RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS

R WORKSHEET 4

This worksheet is an exercise for you to do yourself. Worksheets 3 and 4 are more advanced and are presented for students who wish to learn more about R and explore some of the more advanced features of the software. Before attempting this worksheet you should have completed Worksheets 1, 2 and 3.

Exercise 1.

Produce a plot of the function

$$f(x) = e^{-|x|},$$

for values of x ranging from -3 to 3. This function is sometimes called the *double exponential* function. Make sure to appropriately label your plot and axes.

Exercise 2.

A study on the age and growth characteristics of selected mussel (shellfish) species in two distinct locations i n S outhwestern Virginia, USA was conducted. The data for this study is contained in a data set called `PROB3.8` which loads automatically when you start R. This file contains three columns: the first indicates the location (region 1 or region 2) where the data was collected; the second contains the ages of the selected mussels; and the third column gives the weight (in grams) of the selected mussels. Name the columns of the object "location", "age" and "weight" using the `names` command.

On the same set of axes, plot the weight versus age relationship for each of the two locations, connecting the datapoints within each location with lines. Make sure to use distinct symbols and line-types for the data from the two different locations and that you label the plot appropriately. [Hint: look at the help file on `lines()` for instructions on how to overlay plots.]

Use the `lsfit()` function to fit simple linear regressions to the data from each of the two locations separately. On the same set of axes, plot the weight versus age relationship for each of the two locations, and superimpose the two regression lines. Again, make sure to use different symbols and line-types for the two different sets of points and regression lines and also to appropriately label your plot.

Exercise 3.

Recall the difference between a sample median and a sample mean - the median is less sensitive to outliers, but the mean is better for normally-distributed data. To check this out, create a sample of size 50 from a standard normal distribution, and find the mean and median for the data. Now add 100 to the first observation in your sample, to make it an outlier. How are the median and the mean affected by the outlier?

Now, do some simulations to compare the performance of the median and the mean for the standard normal and the Cauchy distributions. The latter typically produces samples with apparent outliers. First, create space to store the results:

```
means.from.normal <- rep(0, 100)
medians.from.normal <- rep(0, 100)
means.from.cauchy <- rep(0, 100)
medians.from.cauchy <- rep(0, 100)
```

Next, create 100 means and medians from samples of size 50 from the standard normal and Cauchy distributions and put them in the vectors created above:

```
for(i in 1:100) {
    x <- rnorm(50)
    y <- rcauchy(50)
    means.from.normal[i] <- mean(x)
    medians.from.normal[i] <- median(x)
    means.from.cauchy[i] <- mean(y)
    medians.from.cauchy[i] <- median(y) }
```

Now, investigate the results, using histograms, means, medians and standard deviations.

Exercise 4.

This question asks you to write a new *R* function to implement a method called the *jackknife*, which can be used to examine the bias of estimation procedures in many situations. We don't want to go into the theory behind the jackknife here, but we would like to write an *R* function to implement the jackknife bias calculation for the median.

We start with a dataset of size n, say, with data points $X_1, X_2, \ldots, X_n$, which are contained in a column vector `X`. The basic idea of the jackknife is to form n new datasets, each one the same as the original dataset except that the i^{th} new dataset is missing the i^{th} point from the original dataset. The new datasets are therefore each of size $n-1$: for example, the first is $X_2, X_3, \ldots, X_n$ (i.e. the original dataset missing the first data point), the second is $X_1, X_3, \ldots, X_n$ (i.e. the original data missing the second data point), and so on. Now, for each of these reduced datasets, calculate the median and construct a vector containing the n differences of each of these medians from the overall median of the entire original dataset. In other words, create a vector of length n, whose i^{th} component is the difference between the original median and the median of the reduced dataset. The jackknife estimate of the bias of the median is then just $n-1$ times the average of this vector of differences.

Write an *R* function that takes a columns of data, and returns the jackknife estimate of the bias for the median.