Homework #1 – Due Monday, Sep. 26
COR1-GB.1305 – Statistics and Data Analysis

The problems in this assignment require data files, which are available at `http://ptrckprry.com/course/langone/`. You are permitted to work in teams, but each student must independently write up their own solutions (no direct copying).

Electronic submissions are not accepted. Print out and turn in your Minitab plots. See the "Minitab Tips" handout for tips on opening files and printing or saving plots.

Problem 1

The file `CelebrityEarnings.CSV` contains the personal earnings in 2007–2008 for the Forbes Top 10 Most Powerful Celebrities. Use *Stat* ⇒ *Basic Statistics* ⇒ *Display Descriptive Statistics* to compute the mean, median and interquartile range, for the earnings. You will need to obtain the inter-quartile range by hand as the difference between the third quartile (Q_3, the 75th percentile) and the first quartile (Q_1, the 25th percentile). Why do you think that the mean is greater than the median? To help answer this, look back at the data set, and also create a histogram, using *Graph* ⇒ *Histogram* ⇒ *Simple*. Next, remove the two largest observations, for J. K. Rowling and Oprah Winfrey, by replacing their earnings in the Minitab spreadsheet by `*` (a missing value). Recompute the mean, median and interquartile range for the modified data set. Compare with the corresponding values for the full data set. What changed and what stayed essentially the same? Why?

. . . . . .

Problem 2

The file `USDomesticBeer.CSV` contains data on the percentage of alcohol, calories per 12 oz serving and carbohydrates (grams) for 101 US domestic beers.

1. Run the graphical summary for *%Alcohol*, using *Stat* ⇒ *Basic Statistics* ⇒ *Graphical Summary*. Based on the histogram (with best-fitting normal curve superimposed) and the boxplot under it, do the data appear to be normally distributed? If not, what patterns do you see in the distribution? (By the way, Minitab's Kurtosis is really the excess kurtosis. This is a measure of "tail-heaviness", and should be zero for a normal distribution. The more positive the kurtosis, the more prone to outliers the data set is.)

2. Let the cursor rest over the leftmost asterisk in the boxplot. Minitab should respond with a popup giving you the corresponding value and case number. Go back to the data spreadsheet to identify this outlying value. Repeat for the next-leftmost asterisk in the boxplot.

3. Use the values of these two outliers together with the mean and standard deviation to compute the corresponding z-scores. Based on the empirical rule, does it seem plausible that the data set came from a normal distribution?

. . . . . .

Problem 3

Consider `Market.CSV`, in which *MarketReturn* and *IBMRet* are daily excess returns in %/Day. (The riskfree rate has already been subtracted out from both variables, and *MarketReturn* is value-weighted over all NYSE, AMEX and NASDAQ stocks.) Run Minitab's *Descriptive Statistics* for both variables. Which investment (Market or IBM) has the highest mean return? Which has the highest Sharpe ratio? Why do the answers to these two questions differ? (The Sharpe ratio is the ratio of the mean to the standard deviation of the excess return. It measures the return expected for a given amount of risk.)

. . . . . .

Problem 4

The file `NormTemp.CSV` contains data on body temperatures for 130 randomly selected subjects. The first column (*Temp*) contains the temperatures themselves. For each subject, this temperature, in degrees Fahrenheit, represents an average of several readings taken over the course of two consecutive days. The second column (*Gender*) is 1 for male, 2 for female, and the third column (*HeartRate*) is measured in beats per minute. Here, we focus on *Temp*.

1. Make a histogram of *Temp*. Does the data seem to have a reasonably bell-shaped distribution? Do you see any outliers?

2. What do you think the population mean is for body temperatures? (Presumably, youve been hearing this number since you were very young! If you were raised on Celsius, convert to Fahrenheit using $F = \frac{9}{5}C + 32$.)

3. Based on the histogram, does the sample mean seem to be reasonably close to the "known" population mean? Don't actually calculate the sample mean, just look at the histogram. (Hint: If a distribution is symmetrical, then the mean is the center of symmetry.)

4. Use *Descriptive Statistics* to calculate the sample mean. What is the value of the sample mean? Is it reasonably close to the "known" population mean? And how did you define "reasonably close"?

. . . . . .