**Lab 5**

**1. For the iris data set:**

(1) How many features are numerical data? And how many are categorical data

(2) Make histogram for the numerical features, to see how they distribute

(3) Make table for the categorical features, to see how they distribute

**2.** The data set rivers contains the lengths (in miles) of 141 major rivers in North

America.

(1) What proportion are less than 500 miles long?

(2) What proportion are less than the mean length?

(3) What is the 0.75 quantile?

**3.** Fit a density estimate to the data set pi2000 (UsingR). Compare with the appropriate histogram. Why might you want to add an argument like breaks =0:10-0.5 to hist()?

**4.** The data set DDT (MASS) contains independent measurements of the pesticide

DDT on kale. Make a histogram and a boxplot of the data. From these, estimate the mean and standard deviation. Check your answers with the appropriate functions.

**5.** It can be illuminating to view two different graphics of the same data set at once.

A simple way to stack graphics is to specify that a figure will contain two graphics by using the command

> par(mfrow=c(2,1) # 2 rows, 1 column for

graphic figures

Then, if x is the data set, the commands

> hist(x)

> boxplot(x, horizontal=TRUE)

will produce stacked graphics. (The graphics device will remain divided until you change it back with a command such as par (mfrow=c(1, 1)) or close the device.) For the data set lawsuits (UsingR), make stacked graphics of lawsuits and log (lawsuits). Could you have guessed where the middle 50% of the data would have been without the help of the boxplot?

6. Let sex = c(1,1,1,1,2,2,2,2,2,2); smoking = c(1,0,1,0,1,0,0,0,1,1); age=c(31:40) in R. A data frame is constructed as zz = data.frame(sex, smoking, age). Give the results of following R commands:

1. apply(zz[-1,], 2, min)
2. zz[zz[,3]>35,]
3. zz[order(zz["smoking"], zz["age"]), ]
4. subset(zz, zz["sex"]==1)
5. tapply(zz$age, zz$smoking, max)
6. apply(zz[,-3], 1, function(x){ sum(x) })