THE UNIVERSITY OF AUCKLAND **475.110SC**

EXAMINATION FOR BA BSc ETC 2000

STATISTICS

Visualising Information

(Time allowed: THREE hours)

Attempt all SIX questions. All questions are worth equal marks.

1. (a) Colour is characterized by three natural "dimensions" (one of which corresponds to the dominant light wavelength present in the colour). Name and describe these natural dimensions.

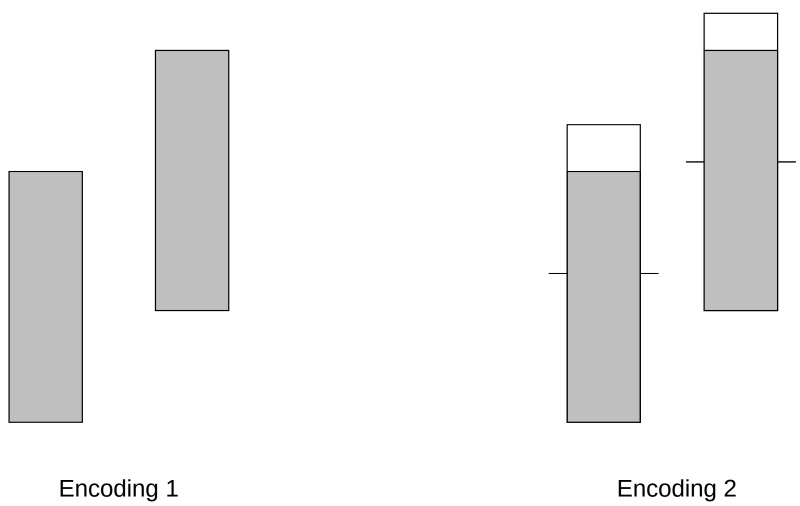
(b) The CIE chromaticity diagram provides a non-subjective description of colour. Explain how the chromaticity diagram is related to the natural dimensions of part (a).

(c) Definite what it means for colours to be *complementary*.

(d) Give four examples of how colour should *not* be used in statistical graphics. Explain your reasons for saying that this kind of usage is undesirable.

2. (a) Explain Weber's law.

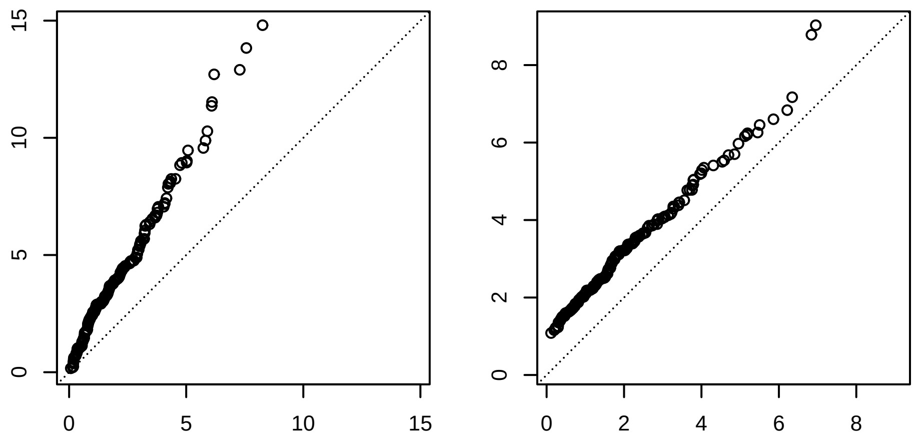
(b) The following graphs show a pair of numerical values encoded in two different ways. Use Weber's law to explain why the second of these encodings is more successful.



(c) The Cleveland-McGill scale ranks a number of methods of encoding numerical values graphically. Write down the ranking.

(d) Give one example of an encoding method *not* included in the Cleveland-McGill ordering. Explain (in detail) where you think this new attribute might fall within the Cleveland-McGill ranking.

3. (a) Each of the following two graphs show quantile-quantile plots comparing two sets of values. Sketch the corresponding pairs of (superimposed) density plots. Explain why you think the plots will appear the way you have drawn them.



(b) What data feature is directly represented in histograms and density traces? Comment on the relative strengths of these two plot types.

