Statistics 335 – Assignment 1
Due: Thursday, October 9, 2003

This first assignment is to give you an opportunity to work with S-Plus or R, dealing with reading in data, its manipulation, performing simple analyzes, and dealing with output. As part of this assignment, you will probably need to access the help pages or the documentation, which is available from the S-Plus / R page. Note that we currently do not have version 6 of S-Plus available but version 5.1.

Please submit your solutions in a word processed format (I don't care which one you use). Also as part of your solutions, please give the commands used to generate the output. I'm curious to the approaches you take as some of the questions have multiple valid ways of getting the answers. For example, I can think of two ways of generating the set of summary statistics desired in one of the early questions. Also in your figures, where appropriate, please give informative axis label, figure titles, etc.

For this assignment, please use S-Plus, not R.

1. Download the data iris.dat from the course web site (either from the Datasets page or the Assignments page) and read the file into S-Plus using the `read.table` command into a data frame named `Iris`. This data set is the famous Fisher Iris data. It contains 4 variables (Sepal Length, Sepal Sepal Width, Petal Length, and Petal Width) on 50 flowers from each of 3 species (Setosa, Versicolor, and Virginica) of iris. All size measurements are given in centimeters. For a more complete description of the dataset, check the description file which is also available at the same locations as the dataset.

2. Calculate the standard summary statistics (mean, standard deviation, median, 1st and 3rd quartiles, min, and max) for the four variables for the combined data set and for each species.

3. Create a new data frame with the measurements given in inches instead of centimeters (use the conversion 1 inch = 2.54 cm). Show the first 5 rows of the two data frames.

4. Draw histograms of the four numeric variables. Please combine the four histograms into a single figure (using `par(mfrow=c(2,2))`). When creating figure, make sure it is clear which variable is being plotted in each histogram with an informative label.

5. For the two Petal variables, draw side by side boxplots for each species with the default settings for boxplot for S-Plus. Figure 1 show the boxplot of Petal Width as generated in R. You should notice that the style of the default boxplot for the two programs is different. Examine the help pages for boxplot (`help(boxplot)`) to find out how to make the S-Plus boxplot output look similar to the R boxplot output. (I don't think you can get them exactly the same.) Just describe how to do this; there is no need to include the plot.

6. Create a scatterplot of Sepal Width (x axis) vs Sepal Length (y axis), superimposing the least square regression line on the plot. In addition, create redo this scatterplot, but with different plotting symbols for each species. For this second plot, add the least squares regression lines for each of species. Use different line types for the different species. Use the legend command to indicate which symbol and line correspond to the different species.

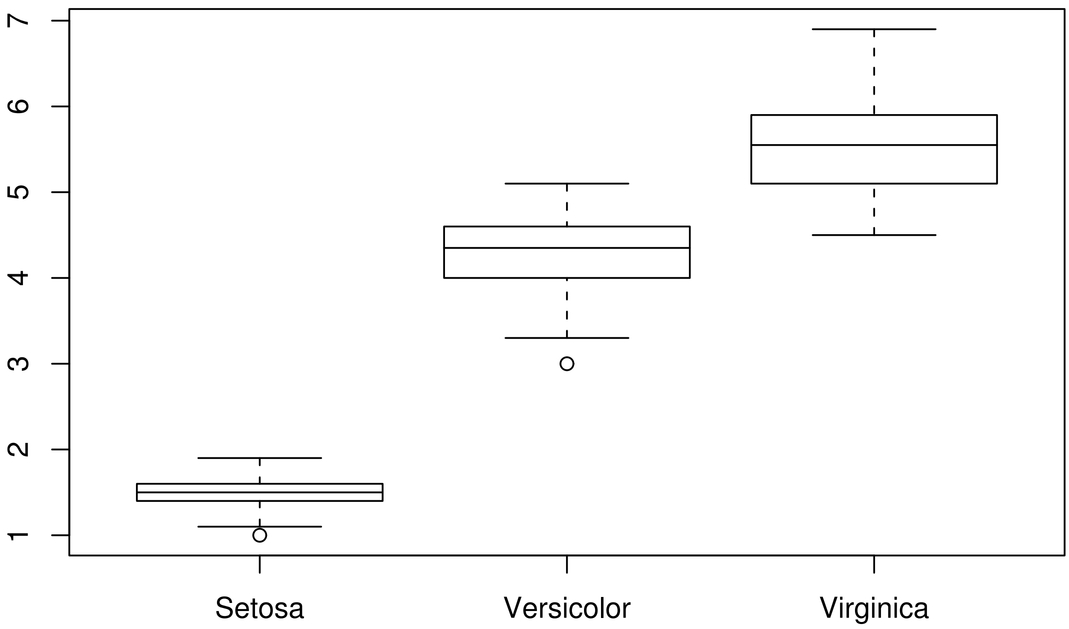

Figure 1: Boxplot of Petal Width generated by R

7. Most of the calculations involved in doing regression involve doing matrix commands. For example, the least squares estimates the parameters of the standard linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; i = 1, \ldots, n$$

are

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

where $\hat{\beta}^t = [\hat{\beta}_0, \hat{\beta}_1]$, $\mathbf{y}$ is a column vector containing the n observations of the response variable, and $\mathbf{X}$ is an $n \times 2$ matrix where the first column contains n 1's and the second column contains the n observations of the explanatory variable. To estimate σ^2, the variance of the ϵ_i's, the usual estimate is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, , n$ are the fitted values. The vector of fitted values can be gotten by $\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$.

Verify that the S `lm` function gives the same estimates as the calculations described above. To run the `lm` command on this data and get the estimated regression parameters and estimated residual variance, use the following commands:

```
petal.lm<-lm(PetalLength ~ PetalWidth,data=Iris)
coef(petal.lm)
var(residuals(petal.lm))
```

8. It is possible to show that one way to generate an exponential random variable (RV) with mean μ is to generate a Uniform(0,1) RV, U and make the transformation $V = -\mu \log U$.

   (a) Generate 100 uniform random variables and store them in a vector `u`. (`u <- runif(100)`). Produce a histogram of the vector `u`.

   (b) Generate 100 exponential random variable with mean 2 using the S function and store them in a vector `v`. (`v <- rexp(100,0.5)`). Note: S parameterizes the exponential distribution in terms of the reciprocal of the mean instead of the mean. That is why 0.5 in the function for generating the random variable.

   (c) Using the 100 uniforms generated in part (a), generate 100 exponentials with mean 2 and store them in a vector `w`. Examine whether the vectors `v` and `w` appear to really have the same distribution by examining summaries and graphs of the data.