THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 1

SPECIMEN PAPER B

AND SOLUTIONS

The time for the examination is 1½ hours. The paper contains four questions, of which candidates are to attempt **three**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in 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 the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

1. The data given below are the numbers (in thousands) of farms in each of the 50 states of the USA in 1987, given in order of increasing numbers of farms.

1	1	2	3	3	4	4	6	7	8	8	9	9
14	14	17	21	24	24	27	28	33	36	36	37	38
39	42	46	49	50	50	52	57	57	61	70	71	73
78	79	82	87	88	93	96	99	109	115	160		

(i) Draw a stem and leaf diagram to display these data. Briefly discuss whether the data appear to be symmetrical or skew.

(7)

(ii) Compute the median, mean, standard deviation and inter-quartile range.

[Note: $\sum x = 2217, \quad \sum x^2 = 164441.$]

(8)

(iii) State, with reasons, which measure of central location you consider more appropriate for summarising these data. Interpret the results for this measure and for the corresponding measure of variability (or spread).

(5)

2. In a survey of British adults (those aged over 16), there were questions about alcohol consumption. One of these questions related to their average weekly alcohol consumption. The table below shows trends in the distribution of average weekly alcohol consumption during the period 1992–1998.

Average Weekly alcohol consumption by sex and age: 1992 to 1998
Persons aged 16 and over

Age	1992	1994	1996	1998
	Mean number of units per week			
Men				
16–24	19.1	17.4	20.3	23.6
25–44	18.2	17.5	17.6	16.5
45–64	15.6	15.5	15.6	17.3
65 and over	9.7	10.0	11.0	10.7
Total	15.9	15.4	16.0	16.4
Women				
16–24	7.3	7.7	9.5	10.6
25–44	6.3	6.2	7.2	7.1
45–64	5.3	5.3	5.9	6.4
65 and over	2.7	3.2	3.5	3.3
Total	5.4	5.4	6.3	6.4
All persons				
16–24	12.9	12.3	14.7	16.6
25–44	11.8	11.4	11.9	11.4
45–64	10.2	10.2	10.5	11.6
65 and over	5.6	6.0	6.8	6.5
Total	10.2	10.0	10.7	11.0

(i) Using the statistics given in the table, draw suitable diagrams to illustrate

(a) the overall trend in alcohol consumption amongst men and women during the period 1992–1998,

(b) similarities and differences between the 1998 age-specific alcohol consumption for males and females.

(10)

(ii) Write a short report, suitable for publication in a serious newspaper, based on the diagrams you have produced in part (i) and any other aspects of the data you consider relevant.

(10)

3. (i) (a) Explain why non-response is a problem in social surveys.

(5)

(b) Suggest how non-response might be reduced in a mail survey by appropriate follow-up procedures.

(5)

(ii) An interview survey of heads of household is to be undertaken and it has been estimated that 600 completed interviews are needed.

(a) Comment on the following two strategies (A) and (B) for achieving 600 completed interviews.

(A) *Give the team of interviewers contact details of a large sample of heads of household and give instructions that interviewing should stop once 600 interviews have been completed.*

(B) *Give the team of interviewers contact details of a sample of 600 heads of household and give instructions that only one attempt is to be made to interview each of these heads. If no interview is obtained from m of these 600 heads of household then give the team contact details of a further sample of m heads of household and give instructions that several attempts are to be made to interview the members of this further sample.*

(10)

4. A society has two grades of membership (Grade I and Grade II) and a worldwide membership of about 7000 individuals. About 20% of the members fall into Grade II, and about 75% of all members are concentrated into three geographical areas (A, B, C), the rest being spread throughout the rest of the world.

The society publishes a journal which is sent by post to all members, and it wishes to carry out a survey to discover if members find the journal useful, and what aspects of their subject they most wish to see covered in the journal. The society is particularly anxious to discover whether members of both grades find the journal useful.

The society keeps its membership list as a computer file, with one record for each member. The records are stored in alphabetical order of members' names, though the secretary is assured that separate lists could be provided for the different grades of membership.

Consider five possible methods for selecting a sample of members to receive, by post, a questionnaire for this purpose:

simple random sampling,
stratified random sampling,
quota sampling,
cluster sampling,
systematic sampling.

For each method, discuss

(i) whether it would be possible to use this method of sampling for this purpose,

(ii) whether the method would be a good one to choose for the purpose.

