MA 2611 A' 01 Solutions, Test 2, Part 1

1. Could something about teaching-perhaps the years of contact with kids and their germs-increase the risk of immune system diseases such as multiple sclerosis and lupus? A study published recently in the Journal of Rheumatology suggests that the rates of such diseases in schoolteachers is substantially higher than it is for people in other professions.

   Below are three alternative designs for this study. Tell what kind of design each is (be as specific as possible). Be sure to justify your answers.

   (a) **(10 points)** Investigators choose a random sample of schoolteachers and another of individuals who are not schoolteachers. They follow both groups over a 20 year period and compare the proportion of schoolteachers who contract immune system diseases during that time with the corresponding proportion of the other group.

   ANS: *This is a prospective observational study, since the groups (schoolteachers or not) are set up at the outset and the outcomes (immune system disease or not) in each group obtained subsequently.*

   (b) **(10 points)** Investigators divide a sample of death certificates into two groups: those which list an immune system disease as a cause of death, and those which do not. They then compare the proportions of schoolteachers in each group.

   ANS: *This is a retrospective observational study, since the groups are based on outcomes (immune system disease or not), and the causes for the outcomes are sought (schoolteacher or not).*

   (c) **(10 points)** Investigators interview a random sample of schoolteachers and a random sample of non-schoolteachers to determine how many have an immune system disease.

   ANS: *This is a sample survey, since they are taking samples from populations to summarize an aspect of the populations.*

2. Figure 2 displays a box and whiskers plot for a set of data.

   (a) **(10 points)** Which of the following data sets could NOT have produced the plot in Figure 2 (For each one you conclude could not have produced the plot, give a reason):

   i. 1 2 4 5 8 9 11 29 33 37 39 40 41 44 47
   ii. 1 2 4 5 8 9 11 29 33 37 39 40 41 44 45
   iii. 1 2 4 5 8 9 11 29 33 37 39 40 41 43 44 45
   iv. 1 2 4 6 8 9 11 29 33 37 39 40 41 43 45
   v. 1 3 4 5 8 9 11 29 33 37 39 40 41 43 45

   ANS: *i. could not have produced the plot because 47 would appear as an outlier. iii. could not have produced the plot because its median is 31, not 29 as shown on the plot. iv. could not have produced the plot because its Q_1 is 6, not 5 as shown on the plot.*

   (b) **(10 points)** Is the box and whiskers plot in Figure 2 a good graphical summary for those data sets from part (a) that could have produced this plot? Justify your answer.

   ANS: *No, it is not a good graphical summary for either ii. or v. since the main feature of these data is that there are two distinct clusters: 1-11 and 29-45, and the box and whisker plot disguises this feature.*

3. Diameter specifications for mylar sheaths produced by the Ebco Company are 15 ± 2 mm. Measurements taken from a large number of production units suggest the diameters Y of the population of sheaths follow a normal distribution with mean 15.1 and standard deviation 0.9.

   (a) **(10 points)** What is the population proportion of sheaths that fall within spec?
   ANS: *If Y is the diameter of a randomly chosen sheath,*

$$P(13 < Y < 17) = P(\frac{13 - 15.1}{0.9} < Z < \frac{13 - 15.1}{0.9}) = P(-2.33 < Z < 2.11) = 0.9727$$

   (b) **(10 points)** The value of the $N(15.1, 0.9^2)$ density curve at 16.2 is one half its value at 15.1. Relate this fact to the relative likelihood of finding a sheath with a diameter close to 15.1 compared with finding one with diameter close to 16.2.
   ANS: *It is twice as likely to find one with a diameter close to 15.1 as it is to find one with diameter close to 16.2.*

4. Newly-picked oranges are sorted into four classifications, which we will label 1 (highest) through 4 (lowest). Let the random variable Y denote the classification of a randomly selected orange. Experience shows that Y has probability distribution $p_Y(y) = c/y, \ y = 1, 2, 3, 4$.

   (a) **(10 points)** Find the value of c.
   ANS: $1 = c + c/2 + c/3 + c/4 = 25c/12$. *Therefore,* $c = 12/25 = 0.48$.

