

Exercise Sheet 1 – Plotting Data and Basic Statistical Measures

1. For the babyboom dataset included in the lecture notes
 - (a) Identify the types of the three variables in the data set.
 - (b) Construct a histogram of the birth times of all the babies.
 - (c) Imagine that the data were extended to include all births in the full week 15–21 December, 1997. How might you expect your histogram of birth times to look? Draw a sketch.
 - (d) Calculate the mean, median, SIQR, MAD and sample standard deviation of the birth weight of all the babies.
 - (e) Construct box plots of weights for boys and girls separately, and compare the two distributions. Comment.

2. The following dataset is from a study of rates of smoking by occupational group in England. For each occupational group a “smoking index” has been computed, which is the ratio of the average number of cigarettes smoked per day by men in this group, to the average number of cigarettes smoked by the whole population of men, times 100; and a mortality index, which is the ratio of the rate of lung cancer in that group to the rate in the population, times 100.

Occupational Group	Smoking	Mortality
Farmers, foresters, and fisherman	77	84
Miners and quarrymen	137	116
Gas, coke and chemical makers	117	123
Glass and ceramics makers	94	128
Furnace, forge, foundry, and rolling mill workers	116	155
Electrical and electronics workers	102	101
Engineering and allied trades	111	118
Woodworkers	93	113
Leather workers	88	104
Textile workers	102	88
Clothing workers	91	104
Food, drink, and tobacco workers	104	129
Paper and printing workers	107	86
Makers of other products	112	96
Construction workers	113	144
Painters and decorators	110	139
Drivers of stationary engines, cranes, etc.	125	113
Laborers not included elsewhere	133	146
Transport and communications workers	115	128
Warehousemen, storekeepers, packers, and bottlers	105	115
Clerical workers	87	79
Sales workers	91	85
Service, sport, and recreation workers	100	120
Administrators and managers	76	60
Professionals, technical workers, and artists	66	51

Which variable types are represented here?

Make a scatter plot of smoking index against mortality index, and comment. What do you think of the design of the study?

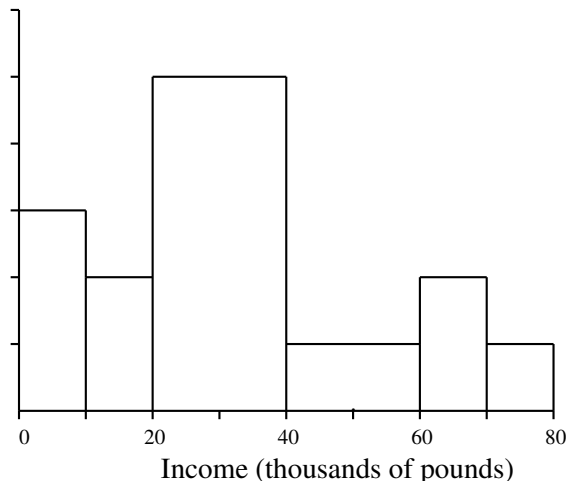
(The data are available at <http://lib.stat.cmu.edu/DASL/Datafiles/SmokingandCancer.html>. They were originally published in *Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970-1972* (1978), and appeared in *Introduction to the Practice of Statistics* by David S. Moore and George P. McCabe (1989).)

3. According to the 1976 Canadian census, the average number of children in francophone families was 1.85, while the average number in anglophone families was 1.95. N. Keyfitz (in Applied Mathematical Demography) made the following table:

Avg. # Children in:	French	English
Quebec	1.80	1.64
Other Provinces	2.14	1.97

In other words, the average number of children in francophone families in Quebec is higher than the average number of children in anglophone families in Quebec; and the average number of children in francophone families in the rest of Canada is higher than the average number of children in anglophone families in the rest of Canada. Does this contradict what we just said, that the average number of children in francophone families in the whole country is **lower** than the average number of children in anglophone families? Why or why not? Think about which set of individuals goes into making up each average.

4. The histogram below represents the distribution of family incomes in a hypothetical town with 20,000 people.



- (a) The vertical axis uses the density scale, but the numbers and units have been left off. Fill these in. What would the numbers and units be if the data were given in percentages?
- (b) What percent of the families earned between £10,000 and £20,000?
- (c) Is the mean income under £30,000, over £30,000, or equal to £30,000? Explain your reasoning, and say what assumptions you need to make.