THE ROYAL STATISTICAL SOCIETY

2007 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, 2007. Question 1

(i) Dear Member,

We are carrying out a survey of all our members. We particularly want to find out why people join our organisation and what their special interests are. The views of all our members are important to us, so please take a few moments to complete this questionnaire.

Please be assured that your answers will be treated in the strictest confidence, and will remain completely anonymous when the results of the survey are reported. No-one will contact you as a result of answering the questionnaire and no other organisation will receive any information from it.

It will only take a few minutes to complete the questionnaire. Some questions simply need a tick in the appropriate box; some need you to circle the appropriate number. A few questions ask you to write a short answer in the space provided.

When you have completed the questionnaire please return it to us in the enclosed reply-paid envelope. Thank you for your help – and please reply soon!

(ii) Explanation of rating in Q1a. Sections in a sensible order, personal questions last. All types of status covered in Q8.

(iii) Instructions to circle a number (Q1a) or tick a box (several times) should be repeated in each question. Responses in Q2a and 2b should have the same response boxes, possibly including "poor" and/or "don't know". The rating score 1 – 10 in Q3 needs fuller explanation. By the nature of the topic, Q4 is not easy to word clearly – some will use it a few times a year, others hardly at all, and a few might use it very often (it depends largely on the information given in it); perhaps give some numbers per year, such as "more than 20", "11–20", "5–10", "fewer than 5". Q7 should say "age in years", with the first box "under 25".

It would be useful to know when members joined, so as to evaluate the effect of any changes in aims or activities that may have occurred.

Besides rating the present scheme in Q3, it would be useful to go into more detail about specific items in it.

Higher Certificate, Module 1, 2007. Question 2

(i) By the end of mailing 3, there were 2099 + 701 + 483 = 3283 responses, which is 68.1%. The three mailings were worthwhile, because the percentage response to mailing 1 was only 43.5; after mailing 2 it was 58.1.

Refusals (only 2 to mailing 1) rapidly increased in further mailings, ending with a total of 168. The percentages of actual responses to mailings 2 and 3 were respectively 25.8 and 24.6, which is reasonable.

The second and third mailings were made to all who had not responded or whose questionnaires were undelivered. Non-delivery raises questions which are discussed in (iii) below. The final percentage undelivered at the end of the mailings was $100 \times 132/4822 = 2.7$.

At the end of the process there were still 1239/4822, or 25.7%, who had not responded at all, even though they had apparently received three requests.

(ii) The two main reasons are (1) to increase the size of the sample that is actually collected, so reducing the estimated variance of means etc calculated from it, and (2) to avoid bias. Bias may arise because those who respond immediately are often those who are most interested in making comments (good or bad) and not a fully representative selection from the whole population. Provided, as here, the follow-up response is reasonable, the results should be improved.

(iii) There is no indication that alternative methods of locating non-responders were used, e.g. telephone. Nor is there any indication whether any reasons for refusal to respond were given or asked for; in a postal survey, the respondents would in any case have had to contact the survey organisers to indicate a refusal.

Undelivered questionnaires were presumably sent to the same address a second (or third) time, which seems rather pointless. Attempts to locate these people could be made by printing a list of "lost contacts" in any magazine or publication the Gulf War veterans are likely to read. Commanding officers could contact any still in service.

These methods would not increase the cost of the survey very much. Some other possibilities would require extra resources, such as telephoning non-responders to either (1) encourage them to reply or (2) offer a telephone interview or a face-to-face interview. Besides extra cost, this would delay completion of the survey. Also, collecting information from this group by a different method from that used for the others could lead to different responses. Whether extra effort would be worthwhile depends on the nature of the survey.

Higher Certificate, Module 1, 2007. Question 3

Part (a)

(i)

Number of visits, x_i	*Number of students, f_i*	$f_i x_i$	$f_i x_i^2$
0	16	0	0
1	21	21	21
2	29	58	116
3	19	57	171
	$\Sigma f_i = 85$	$\Sigma f_i x_i = 136$	$\Sigma f_i x_i^2 = 308$

Population size $N = 987$. Sample mean $\bar{x} = \dfrac{\Sigma f_i x_i}{\Sigma f_i} = \dfrac{136}{85} = 1.6$.

Hence estimate of total number of visits is $N\bar{x} = 987 \times 1.6 = 1579.2$.

Sample variance $s^2 = \dfrac{1}{84}\left(\Sigma f_i x_i^2 - \dfrac{\left(\Sigma f_i x_i\right)^2}{85} \right) = \dfrac{1}{84}\left(308 - \dfrac{136^2}{85} \right) = \dfrac{90.4}{84} = 1.076$.

The estimated variance of the estimated total is $N^2(s^2/n)$, where n $(= \Sigma f_i) = 85$ is the sample size, i.e. it is

$$987^2 \times (1.076/85) = 12331.8.$$

(ii) Estimated proportion = $\dfrac{85-16}{85} = 0.8118.$

The estimated variance of this is $\dfrac{0.8118(1-0.8118)}{85} = 0.001797.$

Expressed in terms of percentages, the estimate is 81.2% with estimated variance $100^2 \times 0.001797 = 17.97$.

