

THE ROYAL STATISTICAL SOCIETY

2010 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

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, 2010. Question 1

[Solution continues on next page]

Part (a)

- Q1. The question asks for a year but "boxes" are also given for day and month. One obvious way to correct this is simply to remove all the boxes and allow respondents merely to write in the year. Alternatively, the question could be reworded as "Please state the date on which you left full-time education"; or, as few people are likely to remember the exact date, "Please state the month and year when you left full-time education", giving boxes for month and year only.

Another point is that larger boxes are needed, as numbers have to be written in them. This is less of an issue for the tick-boxes in Q2 and Q5 as it will not really matter if a tick is larger than its box.

- Q2. There are many other possible categories, including unemployed, home-maker (or similar, but care should be taken that this title will be generally understood) and retired. An "other" category should certainly be included too.

The definition of "employed part-time" is incomplete (is it < 40 hours per week, per month, or what?). Further, as it is likely to be *thought* to mean per week, it is misleading vis-à-vis "full-time", as many full-time employees work, at least formally even if not in practice, less than 40 hours per week. It is probably better not to give any definition, as people are likely to know whether they are full-time or part-time.

There may also be a problem with people who (legitimately) have more than one job. If this is thought to matter for the survey, a solution might be to add "for your sole or main employment" to the request to state employment status.

- Q3. This is a leading question due to the use of "really" which gives a strong suggestion that the job is not enjoyed. Deleting this word improves the question, but it remains leading as there is then a suggestion that the job is enjoyed. A preferable wording would be simply "Do you enjoy your job or not?"

- Q4. This does not make clear to what period of time it relates or whether it is a basic income. A better wording is "What is your annual gross income excluding any additional benefits?"

A further point is that many respondents may genuinely not know it accurately without checking records, and many will be unwilling to give it accurately even if they do know it. A set of "boxes" to capture ranges of income is likely to work better.

The point about sole or main employment (see Q2) may arise here too.

- Q5. There is an overlap here in that 10 employees appears both in "10 or fewer" and in "10-50". This can be avoided by making the first category fewer than 10. However, the categories in themselves are suitable only for those working in places where the number of employees is relatively small. An open-ended question asking for the approximate number of employees might work better.

A separate problem is that it is not clear whether "where you are employed" refers to an entire organisation (which may be a worldwide multinational!) or to some division of it such as a Department. This needs to be clarified, but the nature of the clarification will depend on what the survey intends to measure. A phrase such as "(This refers to the unit, department or division in which you work rather than the entire organisation)" *might* be appropriate. It might be better to provide this clarification via a footnote rather than a lengthy addition to the question itself.

Editorially, more space should be left after each of the first two boxes so that it is immediately clear that a box refers to the category on its left. Q2 is set out better in this regard.

- Q6. The direction of scoring is not stated so there would be no indication of whether a low score is good or bad. This cannot be assumed to be obvious. The question could be amended by adding "where 0 denotes no satisfaction and 10 denotes perfectly satisfied".

Part (b)

- (i) Sensitive questions are likely to be troublesome as respondents might feel embarrassed by them and might not want to give honest answers – or, indeed, to answer them at all. This may extend to a decision not to answer any part of the questionnaire, even those parts that are not sensitive. Topics that are often found to be sensitive are those to do with income, with sexual habits and attitudes, and with activities that are either illegal or widely frowned on without necessarily being illegal (eg use of "recreational" drugs).
- (ii) Hypothetical questions are of limited worth as respondents cannot be sure what they would *actually* do in a particular situation that has not occurred. Examples of such questions include "What would you do if you won a million pounds?" and "How would you react if your house was burnt down?"

Higher Certificate, Module 1, 2010. Question 2

[Solution continues on next page]

The gains in height in cm of daughters of short mothers are 19.5, 21.4, 22.3, 20.0, 20.5 and 23.2.

The gains for daughters of tall mothers are 20.3, 26.3, 25.1, 25.3, 23.0 and 25.2.

The sample means and standard deviations (divisor $n - 1$) are as follows (these can be taken routinely from a calculator, or obtained by standard formulae via Σx and Σx^2).

Daughters of short mothers, age 6: mean 112.53, standard deviation 1.81.

Daughters of short mothers, age 10: mean 133.68, standard deviation 2.23.

Daughters of short mothers, gain in height: mean 21.15, standard deviation 1.42.

Daughters of tall mothers, age 6: mean 119.52, standard deviation 1.91.

