than male undergraduates. We could use our survey results to try to answer this in the following way:

Example 4-sentence summary (note that these results are from the 2012 class and so you will get slightly different numbers)
We are interested in whether females check their email more frequently on average than males do. To investigate this problem, we asked 71 females and 37 males enrolled in an undergraduate statistics class at the University of Colorado how many times they checked their emails per day. We found that females check their emails 5.6 times per day on average (SD=5.56) whereas males check their emails 3.8 times per day on average (SD=3.97). We conclude that female undergraduates do indeed appear to check their email more frequently, although we cannot say for certain if this difference arose by chance or not (exists only in this sample but not in the population).

I used the following R syntax to find these results:

```
males <- survey2013[survey2013$gender=="m",]  #this gave me a dataset of all the males
females <- survey2013[survey2013$gender=="f",] #this gave me a dataset of all the females
mean(males$email.check)
sd(males$email.check)
mean(females$email.check)
sd(females$email.check)
```

A couple of interesting syntax points to understand above. First, if we want to select rows or columns from a dataset, we do that using square brackets. The rows we want comes first, then a comma, then the columns we want. Here, for the "males" dataset, we want the rows where "gender is equal to 'm'". We then have a comma, and then nothing (which means we want ALL the columns). When we want to select a particular value for a variable, we use two equals signs in a row (==). NOTE: you cannot just paste the entire section above into the Script Window because the carriage returns from Microsoft Word will be messed up; paste each line in one-at-a-time.

Choose any dependent variable in the "survey2013" dataset you wish, and compare this dependent variable (its means and standard deviations) across males vs. females. You'll probably want to do this using syntax! Just try to mimic the syntax above, but choose a different dependent variable than "email.check". The dependent variable you choose should be a continuous or interval variable.

A) Identify the independent or quasi-independent variable. If it is a quasi-independent variable, explain why.

B) Identify the dependent variable. Is it integer or ratio?

C) What type of study are you conducting (e.g., an experiment?)? Can you make causal inferences that the independent variable caused a change in your dependent variable? Why or why not?

D) Attach a histogram of the dependent variable you chose for males and do this again for females. Comment on what the histogram shows you (i.e., compare the central tendency, spread, and shape of the two histograms). To make a histogram in R:

```
hist(survey2013$variable)       #where "variable" is whatever you choose
```

Or you can of course use the menu.

E) Use R to find the mean and the standard deviation of your dependent variable for males and for females.

F) What, in words, does the standard deviation mean?

G) When R calculated the standard deviation of your dependent variable, did it divide the sum of squares by n or by n-1? Why?

H) Write a four sentence summary of your findings.

<u>Part 2: To be done by hand:</u>

Question 4:

For the following set of scores: 33 26 208 12 37 25 34 29 26 30 33 15 35 38 31

A) Compute the mean, median, and mode (show your work)

B) What is the range? What is a disadvantage of the range statistic?

C) What is the median, 1st quartile, 3rd quartile, and inter-quartile range?

D) Which measure of central tendency do you think is best of this distribution? Why?

Question 5:

A sample of n = 5 scores has a mean of 10. One new score is added to the sample and the new mean is computed to be 11. **What is the value of the score that was added to the sample? Show your work.**

Question 6: (this is a thinking question)

A) John (an imaginary friend of yours) says that Mark is a jerk. **Given that John's information is the only information you have to go on, how sure are you that Mark is a jerk? What other possibilities might exist for why John said this other than that Mark really is a jerk?** Stay away from discussing how well you know or don't know John; just explain this in terms of it being one person's opinion.

B) Now say that Jill, James, Julie, and Janice also said that Mark is a jerk. **Does this change your subjective guess (i.e., your internal probability) that Mark really is a jerk? Why or why not?** Try to explain this in terms of the concept of "degrees of freedom".

C) Now say that you found out that Jill, James, Julie, and Janice are all friends of John's, and that you find out that John had been talking to the four of them about Mark before they told you that Mark is a jerk. **Does this new information change your subjective guess about whether Mark is a jerk? Why or why not (again, try to explain in terms of "degrees of freedom")?**