STA 437/1005, Fall 2008 — Assignment #2

Due on October 20, at start of lecture. Worth 10% of the course grade.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion of this assignment with any written notes or other recordings, nor receive any written or other material from anyone else by other means such as email.

For this assignment, you will analyse two multivariate data sets using the built-in facilities of R, and and also a set of R functions that I have written for performing the tests discussed in Chapter 5 of the text. The data and these functions are available from the course web page,

`http://www.utstat.toronto.edu/~radford/sta437`

Look in the section for Assignment 2. There are also some hints about R functions that may be useful for this assignment.

For both questions, you should hand in a discussion of what you did and what your conclusions were. Your analysis should include whatever specific things are requested below, but you may also include other things if you think they are of interest. You should hand in a print-out of the file of R commands that you used to obtain your main results. (Of course, you may have just typed in some R commands as well, but you should put the commands that are crucial to understanding how you analysed the data in a file that you hand in.) You should also hand in a moderate amount of R output and/or R plots that justify your conclusions.

Question 1:

In this question, you will look at data on the returns of various sorts of US investments for the years 1949 to 1995. The return is expressed as the value at the end of the year of one dollar invested at the beginning of the year, after adjusting for inflation. (For examle, if \$1.00 invested at the beginning of a year produced \$1.43 at the end of the year, but inflation during that year was 10%, the number recorded in the file would be $1.43/1.10 = 1.30$.)

The data file contains a one-line header with the names of the variables, followed by 47 lines of data. The variables are as follows:

`year`	Year, from 1949 to 1995
`gbonds`	Return for long-term government bonds
`gbills`	Return for short-term government bills
`cbonds`	Return for long-term corporate bonds
`indust`	Return for industrial stocks
`transp`	Return for transportation stocks
`util`	Return for utility stocks
`finance`	Return for finance stocks

You should look at this data to see to what extent the returns in these years can be seen as a random sample from a hypothetical distribution of returns for these investments, which you might expect to be a guide to returns in future years. You should also look at whether or not the returns seem to be normally distributed.

You should consider transforming the returns by taking the log. You should first look at whether the log returns seem to be normally distributed. You should also ask whether knowing the mean of the returns or the mean of the logs of the returns is more useful to an investor. In particular, consider these things an investor might want to know:

- Rather than putting all their money in only one type of investment (eg, industrial stocks), an investor might be willing to split their investment among several types (eg, one-third each in industrial stocks, finance stocks, and long-term government bonds). How would they determine the expected return with such a split?
- Most investors are not interested in returns over only a one-year time-span. Consider an investor who intends to stay invested for many years, say 40 years. How would they determine how much they might expect to have for each dollar invested 40 years earlier?

For both the returns as given in the data file, and the logs of these returns, you should find 90% and 95% confidence intervals for the mean for each of the seven types of investments. Compute three kinds of confidence intervals: simultaneous confidence intervals based on the T^2 statistic, univariate confidence intervals based on the t statistic, and univariate confidence intervals with the Bonferroni correction. You should also find estimates of the standard deviations of the returns and of the log returns, which are measures of how risky these investment are.

Discuss what your results say about what one might expect in the future from these types of investments.

Question 2:

In this question you will look at data on IQ and various physical measurements for 10 pairs of monozygotic ("identical") twins (when they were adults). The question of interest is whether or not the twin born first differs from the twin born second, and if so in what respects they differ.

The data file contains a one-line header with the names of the variables, followed by 10 lines of data. The variables are as follows:

`sex`	Sex of the twins (always same), 1=male, 2=female
`IQ1`	IQ of twin born first
`head1`	Head circumference (cm) of twin born first
`area1`	Surface area (cm^2) of brain of twin born first
`volume1`	Volume (cm^3) of brain of twin born first
`weight1`	Body weight (kg) of twin born first
`IQ2`	IQ of twin born second
`head2`	Head circumference (cm) of twin born second
`area2`	Surface area (cm^2) of brain of twin born second
`volume2`	Volume (cm^3) of brain of twin born second
`weight2`	Body weight (kg) of twin born second

This is paired data, in which we have measurements on both of a pair of monozygotic twins, who have identical genes. We can therefore look for differences between first-born and

second-born twins by taking the differences between `head1` and `head2`, `area1` and `area2`, etc. This should give more precise results than if we had data on the first-born of one set of twins, and the second-born of a different set of twins, so that there was no pairing.

You should create a new data frame that contains the differences between the measurements for the first-born and the second born twin. You should look at this data (and perhaps the original data) and comment on whether outliers are apparent, and whether the data appears normally distributed. You should then use the `Tsq.test` and `Tsq.conf.int` functions that I have supplied, along with R's built-in `t.test` function, to test the null hypothesis that there is no difference in the means of any of these variables between first-born and second-born twins, and to find 90% confidence intervals for the differences in means. You should compute three kinds of confidence intervals: simultaneous confidence intervals based on the T^2 statistic, univariate confidence intervals based on the t statistic, and univariate confidence intervals with the Bonferroni correction.

You should write up your conclusions, addressing the question of whether the data provides evidence for a difference in the twins depending on birth order, and on what a researcher who has some reason to think there might be such a difference should do after seeing these results.