LING280 Problem Set 1: R Warmup
Due Monday April 4 by 4pm to mwagers@ucsc.edu

BACKGROUND
In this problem set, you will be working with a data fragment from a large acceptability study on island constraints. The materials consisted of biclausal constituent questions, designed by crossing two factors. Each factor had two values: (a) the head of the embedded clause: \{that, whether\}; (b) the site of extraction: \{matrix subject, embedded object\}. An example lexicalization of this design is given below.

a. \textit{that, matrix subject}
   Who \_ thinks that John bought a car?
b. \textit{that, embedded object position}
   What do you think that John bought \_?
c. \textit{whether, matrix subject}
   Who \_ wonders whether John bought a car?
d. \textit{whether, embedded object position}
   What do you wonder whether John bought \_?

142 undergraduates gave 2 judgments on each of the resulting 4 sentence types. They made their judgments on a 7 point scale.

PROBLEMS
1. Import the sample data set, which is available as a CSV file here:
   \url{http://ling.ucsc.edu/~wagers/ling280/ps1.csv}

2. Summarize the data in several different ways. For each of the four conditions in the design, determine its mean, standard deviation, median, and inter-quartile range. Include a code fragment with annotations. See Problem 6 for an example of an annotated code fragment.

3. Assuming the data are stored in a dataframe called \texttt{judge}, what does the following command do? Identify each function, argument or operator and explain its role.

   ```r
   > apply( judge, 2, table ) -> judge.tab
   ```

4. For each of the four conditions, what proportion of judgments were assigned a value of ‘6’ or ‘7’? Include the annotated code fragment you used to compute this. Note, you should not build in any constant values: if I run the experiment again with different numbers of individuals (or judgments), I should be able to run this fragment of code and get the correct answer.

5. Create a boxplot of the data. Make sure each condition is represented in the same plot (i.e., don’t turn in 4 separate plots; but one plot containing 4 boxes).

6. What’s the difference between the two commands below?

   ```r
   > hist(judge[, 'whether.matrix'],breaks=7,freq=TRUE)
   > hist(judge[, 'whether.matrix'],breaks=7,freq=FALSE)
   ```
7. Consider the following code fragment. Before you execute it, read through it carefully and try to figure out what it does. After you execute it, examine the resulting boxplot. What is going on? In a paragraph, first explain what the code does. Then explain what can be deduced about the relationship between (a) the number of observations made in an experiment and (b) the precision and accuracy of that experiment's estimate of the random variable. Be specific and justify your claims with reference to the information conveyed by features of the boxplot.

```r
### CODE FRAGMENT FOR QUESTION 6
### Create a matrix, rep.experiment, with 1000 rows and 5 columns
rep.experiment <- matrix(nrow=1000, ncol=5)

### Assign the data from Condition D to its own vector, called island.judgments
judge[, 'whether.embedded'] -> island.judgments

### Run a loop that iterates from 1 to 1000
### Each loop creates a replicant of 5 different experiments
for(i in 1:1000){
    ### SIMULATE AN EXPERIMENT WITH FIVE OBSERVATIONS
    ### Sample 5 data points, with replacement, from island.judgments
    sample(island.judgments, 5, replace=TRUE) -> n5;
    ### Compute the mean of this sample and
    ### store it in row i, column 1 of rep.experiment
    mean(n5) -> rep.experiment[i,1];

    ### SIMULATE AN EXPERIMENT WITH TEN OBSERVATIONS
    sample(island.judgments, 10, replace=TRUE) -> n10;
    mean(n10) -> rep.experiment[i,2];

    ### SIMULATE AN EXPERIMENT WITH TWENTY OBSERVATIONS
    sample(island.judgments, 20, replace=TRUE) -> n20;
    mean(n20) -> rep.experiment[i,3];

    ### SIMULATE AN EXPERIMENT WITH FORTY OBSERVATIONS
    sample(island.judgments, 40, replace=TRUE) -> n40
    mean(n40) -> rep.experiment[i,4]

    ### SIMULATE AN EXPERIMENT WITH EIGHTY OBSERVATIONS
    sample(island.judgments, 80, replace=TRUE) -> n80
    mean(n80) -> rep.experiment[i,5]
}

### Visualize the location/spread of the means from the 5 different experiments
boxplot(rep.experiment,
    names=c("n = 5", "n = 10", "n = 20", "n = 40", "n = 80"),
    ylim=c(1,7),
    ylab="mean rating",
    xlab="sample size",
    main="Simulated outcome of 1000 experiments")
```
8. Here’s an unannotated code fragment. Annotate each line. What does this code compute? And what do you conclude from its output?

```r
numbers<-seq(0,7,by=0.05);
squared.deviation <- vector(length=length(numbers))
k<-1
for(n in numbers){
    sum((n - island.judgments)^2)->sum.squares
    sum.squares /( length(island.judgments) - 1 ) -> squared.deviation[k]
    k<-k+1;
}
plot(numbers,squared.deviation,pch="*")
abline(h = min(squared.deviation) )
```