Math 324 Fall 2004
Assignment 1
Due: Sept 22, 2004

Dr Ben Bolstad
bolstad_math324@bmbolstad.com
http://math324sfsu.bmbolstad.com

This assignment is intended to give you practice using a statistical package and the opportunity to perform some EDA techniques as applied to both real and simulated data. You should submit your solutions to this assignment as a written report. In other words, it is not enough to submit pages of computer output with no interpretation. If you wish to submit your analysis code, please attach it as an appendix to your report.

Part I - Exploratory Data Analysis

Current research states that adults should consume no more than 30% of their calories in the form of fat, they need about 50 grams (women) or 63 grams (men) of protein daily, and should provide for the remainder of their caloric intake with complex carbohydrates. One gram of fat contains 9 calories and carbohydrates and proteins contain 4 calories per gram. A "good" diet should also contain 20-35 grams of dietary fiber.

Download the datafile *cereals.dat* from the course webpage. Note the datafile is a tab delimited text file and the first row contains column names. This datafile consists of measurements of 16 different variables on 77 different types of breakfast cereals. In particular, the variables measured are:

- Name: Name of cereal
- mfr: Manufacturer of cereal where A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina
- type: cold or hot
- calories: calories per serving
- protein: grams of protein
- fat: grams of fat

- sodium: milligrams of sodium
- fiber: grams of dietary fiber
- carbo: grams of complex carbohydrates
- sugars: grams of sugars
- potass: milligrams of potassium
- vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
- shelf: display shelf (1, 2, or 3, counting from the floor)
- weight: weight in ounces of one serving
- cups: number of cups in one serving
- rating: a rating of the cereals

Note that a value of -1 for any nutrient indicates a missing value.

Use the exploratory data analysis techniques discussed in class to learn more about this data. In particular, you should consider using histograms, stem plots or boxplots to explore the data. You may also wish to use scatterplots to look at relationships between variables. Also be sure to report any summary statistics you might compute. Your goal is to discover any interesting features of this data. Some questions you might wish to explore

1. Is there any relationship between the shelf and the grams of sugar? Can you suggest a reason for this?
2. The "ratings" were computed by Consumer Reports. Use scatterplots or other means, to explore the relationship between the each of the nutrient measurements and the ratings values. Describe the relationship, ie increasing or decreasing, strong or weak, linear or non-linear.
3. Are there any cereals that seem to be outliers in any of your plots? Why or why not?

Part II - Histogram binning

You may have wondered about how many bins you should use for a histogram. Suppose that there are n data values. Several common rules of thumb are

- use $k + 1$ bins where k is given by the value of 2^k closest to n.
- choose a bin width of $2IQR/n^{1/3}$
- number of bins is approximately $1 + 3.3 \log n$ for $n \geq 15$

- number of bins $\geq$ (range/binwidth) $\geq (2n)^{1/3}$

Using which ever statistical package you have chosen to do the following:

1. Simulate 10, 20, 50, 100, 1000 random numbers from the standard normal distribution (mean 0 and variance 1)

   For each sample create histograms with various numbers of bins ranging from 3 to twice the number specified by one of the "rules of thumb".

2. Simulate 10, 20, 50, 100, 1000 random numbers from the exponential distribution (mean 1)

   For each sample create histograms with various numbers of bins ranging for 3 to twice the number specified by one of the "rules of thumb"

Report your results, interpret what you have learned. Note you do not need to submit every plot. A representative selection of plots with a good description is sufficient. If possible try to draw your histograms using relative frequencies and use the same size vertical scale on each plot.