(4 marks for each method)

SOLUTIONS

Question 1

(i) Ordered diagram: (stem unit 10000)

STEM	
0	1 1 2 3 3 4 4 6 7 8 8 9 9
1	4 4 7
2	1 4 4 7 8
3	3 6 6 7 8 9
4	2 6 9
5	0 0 2 7 7
6	1
7	0 1 3 8 9
8	2 7 8
9	3 6 9
10	9
11	5
…	
16	0

There is considerable skewness, with a large number in stem 0 and a long tail to the right. There are also gaps.

(ii) The median is between the 25th and 26th in order: $M = \frac{37+38}{2} = 37.5$.

The quartiles are at the 13th and 38th in order: lower quartile q = 9, upper quartile Q = 71. [Other conventions are also acceptable for the quartiles.]

These are in thousands; so we have q = 9000, M = 37500, Q = 71000. The inter-quartile range is then 62000.

$\bar{x} = \frac{2217}{50} = 44.34$ (thousand).

$s^2 = \frac{1}{49}\left(164441 - \frac{2217^2}{50}\right) = \frac{66139.22}{49} = 1349.78$; so $s = 36.74$ (thousand).

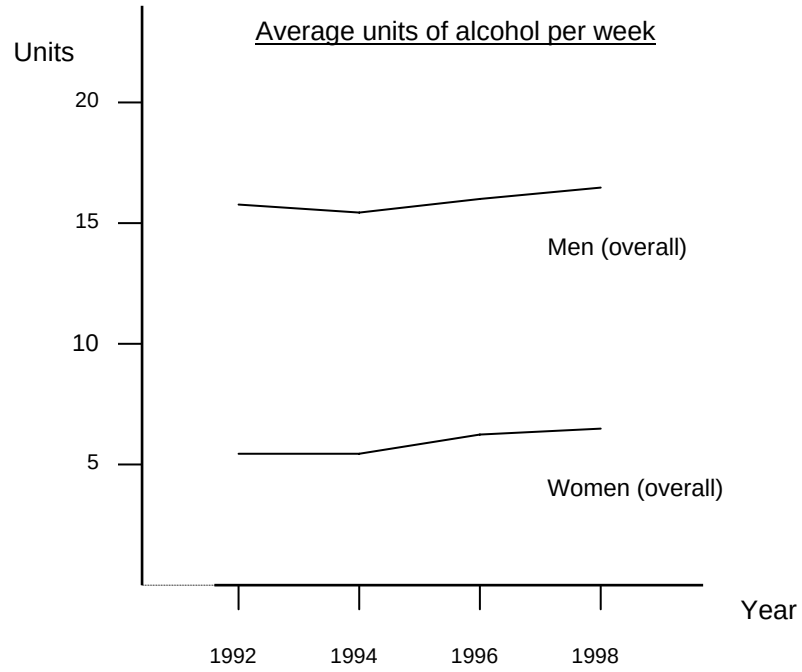
(iii) For reasons given above, the mean and standard deviation will not be good measures of location and dispersion. The median, 37500, and inter-quartile range, 62000 (or the semi-iqr 31000), would be preferred.

The middle 50% of the observations have range 62000. The lower 50% are 37500 or less.

Question 2

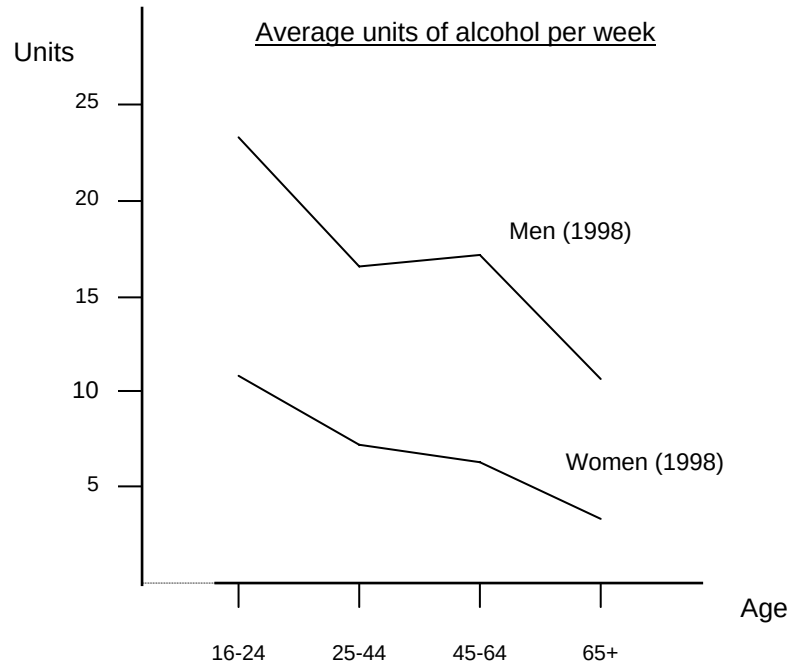
Part (i)

(a)



(The limits of electronic reproduction may make the lines in this diagram and the next appear somewhat ragged.)

