Homework 1

Psychology 310

Instructions. These questions are based on material in G&H, Chapters 1–8, and lectures. Answer the following questions. Show your R code, your input, and your output. Feel free to email me for hints if you get stumped.

1. (10 points). Load the Chapman data into R, using the `read.table` command used in Lab 1. The data are in a file called *CHAPMAN.TXT*, available for download from the website. `attach` the data. Then produce:

   (a) A stem-leaf diagram of the CHOLES data.

   (b) A histogram of the CHOLES data.

   (c) Go to www.statmethods.net, and find the section on boxplots in the graphics area. The first example will show you how to produce a boxplot using one or more grouping variables. Study this example, and also type `?boxplot` in R and take a look at the help file (while trying to remain unintimidated). Following this example, produce side-by-side boxplots of the `CHOLES` data broken down into two groups, based on the `CORON` variable. This variable is 1 if the individual has had a coronary event, 0 if the individual has not. Label the boxplot so that the Y-axis is properly labeled, and the X-axis shows proper labels for the two variables (Hint: look for the `names` option).

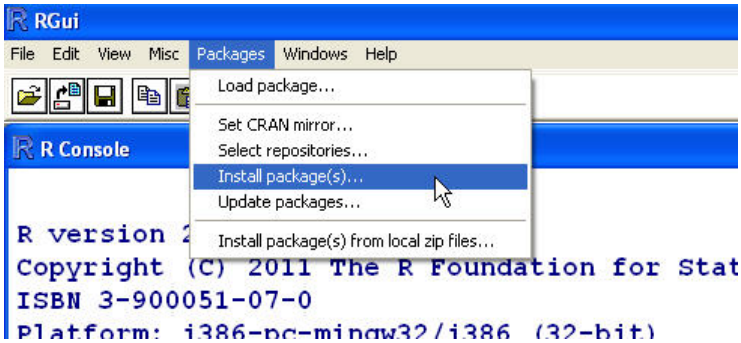
   (d) Describe what the boxplot conveys to you about the difference between the two groups.

2. (10 points.) You need to load the High School and Beyond data set. This is a big data file, with a number of "nominal" or qualitative variables. These are coded numerically, but in the original file they also have what are commonly referred to as "value labels." Value labels are verbal labels that are matched to the integer values (for example, `1=MALE, 2=FEMALE` for the `SEX` variable). Most modern software packages give you the option of using either the value labels or the integer values.

   Reading in the text file *HSB.TXT*, attaching the variable names, and entering the value labels is an extensive operation that is described nicely on the statmethods.net website. However, your textbook also provides SPSS files that have the variable names and value labels already embedded in them. R can read these directly and convert them, so we'll save some time and use them.

   Download the SPSS file *HSB.SAV*, and put it in your workspace. Then follow the instructions on *www.statmethods.net* for importing an SPSS file. Notice that you will need to load the `Hmisc` library with the command `\library{Hmisc}`. If you don't already have this library, you will get an

error message saying that there is no package `Hmisc`. If you get this error message, it means that this package is not yet installed in your computer's version of R, and you will need to download and install it. To do this, open up the *Packages* menu in R, and select *Install Package(s)*.



If this is your first package installation in this R session, you will be prompted to select a download mirror site. I generally choose a U.S. site relatively close to Tennessee. Once you have selected a download site, R should proceed to download and install the package. If you then type `\library{Hmisc}` again, you should see no error message, and the package should be loaded successfully.

Produce a side-by-side notched boxplot like the one on page 28 of G&H. A general guideline is that if the notches fail to overlap, then the medians of the two groups are "signficantly different." Are the medians significantly different for Public and Private schools? Does your plot differ from the one in G&H?

3. (15 points.) Suppose you have the following 5 rows and 4 columns of data in the array X.

	1	2	3	4
1	11	4	9	11
2	11	11	14	6
3	9	8	10	8
4	8	8	11	12
5	11	6	7	10

Evaluate the following expressions

(a)

$$\overline{X}_{\bullet 2}$$

(b)

$$\sum_{i=1}^{5} X_{i3}$$

(c)

$$X_{\bullet\bullet}$$

(d)

$$\sum_{i=2}^{5}\sum_{j=1}^{4} X_{ij}$$

(e)

$$\sum_{i=1}^{5}(X_{i\bullet})^2$$

4. (10 points.) In class, we demonstrated a formula to compute the mean of a combined group from the means and sample sizes of the original groups. This formula generalizes to any number (J) of groups, and can be written as

$$\overline{X}_{\bullet\bullet} = \frac{\sum_{j=1}^{J} N_j \overline{X}_{\bullet j}}{\sum_{j=1}^{J} N_j}$$

The numerator is a sum of products of the Ns and the group means, and the denominator is the sum of the Ns. Write an R function to compute the mean of a combined group. Your function call should follow these specifications:

```
> combined.mean(Ns,Xbars)
```

where

Ns = a vector of group sample sizes Xbars = a vector of group means

Here is the example from G&Hp. 57:

```
> combined.mean(c(24,33,27),c(6.91,7.72,6.75))
[1] 7.176786
```

5. (15 points.) Answer the following questions about the normal distribution:

   (a) A student received a grade of 76 in a course where the class average was 70, and the standard deviation 10. If the class distribution was approximately normal in shape, what was the student's approximate percentile rank?

(b) If SAT scores have a mean of 500 and a standard deviation of 100, approximately what percentage of students obtain SAT scores between 550 and 650?

(c) *Statistical Detective.* Suppose that the class distribution in a large course is almost exactly normal in shape. Joe got an 88 and had a percentile rank of 79.8, while Felicia got a 75 and had a percentile rank of 40.1. From this information, estimate the mean and standard deviation of the class distribution. (*Hint.* The percentile ranks tell you the Z-scores, and they are a function of the mean and standard deviation.)

6. (10 points.) Given the following course grades for 5 students. 78,64,67,91,72. If these grades are lineary scaled to have a mean of 70 and a standard deviation of 12, what grade will the student with a 91 receive? (Show all work, or show all R code used to solve the problem.)

7. (10 points.) X has a mean of 70 and a standard deviation of 10. Y has a mean of 75 and a standard deviation of 12. X and Y correlate .60. Assume that X and Y have a bivariate normal distribution.

   (a) What is the linear regression equation for predicting Y from X?

   (b) Among people who score 80 on X, what percentage score above 80 on Y?

8. (10 points.) You have 9 numbers with a mean of 10 and a variance of 100. If you add a 10th number to this group, and this number is 15, what will be the variance of the new list of numbers?

9. (10 points.) You are the football coach at Nardine High. Nardine has 200 boys. If they are a random sample from a normally distributed population with a mean weight of 150 and a standard deviation of 25, what would you expect the heaviest boy in your school to weigh?