Question 2 (*Chapter 9 Examining the assumptions*)

(a) The data set **energy** in the SSC data subdirectory contains data on the average daily energy intake (in kilojoules, kJ) of eleven healthy women aged 22-30. It is required to find a 95% confidence interval for the mean daily energy intake in the population from which this sample of women was drawn.

(i) Load the data into your SSC worksheet and, assuming normality in the sample, obtain the required confidence interval. [2]

(ii) Now use a graphical method to check your assumption of normality in the data. Comment on the adequacy of a normal model in this case. [3]

Your individual perception of the 'shape' of the data may well differ in detail from that of your fellow students (and tutor). This does not matter. With only eleven data points in the sample, it is difficult to make a clear judgement of the normality or otherwise of the population involved. Nevertheless, some analysts might feel that a transformation of the data would help.

(iii) Show the SSC command to create a data vector **energy2** in which a logarithmic transformation (base e) is applied to these data. Use a graphical method to check the adequacy of a normal model to represent the variation in **energy2**. [3]

(iv) Find a 95% confidence interval for the mean of the transformed data, and retransform to obtain another confidence interval for the mean average daily energy intake in this population. Comment on how this interval compares with the interval you found in part (a)(i). [4]

One of the aims of the study that provided these data was to compare the mean energy intake of this population with the recommended daily intake of 7725 kJ.

(v) Do these data support the hypothesis that the mean energy intake of this population is below the recommended figure? Briefly explain why or why not. [2]

(b) The data set **injuries** in the SSC data subdirectory gives the number of accidents experienced by 414 different machinists over a given period of time. The data were collected as part of a famous study of 'accident proneness'.

(i) It is required to test the fit of a Poisson model to these data. As a quick precursor to a formal test, find the sample mean and sample variance for these data, and comment on whether these statistics provide support or otherwise for the adequacy of a Poisson model. [3]

(ii) Using SSC, obtain a vector of observed frequencies for these data. Assuming a Poisson model, obtain a vector of expected frequencies to compare with the observed frequencies. Have you any comments at this stage on the quality of the fit of the model? [4]

(iii) Perform a chi-squared test of goodness of fit of a Poisson model to these data, pooling adjacent counts where necessary to satisfy the requirements of this test. Report your conclusions carefully. [5]

(c) In a study of the effects of a drug, halofenate, on the levels of certain compounds in blood, a number of individuals had their level of serum triglyceride measured before and after a dose of either the drug or an inactive placebo. The differences between After and Before measurements are given in two vectors in the data set **trigly**, in the SSC data subdirectory.

(i) Load the data into your SSC worksheet. By producing appropriate plot(s), explain why it would be inappropriate to make an assumption of normality for these data. [3]

(ii) Briefly explain why a simple logarithmic or square root transformation would be inappropriate for these data. [2]