Homework #1: Basic concepts, distributions, and central tendency.
Due: In lab, week of Jan 28-Feb1

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.

Entering Data, Measures of Central Tendency, Measures of Spread
1. First, start R. In the R Console window, type:

```
library("Rcmdr")
```

This will launch the R interface, "R commander," which we'll be using throughout the semester.

These are the average high temperatures for two US cities depending on the month of the year:
Independence, MO: 35° 42° 56° 66° 76° 84° 92° 88° 80° 70° 52° 39°
San Francisco, CA: 55° 58° 60° 64° 66° 70° 74° 72° 71° 70° 62° 58°

a. Use R to find the mean average high temp in each city across the months of the year. To do this, you'll need to enter the scores into R Commander. In the Script Window (top), first create a dataset called "temps" with one variable being "indep" and the other variable being "sanfran", like this:

```
temps = data.frame(indep= c(35,42,56,66,76,84,92,88,80,70,52,39)
,sanfran=c(55,58,60,64,66,70,74,72,71,70,62,56))
```

That is all a single line of code (no returns carriages!), even though it shows up as two lines in a Microsoft word document like the one this homework was written in. [Note: I'll put code that you type into the Script window in red in homeworks]. To run that line of code, put your cursor on that line of code and hit <control> < r>. It will run in the "Output Window" below. Alternatively, you can put your cursor on that code and hit the "Submit" button. If you made an error entering the code above, the error message will show up in the "Messages" (bottom) window.

This command creates a *data set* named "temps," which in this case contains two vectors of length 12. Let's tell R that we want to work with the "temps" dataset. To do this, hit the button to the right of "Data set:" and choose "temps". This has loaded this dataset for us to work on.

Look at the dataset "temps" by hitting the button, "View data set." You can also change any entry in "temps" by hitting the button "Edit data set" – just like in Excel. There are buttons at the top of the window that pops up after hitting "Edit data set" that also allow you to add or

subtract rows/columns. All this can be done using scripting language too, but in this class we'll focus mainly on using the R-commander interface.

Hit the button "Edit data set" and **change the final (December) value of sanfran from 56 to its correct value of 58.** Notice that when you hit "Edit data set," the actual R code that can accomplish the same thing shows up in the Script Window. This is an extremely useful feature of R commander, because it allows you to learn the scripting language without all the frustration of having to type it in yourself. Make sure to close the editing/viewing window before you try to do anything else in R. So go ahead and close the editing windows now.

Now, find the mean temperature for San Francisco and the mean temperature for Independence. To do this, go to "Statistics" -> "Summaries" -> "Active Data set". This produces not only the mean of each variable, but also the min and max value along with the 1st and 3rd quartile (which we'll cover later). Notice again that the code for doing this is in the Script Window, so you could have typed:

```
summary(temps)
```

to get the exact same output. Alternatively, if you want the mean, go to "Statistics" -> "Summaries" -> "Numerical Summaries" and choose both variables in the window and put a check in the box next to "mean". **What are the means?**

b. There is a striking difference between the cities' average (or mean) high temperatures, but it isn't in the mean. **Describe in words what the difference is**. It might help to look at the data again (using "View data set", or alternatively, just typing "temps" in the Script window and hitting <control> <r> when your cursor is on "temps".

c. For a randomly chosen day out of the year, in which city are you more certain what the daily high will be? **Why?**

d. Create two histograms of the temperatures of Independence and those of San Francisco and paste them into your HW. Do this by going to "Graphs" -> "Histogram" and choosing each variable, one at a time. To save the figures as PDFs, go to "Graphs" -> "Save graph to file" -> "as PDF/Postcript" and then choose the smallest size possible and a unique name for each histogram. **Paste these histograms into your MS word homework document** by simply dragging the PDF into the open file and resizing it thereafter. Describe in words what these histograms are telling you (e.g., what is each bar, and what does the height of each bar tell you, and what is the x-axis?). If you're not familiar with histograms, which we'll cover in more detail later, ask a neighbor or your TA to help you.

e. The problem with the procedure above is that the limits (the extreme ends) on the x-axis are different for the two histograms. Sometimes, there are limitations to what you can do using R-commander. Let's make sure the x-axes are the same for the two histograms by modifying the code in your Script Window like this:

```
Hist(temps$indep,xlim=c(30,100))
Hist(temps$sanfran,xlim=c(30,100))
```

