THE ROYAL STATISTICAL SOCIETY

2008 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

(MODULAR FORMAT)

MODULE 1

DATA COLLECTION AND INTERPRETATION

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Module 1, 2008. Question 1

(**Solution continues on next page**)

(i) Every survey should be aimed at a target population of individuals or organisations or groups whose characteristics are being studied. The sampling frame should be a complete list of items in the target population. Sampling frames can be imperfect in several ways, such as being out of date, having duplication or omissions or other inaccuracies, or not matching the target population. These are inter-related to some extent.

Every sampling frame is out of date as soon as it is published. The longer the time since a frame was drawn up, the greater the discrepancy between it and the target population. An out of date list is likely to include elements that no longer exist and exclude elements that have joined the population since the list was compiled; for example, in a list of buildings some might have been demolished and some new ones might have been built in the intervening period. Duplication occurs when the same element is listed more than once; for example a mail order firm's list of customers might include the same customer with first name and surname and also with initials and surname. Omissions occur when an element that should have been in the frame at the time it was drawn up is not listed; for example some entries might be missed when transferring clerical records to electronic form. Inaccuracies occur when details about elements are incorrect; for example a school in a sampling frame of schools might incorrectly state that it had a sixth form. A frame might also fail to match the target population because the elements do not match; for example the electoral register for an area is not a frame of adults as it excludes prisoners and a few other categories of people.

(ii) (a) Neither frame is a sample frame of households with school age children, so neither matches the target population. Similarly both frames could be out of date unless drawn up very recently.

There is a fairly good relationship between addresses and households, in that once non-private addresses have been eliminated from a list of addresses there will mostly be one household per address. However, a sample frame of addresses could contain addresses which no longer exist, or omit addresses corresponding to new buildings or new subdivisions of buildings into separate units. There could be duplication if a building which was formerly subdivided, giving rise to more than one address, is no longer subdivided. Some addresses will obviously be non-private, for example they might include the name of a shop; local knowledge may indicate whether there is also a residential space (maybe above or behind it), and may also be useful in identifying other such residential addresses. A sample of addresses could be drawn from those remaining on the list after an initial cleaning up and pruning.

When the addresses are surveyed, questions will need to be asked as to how many households live at the address and whether the households contain children of school age. If there is more than one household with school-age

children then one of these might be chosen according to a previously assigned random number to ensure that every household has the same chance of being selected.

Selecting a sample of children from a sample frame of school children could be adapted to produce a sample of households with school age children by obtaining the addresses where the children lived most of the time, and the name of a responsible adult at the address. Not all children live with parents and some children spend part of the week with one family and part with another. Others live in institutions.

A sample frame of school children would lead to duplication if it included two or more children from the same household. It would lead to omissions of households which had moved into the area, and it could include households which had moved away and those where the children were no longer of school age.

(ii) (b) Either choice is acceptable if reasonable justification is given. Using a sampling frame of addresses is likely to lead to more wastage in that it will include addresses with no households in them and many households with no children of school age. Most children selected from a sample frame of school children will be in households with school age children, but a possible difficulty with this frame is obtaining the addresses of the households.

Higher Certificate, Module 1, 2008. Question 2

(i) Cluster sampling would be an obvious choice of sampling method for the first stage of sampling. As lists of hospitals are available for each region, either a simple random sample of regions could be taken at the first stage, or a sample of regions with probabilities proportional to the number of hospitals in the region. At the second stage, hospitals in the selected regions could be stratified into teaching and non-teaching hospitals and a simple random sample of hospitals taken from each stratum.

The advantage of cluster sampling is that, if members of the research organisation need to travel to the hospitals or talk face to face with patients in connection with the survey, there will be less travel involved if selected hospitals are in a few regions only. A possible disadvantage is that if the patients in the different regions vary a great deal, as regards their perceptions of post-operative care and other characteristics, the selected regions are not necessarily going to represent all of this variability. For cluster sampling to work well, clusters need to be similar to one another so that any one cluster is representative of the whole. Another disadvantage is that estimators under cluster sampling tend to have high variability.

It is likely that teaching hospitals and non-teaching hospitals vary, both as regards the conditions for which patients are admitted and in the amount of time that they are able to spend interacting with patients. Patients' perceptions of post-operative care are therefore likely to reflect these differences. Having teaching and non-teaching hospitals as strata ensures that both types of hospital are in the sample. In addition it would be easy to obtain estimates for each stratum alone as well for all hospitals combined. Stratified sampling works well when elements within each stratum are like one another but elements in the different strata are dissimilar. A disadvantage of stratified sampling is that elements have to be sorted into strata in order to obtain the separate estimates.

