

# **THE ROYAL STATISTICAL SOCIETY**

## **2006 EXAMINATIONS – SOLUTIONS**

### **ORDINARY CERTIFICATE**

#### **PAPER I**

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.

Ordinary Certificate, Paper I, 2006. Question 1

- (i) The description of the project indicates that a large area is involved. Villages chosen should represent all distinct types of agriculture and land, such as high and low land, mixed crop and animal farms, farms that deal only with crops and those that are only animal farms. Size of farm should also be considered, perhaps split into large and small. If the area is large enough to have a variable climate, for example if some parts have less rainfall than others, this is also important. If possible, reasonably accessible villages should be used. But with so many constraints and only 15 villages, there may be some difficulty making a good choice.

(ii) Method A

Advantages are that those attending are likely to be interested in the project and less likely to refuse to be interviewed; there is no "not at home" problem. Provided they are willing to stay, the necessary number can be obtained – if there are enough at the meeting!

Disadvantages are that a meeting may be poorly attended; unless there are a good number of interviewers, people may not be prepared to wait for their interview; and the sample is in any case self-selected, through interest in the project, enough spare time to attend, and perhaps being more articulate or literate. Good publicity beforehand and plenty of interviewing staff would help considerably. There could be other activities while waiting for interviews (e.g. experts whom they could consult), or some refreshments, or the promise of assistance by advisory services when needed.

The choice of random sample members also has problems. This cannot be done in advance; it must be done at the meetings. As there is no list, some form of selection has to be based on numbering of seating or distributing tickets to people as they arrive and then drawing numbers out of a hat when everyone has arrived. More than one member from the same farm could be selected unless this was deliberately avoided.

(iii) Method B

Absence of maps makes it hard to know whether all farms have been located and listed, and also to fix a suitable sampling interval in systematic sampling. Asking local inhabitants where neighbouring farms are could help to complete a list, and a preliminary tour round the area could then be used to fix the sampling interval. Non-response through being unwilling or unavailable is likely. Considerable time will be needed for this method.

It is however a properly random method, in which all farmers have a chance of being interviewed, and the interviewer can record other useful information as well as answers to questions. Choice of calling time could help to reduce non-response, especially by not going at busy times during the working day. Some form of incentive could be offered to obtain answers, though this would cause ill-feeling among those not in the sample.

Ordinary Certificate, Paper I, 2006. Question 2

The required questions could be covered on a form such as the one below, and instructions to interviewers might be printed in italics to avoid confusion.

*Interviewer, say the following: Good morning (or afternoon, evening, as appropriate). The ... organisation is undertaking a project on the control of XX in this country, and as part of this project would like some information about farms and farmers in this village. It would be very helpful if you could answer some questions about your farm. This will take at most half an hour.*

*Interviewer: If the farmer agrees, say "Thank you" and then start asking the questions below. If the farmer refuses, apologise for taking his/her time and try to arrange an alternative time of interview.*

Q1. Please tell me your name.

Q2. How old were you last birthday?

Q3. Is your highest educational level primary, secondary, or higher? (*Interviewer: delete as necessary*).

*Interviewer: Ask questions 4–7 in turn, recording the answers to all four questions on the grid below. Insert the name of the household member where indicated.*

Q4. Please could you tell me the names of all members of your household.

Q5. What relationship is (*give name, working through all members of the household in turn*) to you?

Q6. How old was (*give name, working through all members of the household in turn*) last birthday?

Q7. I am now going to ask you about the education of each member of your household. (*Interviewer: Say this once only.*)

Was the highest educational level of (*give name, working through all members of the household in turn*) primary, secondary, or higher?

<b>Name of household member</b>	<b>Relationship to farmer</b>	<b>Age</b>	<b>Highest educational level (delete as necessary)</b>
<b>1</b>			primary/secondary/higher
<b>2</b>			primary/secondary/higher
<b>3</b>			primary/secondary/higher
<b>4</b>			primary/secondary/higher

**Solution continued on next page**

Q8. How many hours per week do you usually work on the farm this time of year?

*Interviewer: Record answer on grid after Q.10.*

*Interviewer: Ask questions 9–10 in turn, recording the answers on the grid below.*

Q9. Which members of your household work on the farm this time of year?

Q10. How many hours per week does (give name, working through all members of the household who work on the farm in turn) usually work at this time?