Once again, **paste these new histograms in your MS document**. **Describe how the histograms can visually tell you the major difference in average high temps between Independence and San Francisco.**

f. Before moving forward, you probably have lots and lots of code in your Script Window. Spend a few minutes cleaning it up so that only non-redundant and useful code is in there, then save that code by going to "File" -> "Save script as" and then saving it somewhere on your computer. This is very useful because you can recreate everything you've done above simply by re-running all that code – no need to remember all the mouse clicks, etc. If you want to make a note for yourself in the script (e.g., "here is how you place new limits on the x-axis"), you can do that by starting off your note to yourself with a hash mark (#). This is called *commenting* your script, and all good scripters do this.

Anyone who does not save their R scripts gets an F on homework! (**No need to turn in the R script or provide an answer here** – we trust that you're doing it! Only a fool wouldn't. Please note that you'll REALLY want these saved scripts for your lab final you'll take the last week of school)

Visualizing Distributions

2. For this next problem, will import real datasets into R to analyze. Do this by using the following code in the R-commander Script Window:

```
source("http://www.matthewckeller.com/HnrStat/loadData.R")
```

This will take a few seconds so be patient. Many datasets and new functions are pulled into your R session when you use the command above. In this case, we have imported a number of datasets and functions, including data about infants Apgar scores (named for Dr. Virginia Apgar). Apgar scores are given to infants when they are born. They measure a newborn's physical condition to determine if they need extra medical care. Low scores are worse; those above 8 are deemed "normal."

a. Load the apgar dataset by hitting the button next to "Data set" and changing it from "temps" to "apgar". To get an idea of what kind of data is in the "apgar" dataset, look at the first 6 lines using the function "head":

```
head(apgar)
```

If we want to see how many rows (infants) and columns (variables measured on each infant) there are in apgar, we could go to "View data set", or we could just use the function below:

```
dim(apgar)
```

The variable "score" is the one we're interested in here now. Use a histogram to look at the apgar scores of the 60 newborns. You can do this using the point and click method, as above, or you could type this into the Script Window and run it:

```
hist(apgar$score)
```

The dollar sign ($) following the name of the dataset is the way we tell R *which* variable in the dataset we're interested in looking at.

Describe the "shape" of the distribution of apgar scores –i.e., the central tendency, the spread, and whether the distribution looks normal, bimodal, skewed, etc.

b. Give the mean, min, and max of apgar scores.

c. Once again, clean up your code in the Script Window and save it. Take this code with you (e.g., email it to yourself) so that you have it for future reference (hint, hint!). **No need to give an answer here.**

Problems to be done by hand:

3. Use the data chart given below to answer the following 6 questions BY HAND. Make sure you can do these and similar types of problems by hand with ease – they'll be used from here on out and on all tests. If you need more practice, test yourself with new distributions!

i	X_i
1	8
2	4
3	1
4	3

a. Solve $\sum X_i$

b. Solve $\sum X_i^2$

c. Solve $\bar{X} = \frac{\sum X_i}{4}$ (this is the mean of *X*, denoted $\bar{X}$)

d. Solve $(\sum X_i)^2$ (note the order of operations - this answer is NOT the same as in b!)

e. Solve $\sum (X_i - \bar{X})^2$ (just substitute your answer in c above for $\bar{X}$; this is called the sum of squares of *X* in statistics, even though b above technically was also a sum of squares)

f. Given that the variable degrees of freedome (*df*) *df*=3, **solve** $\frac{\sum (X_i - \bar{X})^2}{df}$**.** Note that you don't know what a degree of freedom is yet – but you can still solve this. Just treat "df" as a variable!

Thinking problem

4. We are interested in testing whether Extra Sensory Perception (ESP) is real. In particular, we're interested in whether people can predict future events at levels better than chance. **In ~ a paragraph for each study, explain your design for two studies, one experimental study**

and one non-experimental study, that test this. After a brief description of the study, identify the hypothesis, the independent variable, the dependent variable, whether or not the study is an experiment, whether or not we can draw causal inference, and one limitation of each of the studies you've designed. Example limitation can be "we cannot draw causal inference" or that "the way we've studied this phenomenon may not apply to how it works in the real world (i.e., our construct has low external validity)" or "the sample we have used is not representative of the population", etc.