[*Other reasonable two-stage schemes were acceptable in candidates' answers.*]

(ii) The research organisation would need to know the way in which records are kept – for example in electronic form, on paper, on record cards, etc. The organisation would also need to know the order in which records are kept (alphabetical, geographical, date of admission, ward in which treated, etc) and whether it is possible to sort out those who have had operations and those who left hospital in any specified week before taking a sample. They would also need to know the approximate numbers of patients in groups – in the group of interest, and, if this group cannot be sampled directly, in the larger group or groups containing the patients of interest.

Solution continued on next page

(iii)

1. Did the nursing staff check whether you were in pain when you came round from the operation?

   Yes □ No □ Cannot remember □

2. Do you think that the nursing staff paid you sufficient attention in the 24 hours after your return to the ward?

   Yes □ No □ Don't know □

3. If you were in pain at any time, were you offered pain relief medicine?

   Had no pain □

   Was given pain relief whenever I was in pain □

   Was given pain relief some of the time I was in pain □

   Was not ever given pain relief when I was in pain □

4. Were you visited by a doctor in the 24 hours after your return to the ward?

   Yes □ No □ Don't know □

5. Did a doctor agree that you were fit to go home before you left the hospital?

   Yes □ No □ Don't know □

6. Were you given any information or advice concerning after-effects of the operation when you left the hospital?

   I was given both information and advice □

   I was given information but not advice □

   I was given advice but not information □

   I was given no information or advice □

Higher Certificate, Module 1, 2008. Question 3

(i) The increase in the number of women in the population from 1996/7 to 2003/4 was 22913 – 22289 = 624, so the percentage increase in the population of women was 100×624/22289 = 2.8.

The corresponding increase in the number of men was 21399 – 20591 = 808, so the percentage increase in the population of men was 100×808/20591 = 3.9.

In both years there were more women than men.

The distributions of incomes of women and men are different from one another in each year, but the patterns of the distributions for the same sex for each year are very similar. In the case of women in 1996/7, 28% had incomes in the bottom quintile group as defined by the distribution for all adults in that year, and 53% had incomes in the bottom two quintile groups compared with 40% of all adults and 25% of men. In contrast 56% of men had incomes in the two highest income groups in 1996/7, compared with 25% of women. The two distributions are mirror images of one another. This shows very clearly that the total incomes of women were concentrated at the lower end of the income distribution. The percentages in the 3rd (the middle) quintile group were similar for both sexes, and close to the theoretical 20% in this group.

The picture is very similar in 2003/4 with 52% of women having incomes in the lowest two quintile groups compared with 27% of men, and 54% of men having incomes in the highest two groups compared with 27% of women. Thus there was virtually no change between the two years to which the tables relate.

The mean total weekly individual incomes for both women and men increased over the period. That for women increased by £50 per week and that for men by £61 per week in absolute terms, that is a percentage increase of 28.2 for women, and 17.6 for men. It is appropriate to compare the means in this way as the means for both years are quoted in terms of 2003/4 prices. Men's mean weekly total income was 1.96 times that of women's in the earlier year, and 1.80 times women's in the later year. It appears that women are catching up with men. However, a mean can hide the actual picture. It could be that women who had the highest total incomes in the earlier year had a greater increase than women who had the lowest incomes in the earlier year, that is there could be inequalities as regards the changes in the distributions of incomes between the two years.

Solution continued on next page

(ii) In general people are reluctant to give details of their incomes, perhaps fearing that doing so would mean they have to pay more tax. Income in itself is complicated as it might include, for example, benefits as well as salaries, and income from investments. Different sources of income are received at different intervals of time, for example salary might be paid monthly but interest on a building society account might be paid annually. The time period for which income is required needs to be specified. Income can be quoted as gross (before tax) or net (after tax). Tax adjustments are sometimes made in arrears. Designing a questionnaire to obtain accurate information about incomes is therefore complicated, whether it is to be completed with the help of an interviewer or as self-completion. A detailed sheet of instructions will be needed, especially for the case of self-completion. Finding all the information needed is time-consuming for respondents. The response rate may be rather low.

Higher Certificate, Module 1, 2008. Question 4

Part (i)

(a) The mean is the total of all the observations divided by the total number of observations. It is a "sharing out" average. As there were 97 boys and the mean time they spent watching TV was 1.799 hours, in total they must have spent 174.5 hours watching. 1.799 hours is the number each would have watched if they had all watched for exactly the same amount of time, found as 174.5 divided by 97. The calculation is similar for the girls, whose mean watching time was 1.473 hours.