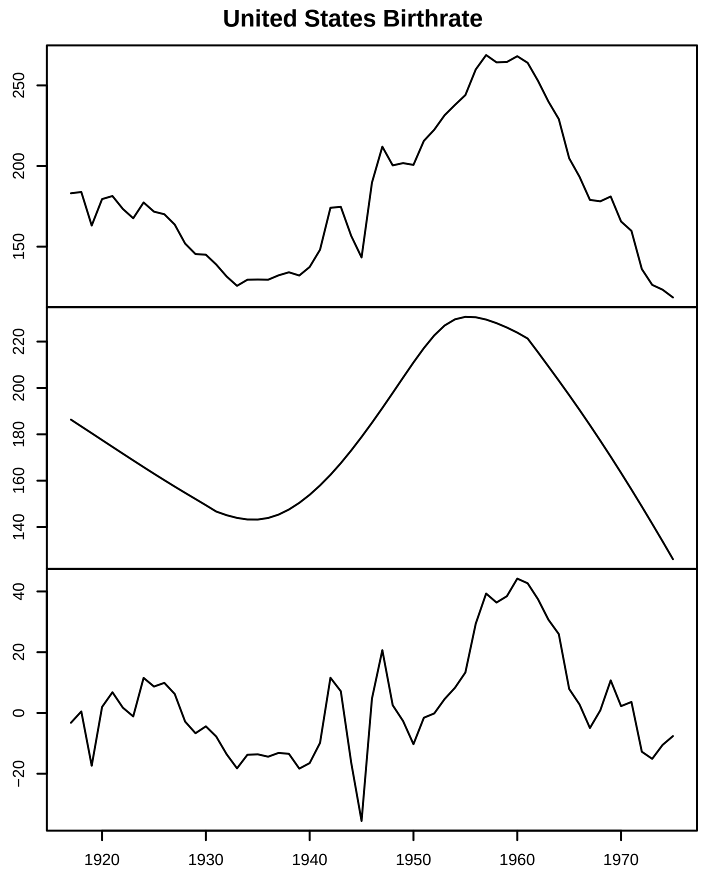
(c) Boxplots provide a simple representation of a set of numbers. Describe how to draw a boxplot.

(d) Are boxplots *good* plots? Justify your answers using ideas from graphical perception theory.

4. (a) When a data set contains three or more variables, a simple scatterplot is of limited use. Describe two techniques for displaying the structure of such multidimensional data sets. What are the relative strengths and weaknesses of these techniques?

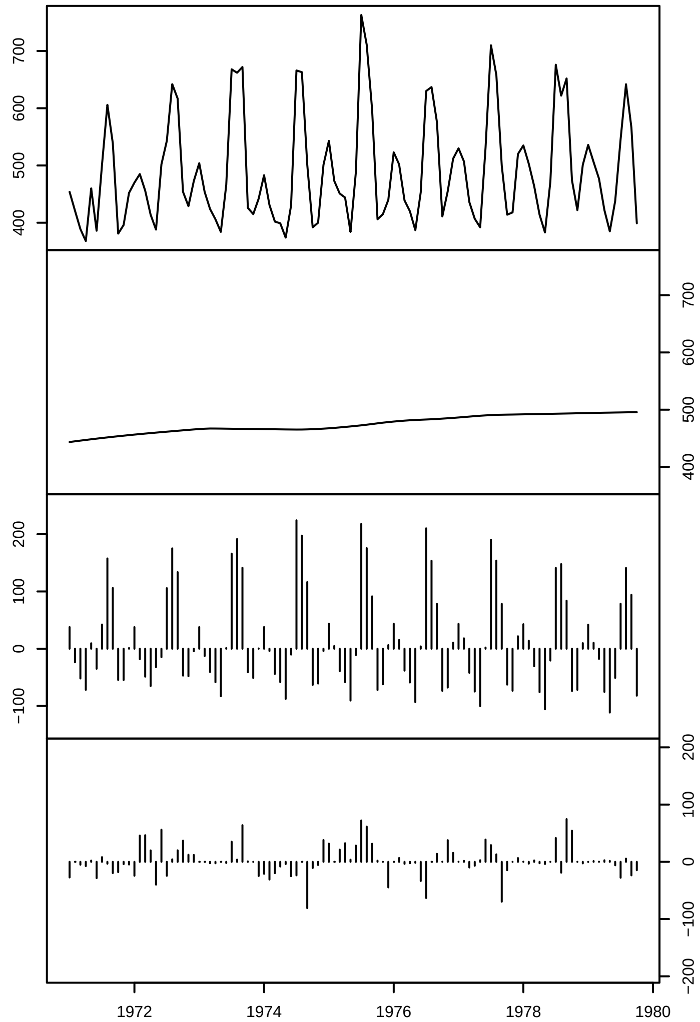
(b) In the special case that we are interested in how a one variable might depend functionally on several others, describe a technique which can be used to get some insight into the nature of the relationship.

5. (a) The following graph shows the times series of the numbers of births per 10,000 23-year-old women in the United States over the period 1917–1975, together with its decomposition into a smooth trend and residual series.