Daughters of tall mothers, age 10: mean 143.72, standard deviation 2.41.

Daughters of tall mothers, gain in height: mean 24.20, standard deviation 2.19.

All measurements in this report are in cm. The report is based on data from two small samples of girls, some with short mothers and some with tall mothers; each sample was of size only 6. The girls' heights were measured at ages 6 and 10.

Perhaps the most noticeable feature of the data is that all the girls with short mothers were shorter than all the girls with tall mothers both at age 6 and at age 10.

For girls with short mothers, the range of heights at age 6 was 110 to 114.5, a range of 4.5. At age 10, it was 130.5 to 136, a range of 5.5. The range of increases in heights was 19.5 to 23.2, a range of 3.7. The average height at age 6 was 112.53 and the standard deviation of heights was 1.81. At age 10, the average was 133.68 and the standard deviation was 2.23. The average increase in height was 21.15 and the standard deviation was 1.42.

Unsurprisingly, heights were all greater at age 10 than at age 6, by around 20 cm. There was no overlap of the height ranges at the two ages. In addition, for these girls there was more variability in heights at age 10 than at age 6, as shown both by the increased range and by the standard deviations. At age 10 there was a noticeable gap between the heights of the two shortest girls (130.5 and 131.4) and those of the others. Such a gap was not apparent at age 6. The two shortest girls at age 10 had also been

shortest at age 6, but overall the rank orderings of heights were not consistent over the two ages.

For girls with tall mothers, the range of heights at age 6 was 115.9 to 121.0, a range of 5.1 – a slightly wider range than for the girls with short mothers. At age 10, it was 140.7 to 146.5, a range of 5.8 – again slightly wider than for girls with short mothers, but only a small difference in this case. The range of increases in heights was 20.3 to 26.3, a range of 6.0 – considerably more than for the girls with short mothers. The average height at age 6 was 119.52 and the standard deviation of heights was 1.91. At age 10, the average was 143.72 and the standard deviation was 2.41. The average increase in height was 24.20 and the standard deviation was 2.19.

Again, heights were unsurprisingly all greater at age 10 than at age 6. The increases tended to be more, and were more variable, than for the girls with short mothers. There was no overlap of the height ranges at the two ages. For these girls there was again more variability in heights at age 10 than at age 6, as shown both by the increased range and by the standard deviations. It was again the case that two girls at age 10 were noticeably shorter than the rest; and it also looked as though two girls at age 10 were beginning to pull away from the rest in growing taller. There had also been something of a gap between the shortest girl and the rest at age 6. This shortest girl had not been much taller than the girls with short mothers, and remained among the shorter girls of tall mothers at age 10. Again the overall rank orderings of heights were not consistent over the two ages.

Higher Certificate, Module 2, 2010. Question 3

- (i) In stratified sampling, the population is divided into homogeneous groups called strata which, it is thought, might show consistent differences from each other in respect of the characteristics being investigated in the survey. A common example in human populations is to divide into males and females. A sample, usually (simple) random, is taken from each stratum. The advantages of this method are that variances of estimators for the entire population are likely to be smaller than those obtained using other methods of sampling, and also that every group is represented and information is available about each group separately as well as about the population as a whole. A disadvantage is that more work is commonly required in setting up the sampling scheme, as it is (usually) necessary to know which stratum each element of the population is in before the sample is taken.

In cluster sampling, the population is divided into heterogeneous groups called clusters which, it is hoped, are themselves representative of the entire population – in particular, that they capture the *variation* in the entire population. A random sample of clusters is then taken to form the sample. A common example in educational research is to divide a country into Local Education Authority (or equivalent) areas which, it is hoped, replicate the characteristics of the entire country. There might be further stages of sampling within the clusters themselves (though it is quite common for only a few clusters, perhaps only one, to be selected and then complete enumeration carried out within the chosen cluster(s)). The advantages of this method are that details of the elements to be sampled are needed only for the selected clusters, and costs are generally less than with other methods, especially if face-to-face interviews are used. A principal disadvantage of this method is that variances of estimators are usually greater than with other methods.

In quota sampling, the population is notionally divided into strata and interviewers are asked to find samples from each stratum. It is thus a non-random method of sampling and analysis of the results cannot proceed in a straightforward way using the usual probability arguments. Further, there is no guarantee that elements will be correctly identified in respect of which stratum they belong to. Whereas in stratified sampling the stratum membership is known before selection, in quota sampling membership is identified by sight or by asking questions as part of the interview. Advantages are that the target sample sizes can be achieved as interviewing can continue until quotas are filled, and (often) the speed with which the survey can be completed. Also, in practice quota sampling is often the only realistic way of carrying out a survey, especially if the strata are not particularly well defined, especially in respect of their sizes, in advance. Disadvantages are the reliance on interviewers in filling the quotas correctly, a risk of interviewer bias, and frequently that there are problems in filling quotas (especially in rarer groups) as the survey draws to a close.