Solution continued on next page

Part (b)

A stratified random sample selected from library records takes considerable staff time; and posting a questionnaire will not lead to 100% response (see question 2). There is also the problem that there is no guarantee that the questionnaires returned will have been completed by the selected sample members; someone else in a household may have done so. Stratification does help to obtain male and female readers in the correct proportions of those registered, and to obtain estimates for each sex. If a number of questions are to be answered, there may be different patterns of response between the sexes for different questions, and it may be important to know about this. Occasionally the records of names might not make it clear which sex a person is (e.g. "Pat" could be male or female), while it might be that records do not indicate a first name (perhaps only an initial is used) and/or do not indicate the sex. Presumably any school students, not technically "adults", would be on a different register or would have ages recorded. Dates of the "last four weeks" should be specified clearly, although memory is not likely to be good on visits four weeks ago (or five weeks or more by the time the questionnaire is completed).

A systematic sample on one day would be easier to carry out on the spot by one or two people, and thus much cheaper. Forms would be given out for immediate completion. Having a box for completed questionnaires would preserve anonymity and so possibly improve accuracy. The same "memory" problem exists, and there will also be refusals (which may be more numerous among certain sections of the population – e.g. on the way home to a meal!). "Systematic" should mean every 10th, or 20th, or some convenient gap, but this will not be easy to implement at busy times or when a group of people all leave at the same time. Conversely, it could waste time at slack periods. Choosing just one day for the survey will exclude users whose regular weekly routine does not include a library visit that day. A major problem is that not all those leaving the library will be registered users.

Higher Certificate, Module 1, 2007. Question 4

(i) *D* (dissolution) data in order:

4.10, 4.10. 4.52, 4.52, 4.81, 5.09, 5.37, 5.51, 5.51, 5.94, 6.22, 6.22

q *M* *Q*

There are 12 items. The median *M* is midway between the 6th and 7th, i.e. 5.23. The lower quartile *q* is midway between the 3rd and 4th, i.e. 4.52, and the upper quartile *Q* midway between 5.51 and 5.94, i.e. 5.73. [Other conventions for the detailed locations of the quartiles also exist.]

H (human) data in order:

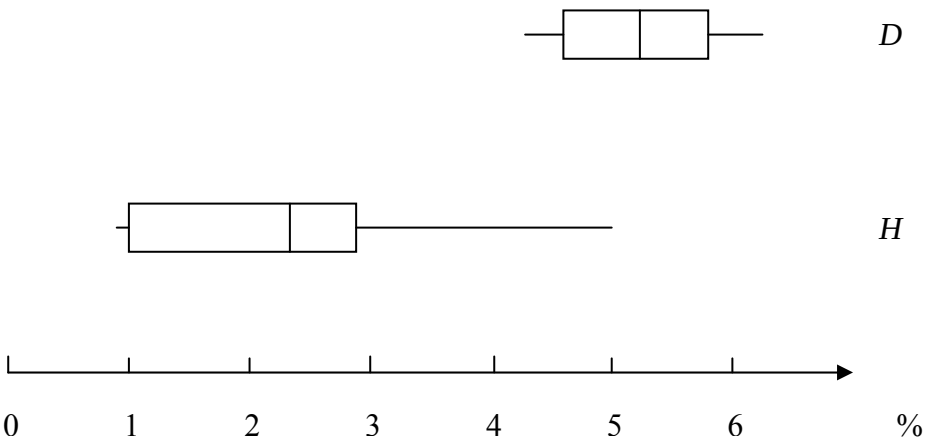
0.84, 0.96, 0.96, 1.55, 2.20, 2.51, 2.52, 2.84, 3.92, 5.01

q *M* *Q*

Here there are 10 items. By similar arguments we have $q = 0.96$, $M = 2.355$ and $Q = 2.84$.

(8)

(ii) The boxplots below are shown horizontally, but vertical plots are equally good. [Note. The limits of electronic reproduction may mean that the plots do not appear *exactly* accurately.]



Solution continued on next page

(iii) For D, sum = 61.91, n = 12, sum of squares = 325.7425.

Hence mean $\bar{x}$ = 5.16 and standard deviation s = 0.759.

[Note. s is calculated as $\frac{1}{11}\left(325.7425 - \frac{61.91^2}{12}\right)$.]

For H, sum = 23.31, n =10, sum of squares = 70.9739.

Hence $\bar{x}$ = 2.33 and s = 1.360.

(iv) The in vitro (D) figures are mostly higher than those for in vivo (H) and are much less widely spread. For each of D and H the mean is roughly equal to the median, but it appears that H is more skewed. This is due to H having two high values some distance away from the others, so its range is much larger than D's. For the same reason, the standard deviation of H is the higher one.