Name of household member	Hours per week
Farmer	

*Interviewer: Ask questions 11–12 in turn, recording the answers on the grid below.*

Q11. What kinds of livestock do you own?

*Interviewer: Record answers in table in Q.12. If answer is none, go to Q.13.*

Q12. How many (give name of type of livestock, working through all listed in turn) do you own?

Type of livestock	Number

*Interviewer: Ask questions 13–14 in turn, recording the answers on the grid below.*

Q13. What crops do you grow?

*Interviewer: Record answers in table in Q.14. If answer is none, go to Q.15.*

Q14. What size of area is planted with (give name of crop, working through all listed in turn)?

Type of crop	Area

Q15. Would you describe the pest risk from XX in this area as high, medium, or low?

Q16. What measures do you take to control XX?

**Solution continued on next page**

*Interviewer, say the following: "That is the end of the questions. Thank you very much for taking the time to answer them. Is there anything you would like to ask me or anything else that you would like to add?" Record responses. Answer any questions if you are able to do so, or say you will try to find out the answers and that someone will get back to the interviewee. Record the action. Record anything interviewee adds.*

Ordinary Certificate, Paper I, 2006. Question 3

- (i) Clearly one possibility is to allow for five entries for every farm, i.e. one for each potential household member. When households are smaller than five, this wastes a large amount of space and leads to large unwieldy files. [It also requires care in programming when calculating means (or other statistics). For example, suppose five columns (one for each potential member) are used for each farm, but a particular farm has only four members. The mean age, say, must be calculated using just these four members, i.e. assuming the fifth is "missing" (and not erroneously taking the fifth as being present but zero). A total (e.g. total number of hours worked) assumes a zero in the "missing" column, but clearly makes no sense on its own without knowing the number of members. It might help to use a code (different from any code for an individual missing value) to indicate those columns in which there are no entries at all because there are fewer than five members at that farm.]

If an additional variable, the number of household members, is created for each farm, and only that number of entries is allowed for, space will be saved but care will still be needed when calculating means and totals (and other statistics) to ensure that the calculation takes the correct number of members into account.

- (ii) Livestock: discrete variable (integer), showing number on farm – as many variables as there are types of animal in the whole survey, score 0 if a particular farm does not have a particular animal type.

Crops: continuous variable, measured area under each crop – as many variables as there are crops in the whole survey, score 0 if there is none of a particular crop on a particular farm.

In each case, could use an "other" category, to save space, if there are particular types of crop or animal mentioned by only very few farmers – include all "rare" types in "other".

Pest risk: an ordinal variable coded 0 (low), 1 (medium), 2 (high) for each farm – only one entry needed. [If there were any "don't know" answers, they would be treated as missing values.]

Pest control: a nominal variable coded 1 (used), 0 (not) for each possible control method for each farm.

Ordinary Certificate, Paper I, 2006. Question 4

- (i) Sample size  $n = 20$ . There are 92 children altogether. The proportions in the schools are as follows.

$$\text{School 1: } 25/92 = 0.2717$$

$$\text{School 2: } 30/92 = 0.3261$$

$$\text{School 3: } 37/92 = 0.4022$$

The sample sizes from the three schools should be as near as possible to  $n \times 0.2717$  etc. These are as follows.

$$\text{School 1: } 5.43 \quad \text{School 2: } 6.52 \quad \text{School 3: } 8.04$$

Therefore take samples of sizes 5, 7, 8.

- (ii) The values of (population size  $\times$  SD of  $m$ ) are as follows.

$$\text{School 1: } 255.0 \quad \text{School 2: } 231.0 \quad \text{School 3: } 329.3$$

The total of these is 815.3, so the required sample sizes are  $\frac{255.0}{815.3} \times 20$  etc, i.e.

$$\text{School 1: } 6.26 \quad \text{School 2: } 5.67 \quad \text{School 3: } 8.08$$

So now the sample sizes should be 6, 6, 8.

- (iii) The sample sizes are almost the same. Method (ii) gives an extra item from the more variable location which could be an advantage. Method (i) is simple to carry out and does not require information from the previous study, since when conditions could have changed. The locations are in (as nearly as possible) the same proportions in the sample and in the population. The estimate for the whole population is simple to calculate. With method (ii), sample results must be kept separate for the three locations in order to calculate population estimates, using strata membership information for weighting.

