Question 2 (*Chapter 2 Models for Data I*)

(a) The uniform distribution (see, say, Figure 2.5 on page 53 of the course text), the normal distribution (Figure 2.12 on page 63) and the triangular distribution (Figure 2.14 on page 65) are different theoretical models for the variation that may be exhibited in measurements on different random variables. In any given data context, one model is usually more appropriate than any other. *In non-technical language*, briefly describe the main characteristic features of the variation that these distributions might usefully model. [4]

In the remainder of the course, you will learn about many more models for variation, and by the end of the course you will have available to you a substantial number of different models to draw upon to express variation.

(b) Many sorts of children's sweets come in packets (bags or tubes) containing 40 or so bite-sized items ('lozenges', 'buttons', 'beans', 'pastilles', and so on). Often the only difference between the individual items in a packet is their colour, and there may be considerable variation here (e.g. red, orange, yellow, green, purple, pink). There is not usually any reason to suppose that one colour should occur more frequently than any other in a packet.

(i) Use the command `dice()` in SSC to simulate the actual colours observed in one tube of 40 coloured chocolate beans, where each bean may be one of eight different colours with equal chance. Making the identification 1 = *red*, 2 = *orange*, 3 = *yellow*, 4 = *green*, 5 = *blue*, 6 = *violet*, 7 = *pink*, 8 = *brown* show how you would use SSC to count the number of yellow beans in your simulated tube. (So what is required here is a vector of length 40 containing the digits $1, 2, \ldots, 8$ at random, to represent the contents of the tube.) [4]

(ii) In any tube of 40 beans, it would be remarkable if each of the eight different colours of bean occurred exactly five times, though the assumption of uniformity suggests this. (Actually it can be shown that this remarkable event would happen only once in every 70 000 tubes or so.)

In fact, the command sequence in part (b)(i), if repeated many times, would permit you to get some idea of the random variation in the colours that one might reasonably expect to see. Alternatively, there is a useful command `rmn()` that enables the colours in many tubes to be tallied. First create the vector of probabilities

```
vP= (1,1,1,1,1,1,1,1)/8
```

which gives the probabilities for each colour (in this case, 1/8 each); and then use

```
mm= rmn(100,40,vP)
```

to create a 100×8 matrix called `mm`. The 100 rows represent the contents of 100 different tubes of chocolate beans; the columns represent the frequencies of each colour.

It may even be that in some tubes a colour is not represented at all. Use your results from `rmn()` to estimate the proportion of occasions when at least one colour is absent in a pack. [4]

(c) The data set `spikes` in the SSC data subdirectory lists 100 waiting times (measured in milliseconds) between consecutive spikes of neural activity in a monkey at rest. These data were collected as part of an investigation into what level of activity would constitute evidence of neurological disturbance.

Assume that it is required to use the triangular model as a theoretical representation of the variation that might be observed in such waiting times.

(i) Obtain a histogram of the data set and comment in general terms on the usefulness of the triangular model. (*Not more than a couple of sentences*