(b)



The solution to part (ii) is on the next page

Part (ii)

For overall consumption of alcohol, there was a slight increase over the time period as a whole, but not a very regular pattern.

Both for men and for women, the 16–24 age group showed a distinct rise between the first two and the last two data sets (i.e. between 1992/94 and 1996/98).

Both for men and for women, the 65 and over age group showed a fall over the last two periods (i.e. from 1996 to 1998). This was also true of the 25–44 age groups, but the other age groups showed quite substantial increases.

Overall, there is a general decrease in consumption with age, though this is largely explained by markedly high 16–24 and low 65+ figures.

Overall, men drink two to three times as much as women.

Question 3

(i) (a) Non-respondents tend not to be a random part of the whole population, but instead particular types of person are more likely to fail, or refuse, to reply. Their responses, if known, would quite likely be different from other parts of the population. Unless they are represented, the results of the survey will not validly apply to the whole population. Besides introducing this bias, intended sample size is reduced by non-response and so precision suffers.

(b) Some possible procedures are as follows.

- Send the questionnaire again, once or twice more
- Send reminder letters (without the questionnaire)
- Telephone people who have not responded
- Visit those who have not responded, perhaps only a sample of them

These all require identification of non-responders, usually by means of a number on the questionnaire which is kept separate from the answers to preserve anonymity.

(ii) (a) Strategy A will lead to a sample consisting of those who are easiest to locate, so even if the list is constructed in a properly random way the actual members used will not have been selected at random from the list. Any who refuse at first request will be ignored rather than any attempt being made to persuade them. Since there is a team of interviewers, the more efficient of these may carry out a higher proportion of the 600 (more quickly). Or, alternatively, quickly completed interviews may not have been done so thoroughly. Why stop at 600 instead of attempting to get as many as possible of the originally selected list?

Strategy B will also lead to willing and easily available heads of household being selected, so that the first sample of $(600 - m)$ could suffer considerable bias. The second sample of m ought to be more representative, but the quality of the final data will be affected by how large m is (the larger the better to avoid bias). Office work is also increased by this method.

Question 4

Note that there about 5600 members in Grade I, and 1400 in Grade II; also about 5250 in areas ABC and 1750 in the rest of the world.

Simple Random Sampling from the alphabetical list of members would be easy to organise; questionnaires could be distributed separately or perhaps by including them in the appropriate copies of the next issue of the journal (or in any other regular publication such as a newsletter). It may not be a very good method because the Grade II members are a small proportion, as are "rest of the world" ones. These groups could be in danger of not being sampled very well.

Stratified Random Sampling would be less easy but far more satisfactory because not only these smaller groups but also the A/B/C groups could be examined satisfactorily according to likely variability, cost, proportion satisfied, as well as having appropriate numbers from each group. Lists subdivided more than just by grade would be useful, and modern data storage methods should make identifiers for subgroups easy to provide.

Quota Sampling is totally infeasible. It would be very desirable to split into several groups as suggested above, and if it were possible quota sampling would produce the required sample numbers. As it is not possible, reminders to non-respondents would be the only way of achieving reasonable sample sizes in subgroups.

Cluster Sampling is not feasible because there is no obvious way of splitting into clusters, nor could enough of them be produced to make sampling from them a reasonable process. There do not seem to be any theoretical grounds for wanting to sample in clusters either.

Systematic Sampling from the original alphabetic list would be very easy, and probably just as satisfactory as simple random sampling. However, there is a distinct risk that some surnames would be especially associated with some areas, so that stratification would be better. If the sample method is going to involve producing lists in different groups to sample from, systematic sampling could be used instead of random choice because it may be quicker.

Possible Groups would be Grades I, II, each split into A, B, C, "rest". A 'good' method must compare Grades satisfactorily. This seems to be the most important requirement in the specification.