The median is a "half-way" measure in the sense that half of the observations are less than or equal to it and half are greater than or equal to it. It is the middle observation when an odd number of observations are ranked in order of magnitude, and is taken as half-way between the two middle observations when an even number of observations are ranked in order of magnitude. For both the boys and the girls the median number of hours, as given in the table, was 1.25. There were 97 boys so the middle observation is the 49th. It is possible to count the dots in the dot-plot; we can see that the 49th observation is one of the 6 at 1.25 hours. There were 92 girls, so we take the median as half-way between the 46th and 47th observations. Counting dots on the dot-plot confirms that both of these observations are among the 11 at 1.25 hours.

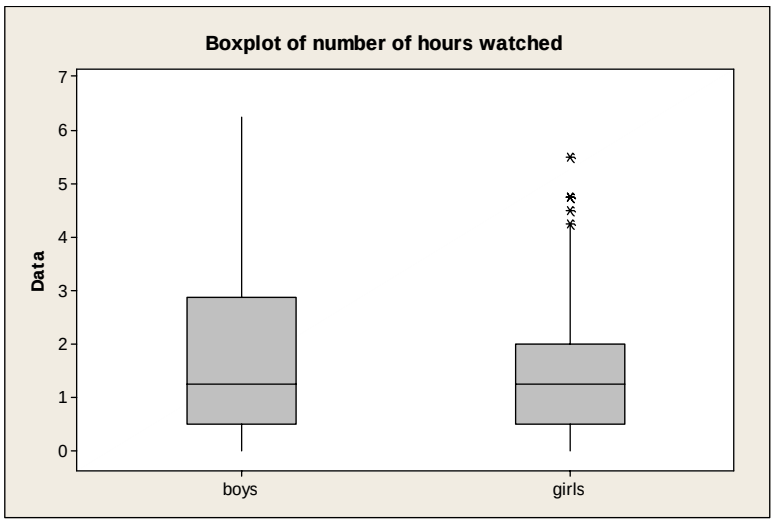
The mode is the value of highest frequency. The mode for boys is 0 hours which has a frequency of 15. More boys did not watch at all than the number who watched any specified number of hours. There are three modes for girls, each with a frequency of 11, at 0, 1.25 and 1.5 hours. More girls watched these numbers of hours than any other number of hours.

A complicating factor in the data is that the answers are given to the nearest quarter-hour. This is immediately apparent from the dot-plots, which show that the observations take only discrete values at the quarter-hours. While it was no doubt intended that a response of, say, "half an hour" was meant to mean anything between 22½ and 37½ minutes, it is likely that some respondents may have interpreted it as from 30 to 45 (exclusive) minutes while others may have interpreted it as 15 (exclusive) to 30 minutes. Further, the "0" entry cannot possibly mean a negative number of hours; presumably it was meant to mean anything from 0 to 7½ minutes, but we cannot be certain that every respondent made this interpretation. It is difficult to know how the calculations might be adjusted to try to take this into account. It may be better simply to take them at their face value of exact quarter-hours, which is what has been done in the dot-plots and in the computer calculations leading to the table.

Solution continued on next page

(b) As discussed above, the responses to the variable concerned are restricted to discrete values. The ranges of the values are relatively small (0 to 6.25 for boys and 0 to 5.5 for girls). The dotplots clearly show the individual values, including the tails of the distributions. They also show that quite a few boys and girls had not watched TV after school on the day for which they responded to the question. Histograms would not bring out these details. In particular it is likely that a histogram would group together values greater than 3 or thereabouts, making it difficult to discern the maximum value.

(c)



[*Note. This diagram was produced by Minitab. In the examination, candidates were not necessarily expected to show the outliers for the girls by separate * characters, though not recognising them as outliers might prejudice the discussion below.*]

The box and whisker plots show that both distributions are of positive skewness, i.e. they have longish tails of high values. This is more so in the case of the distribution for boys than that for girls. The distribution for girls is more tight in the sense that the middle 50% between the first and third quartiles falls in a smaller range than the middle 50% of the boys' responses. In the distribution for girls, there are four values that are unusually high compared with the bulk of the observations. These count as "outliers". For the boys, there are several high values but they are well scattered and do not appear as outliers in the diagram as shown above [produced by Minitab].

Solution continued on next page

Part (ii)

As use of drugs is dangerous, and drugs are expensive and difficult to obtain legally, children are not necessarily going to answer truthfully to questions about their use of drugs. It will be necessary to get their confidence and to assure them that their answers will be confidential and will not be revealed to anyone outside the survey organisation. They should also be told that it will not be possible to identify individuals in any reports of the survey. It might be sensible, when getting the permission of the school and parents/guardians to run the survey, to make clear that questions on drug use are to be asked. The survey and questionnaire might need approval by an ethics committee. Talking to classes of children about the survey could help persuade children to give honest answers. A plan as to what action to take if a child is suspected of drug abuse as a result of his or her survey responses should be made in advance of the survey. Randomised response techniques might be considered.