Homework 1

To do the HWs you will need to either use JMP in the lab or download it on your computer.

On my website I have posted the M&M data. Download this data set and upload it into JMP (the JMP instructions are also given on my website). Do not hand in the JMP commands (we do not need to know this, and it wastes trees). Just hand in the relevant plots and solutions.

(1) What sort of variable is

(i) The type of cycle a person owns.
(ii) The height of a person.
(iii) The gender of a person.
(iv) The number of a bus.

(2) What is the mean and standard deviation of the following samples:

(a) 1, 1, 1, 1, 1.
(b) -99, -49, 1, 51, 101.

(3) Sidney observes the data

1, 1.5, 2, 3, 3.5, 3.8, 6, 6.5, 7, 8.

The sample mean and standard deviation of this sample is 4.23 and 2.47 respectively. What happens to the mean, standard deviation and quartiles of the sample if the last observation 8 is replaced with a much larger number?

(A) The mean and standard deviation stay above the same, but the quartiles and median change.
(B) Only the mean will change.
(C) The first quartile and median stay the same, but the third quartile, mean and standard deviation become larger.
(D) The quartiles and median stay the same, but the mean and most likely the standard deviation will become much larger.
(E) Unless we know the actual new number it is impossible to say what will happen to the values.

(4) Input the M&M data into JMP. We want to see whether there is any difference in the distributions of number of Peanut, Peanote Butter and Milk chocolate M&Ms (later on we will use statistical tools to see if these difference are ‘real’).

(i) Make boxplots and histograms of the total number M&Ms for each type (Peanut, Peanut butter and Milk chocolate).

Go to Analyze, Distribution and put Total into Y. column, then put Type into By.
(ii) Compare their distributions (through histograms and boxplots). Are the distributions roughly the same (in the sense of mean, median, IQR (the difference between the third and first quartile) and standard deviations)?

(5) As we go through the course, of the fundamental results will be that the sample mean (average based on the sample) is random with a distribution. This distribution is special, it's spread (standard deviation) is less than the sample itself (original observations) and its shape is quite unique. The only problem for most real life problems we won't be able to 'see' the distribution of the sample mean, since we have only one sample mean (the histogram of one number is not very informative). Therefore, I artificially make several small samples from a big sample (by grouping the 170 bags of M&Ms into bags of five).

(i) Calculate the standard deviation of the total number of M&Ms (you can do this in JMP).

(ii) In the fifth column each of the M&M bags have been put into one of 34 groups (there are 5 bags in each group). Each group is labelled by a number going from 1-histo. For each group (of 5 bags) we want to calculate the average number of M&Ms (so in the end you will have 34 averages). It is quite cumbersome to do this by hand, so here is how to do it in JMP (it is just like the the average weight of calves for each group that we did in class).

In JMP Going to Tables → Summary, select Group to go into the Group box, highlight total, then in the Statistic you should choose mean (this is a rather complicated, go to a lab session if you need help). You should now get a new table containing the average in each group, it will have 34 rows. You will have to data sets, the original M&M data set and also the data sets containing all the means that you created.

(iii) Using the new table calculate the mean and standard deviation of the groups means (this column is denoted as Mean (Total)).

(iv) Compare the standard deviation for the entire class and the standard deviation of the averages. Comment on their differences.

(v) Make a histogram for the total number of M&Ms and also the averages. Compare the shapes of the two the histograms, are they similar or very different?