Perhaps there is little to choose between the methods overall. The final decision could be made on practical grounds, e.g. ease of reaching the schools.

Ordinary Certificate, Paper I, 2006. Question 5

(i) Stratified sampling

Urban and rural districts are two obvious strata that could be used. Simple random samples in each stratum should give districts with different types of employment pattern, although this cannot be guaranteed unless a more complicated method is used.

Cluster sampling

Regions could be used as clusters, each containing several districts. First some regions will be chosen for a random sample of clusters, and then a random sample will be selected within each cluster, giving administrative districts as the units, which again would have a variety of employment patterns.

- (ii) Stratified sampling controls the urban/rural factor better. Cluster sampling would be easier to administer as both the sampling and the interviewing of individuals are carried out in more concentrated areas rather than over the whole country, saving time and resources.

Ordinary Certificate, Paper I, 2006. Question 6

Simple random sampling from voters' register, especially one a year old

The target population is all residents. However, only those who were voters a year ago will be listed (probably not all of them), so any not eligible to vote will be missed, as well as voters who were not listed which will include all newcomers. Thus the sampling frame is biased. Further, there will be non-response due to deaths or people moving away. When a sample has been selected, it may not be easy to contact all the selected members, and some of those contacted may refuse to respond anyway.

Quota sampling

If this method is to be used easily and successfully, there must be some central facility where a wide selection of people will go, e.g. a shopping centre or a main railway station. It is very unlikely that the whole population of the area could be captured at any one place, or even a representative part of that population – some could only be located at home or work places at convenient times.

Bias through interviewer choice of individuals is likely, refusal is more likely from some groups than others, interviewers need good training (and occasional checking) to make sure they correctly allocate individuals to quotas. Refusal rates may be no lower than in random methods, and bias in asking questions is possible (as with any personal interview method).

Ordinary Certificate, Paper I, 2006. Question 7

- (i) Advantages of a diary – it does not usually rely on memory to any extent, does not usually involve any interviewing if a good layout is used and data can be taken directly from it, there is a reasonably accurate record of the activities in each time unit. Headings could make completion easier, e.g. eating, working, leisure, codes for common occupations.
  
- (ii) Disadvantages – in some circumstances instant completion is not possible, risk of some participants losing it, ceasing to keep it up to date, not giving details. Fewer people may be willing to join up in the first place, and some of those who do will not complete it. Also the act of keeping a diary may lead some respondents to change habits, permanently or temporarily. A tedious task like this may not always be done fully. Coding of responses for analysis could be difficult and time-consuming, and analysis more complicated.

Ordinary Certificate, Paper I, 2006. Question 8

(i) Obviously for the youngest age group this is likely to be the best way of collecting information, but there are a number of difficulties. Some length of time will be needed for each child, perhaps a full day (or the equivalent such as a morning and another afternoon/evening). This could easily alter the child's pattern of activities. A sample of several children will be required to find out the full variety of activities in that group, and to discover whether rural and urban differ and in what ways. Observing just one child for a week, for example, would be a waste of the observer's time. Basic information could be collected from parents, for example whether the child is at school or, if relevant, nursery and play school attendance although observation in such activities is not encouraged for legal reasons. Certainly by the time a child is at school "full-time", detailed observation would neither be possible nor necessary. Observation would be limited to out-of-school hours and holidays. An observational study needs fully trained observers for long periods of time and will be very expensive. For the (say) 11–16 age group, a diary (see qu 7) might be sufficient.

(ii) (a) Depending on the access to the web in the region being studied, it is probably better to concentrate on school-based data collection. For those who have access at home, this could be added. Schools would therefore have to agree to take part and to direct children to a web page containing a questionnaire. The answers would vary in quality according to whether the children took it seriously or not, and whether they were supervised or not; supervision might lead to less-than-honest answers for some questions.

For home use, pop-ups while browsing the web could be used. This misses those whose equipment does not accept pop-ups and those who only use the web occasionally for specific enquiries. Data quality by this method must be in serious doubt.

(b) As for any questionnaire, it will be necessary to check that the questions can be understood, and answered easily in the absence of an interviewer. Children in the appropriate age group must be used in the pilot, to test the wording, coverage of appropriate topics and ease of access to the questionnaire. It is also necessary to examine the interface between the questionnaire and the storage of answers. Open-ended questions will need great care in automatic collection methods. A "focus group" could perhaps be useful.