   (b) **(10 points)** Sketch the population histogram of this random variable.
   ANS: *See Figure 1.*

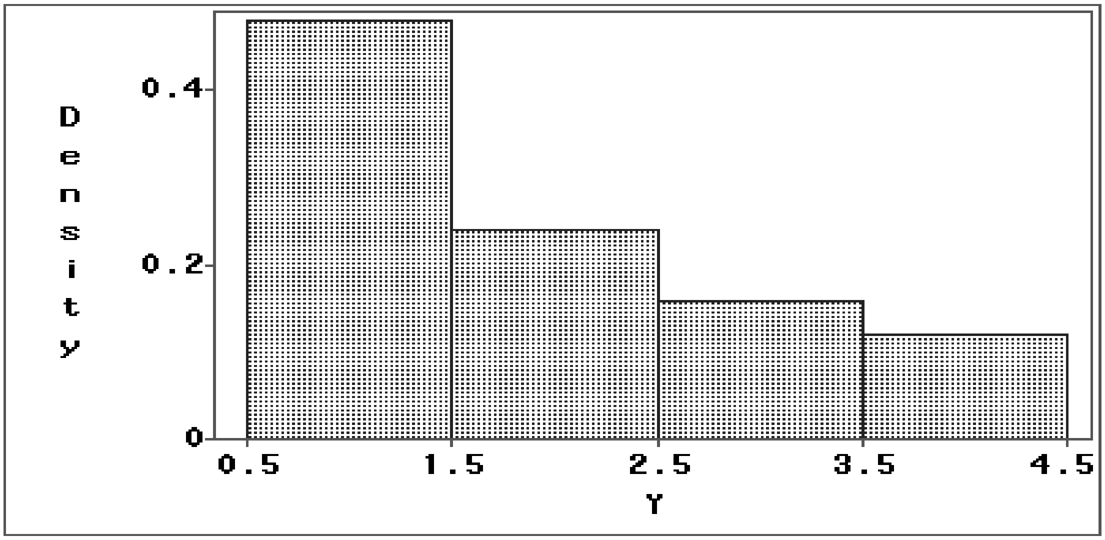

Figure 1: *Population histogram for question 4.*

   (c) **(10 points)** Use the information from (a) to estimate how many of the next 10,000 oranges harvested will be given one of the top two classifications.
   ANS: $10000 \times (0.48 + 0.24) = 7200$

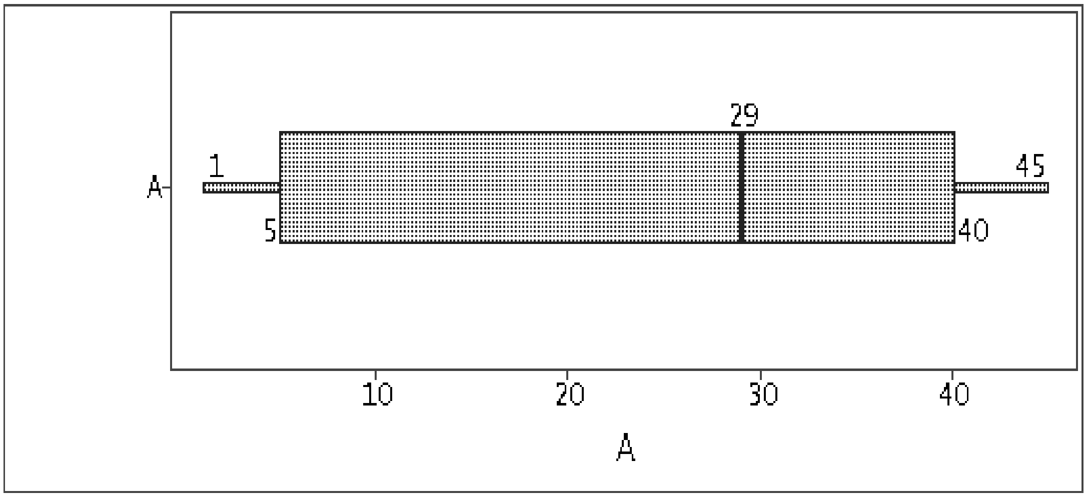

Figure 2: *Box and whiskers plot for question 2.*