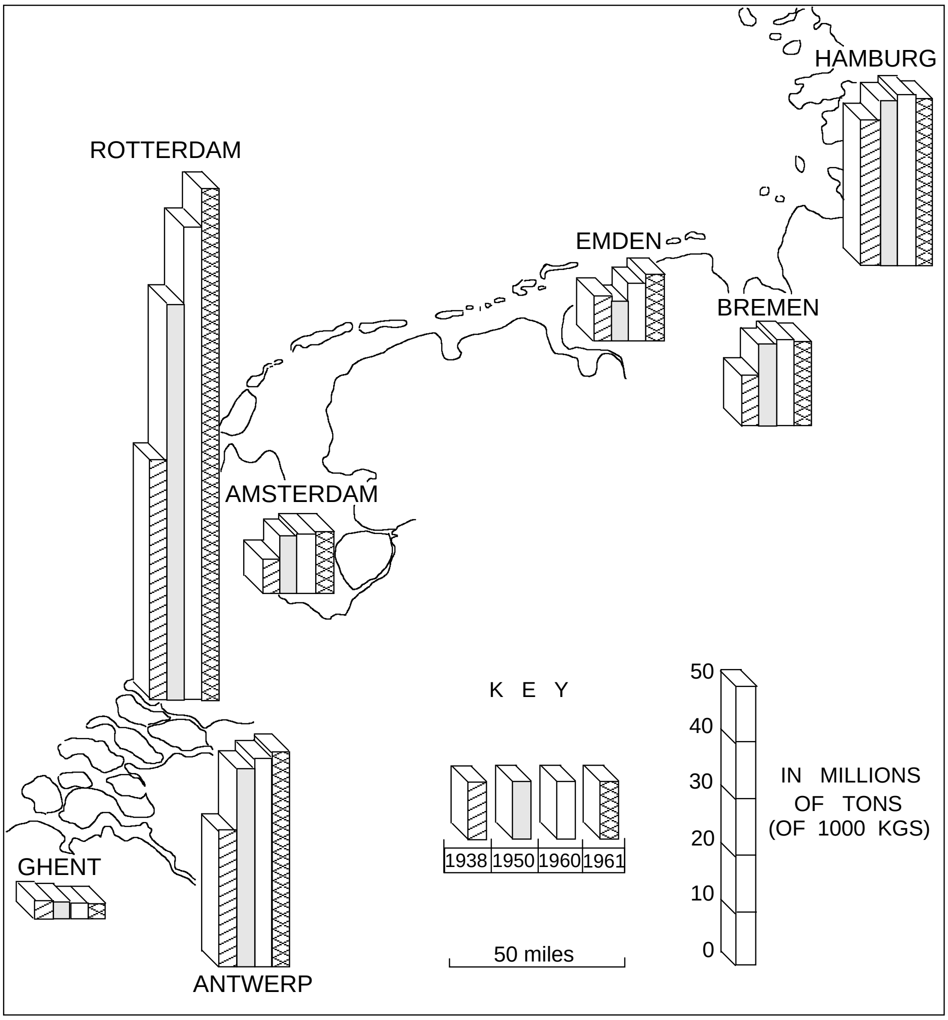
There are a variety of ways in which this graph could be improved. Explain what you see as the problems in the graph and indicate (as precisely as you can) how they could be fixed.

(b) The following graph shows a display of the monthly average residential electricity usage in Iowa City (in kilowatt-hours), January 1971 to October 1979.



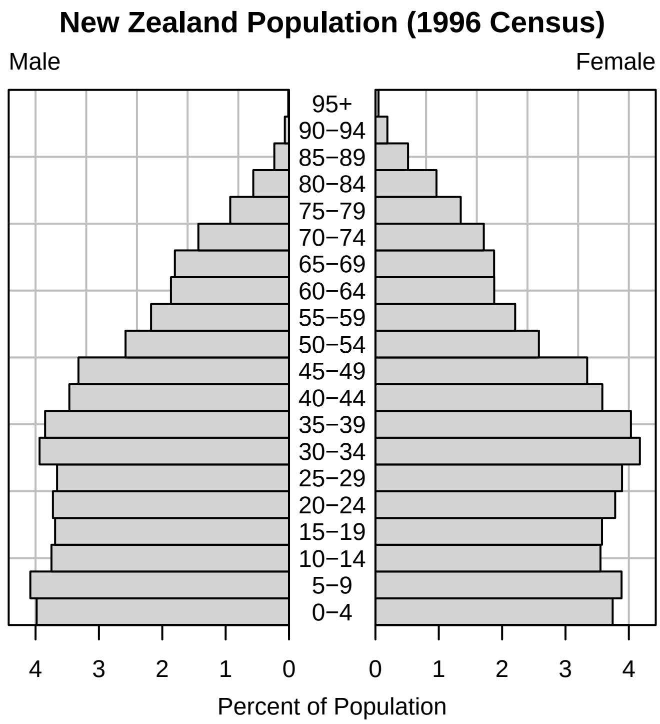
Explain what this graph is showing and indicate carefully how it should be labelled. (It may be helpful to use a sketch of the plot.)

6. (a) The following graph is presented in the book *Maps and Diagrams* by F. J. Monkhouse and H. R. Wilkinson. It shows the amount of shipping through a number of ports in northwestern Europe.



What are the good and bad features of this graph? If you see any problems in the graph, you should indicate how they could be overcome.

(b) The following graph shows a population pyramid for the New Zealand population (as computed in the 1996 census).



What elements (if any) of the Cleveland-McGill scale are being used to encode the data values in the graph? The clear intent of the graph is to make it possible to compare the male and female populations. Can you suggest an alternative way of graphing the data values to make this comparison easier?