Solution continued on next page

- (ii) The regions of the country can be considered to be clusters. The first stage (cluster sampling) is to select one or more of these, presumably on the basis of simple random sampling. In the selected regions, the urban and rural areas can be considered to constitute two strata. The second stage (stratified sampling) is to select, again presumably at random, a sample of areas from each of these strata. In the last stage of sampling (quote sampling), carried out within the selected urban and rural areas, quotas of say males and females in different age groups might be defined, and interviewers asked to obtain samples of specified size according to the quotas set.
- (iii) A systematic sample of "about 50" from 506 is asked for. $506/50 = 10.12$, so we use an interval of 10. Choose a starting point at random in the first 10 names listed and then select every 10th name after that.

If the starting point is in the first 6 names in the list, a sample of size 51 will result. Otherwise, the sample will be of size 50.

Systematic sampling is simple to describe and easy to carry out. It has intuitive appeal. Users of the survey may readily be satisfied that this is a good method of sampling.

However, while a systematic sample may have the appearance of being random, and may in practice behave as though it is random, in fact it is not. Only the starting point is chosen at random; once it has been chosen, all the other members of the sample are fixed. The familiar theoretical results of simple random sampling do not strictly apply.

Further, if there is some cyclic pattern (possibly unknown) in the way the list has been drawn up, the members chosen may coincide with it. This would happen here if, for some reason, every 10th person in the list is similar in some way; the selected sample would then be too homogeneous and would not capture all the variability in the population.

Theoretical results for systematic sampling are derived by considering it as a form of cluster sampling. Using the present case as an example, the population is divided into ten clusters: the first cluster consists of inhabitants numbered 1, 11, 21, ... in the list, the second cluster consists of inhabitants 2, 22, 32, ..., and so on. One of these clusters is then chosen (at random) to be the sample.

Higher Certificate, Module 2, 2010. Question 4

- (i) Those who refuse to answer questions and those who are not contacted might be very different as regards characteristics and attitudes from those who take part in the survey.

As an example, people may refuse because they are short of time. But busy people are likely to have different characteristics and attitudes compared with people with spare time on their hands.

Another example is that people may refuse because they are not interested in the topic of the survey. In this case, the people who do respond will tend to be those who are interested in the topic, and their responses commonly tend to be more extreme (one way or the other) than those of the population as a whole.

Assuming the interviewing is done at people's homes, those who are not contacted tend to be those who are away from home a great deal, and here too it is likely that characteristics and attitudes will be different from those who are at home a great deal.

In the case of an outright refusal, reasons for the refusal might be sought. For example, the refusal might be because the person did not have time immediately but would agree to an interview at a later date. Alternatively the person might agree to a short telephone interview or to complete a self-completion questionnaire. Or the person might agree to answer a few basic questions so that at least a partial response is obtained. Substitution of results from those with similar characteristics might then be considered and this might help reduce the bias.

In the case of non-contact, call-backs should be made at different times. If it is possible to obtain, information from neighbours might help determine when the person was likely to be in. A self-completion questionnaire might be left, or if a telephone number is known the potential respondent might be telephoned either to arrange an appointment or to conduct a telephone interview. Simply leaving information about the survey and asking the potential respondent to telephone an interviewer or the survey office might also result in a contact, but this is less likely to be successful.

Training of interviewers can also help reduce refusals and non-contacts. Interviewers need to be advised both about how to conduct themselves in terms of appearance (dress, etc) and about how to approach respondents. They also need to be advised about the best times to call to minimise non-contacts and refusals.

Solution continued on next page

(ii) 1000 people are selected to take part.

At the first attempt to make contact, 725 are at home but 52 refuse to take part. So the response rate at this first contact is $673/1000 = 67.3\%$.

The second attempt to make contact consists of a call-back to the 275 people who were missed at the first attempt. 30 of these are again missed and, of the remaining 245, 49 refuse to take part. So the response rate at this second contact is $196/275 = 71.3\%$.

Overall, $673 + 196 = 869$ people take part. So the overall response rate is $869/1000 = 86.9\%$.