THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 1

SPECIMEN PAPER A

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. In a project designed to see whether testing procedures in different laboratories give similar results, fatigue tests at three different strain levels (level 1 lowest, level 3 highest) were carried out on samples from the same batch of a composite material. The tests were carried out at 9 different laboratories across Europe. The results (cycles to fatigue) from each laboratory and the means of the samples at each strain level are shown in the table below.

laboratory	*strain level 1 results*	*mean (1)*	*strain level 2 results*	*mean (2)*	*strain level 3 results*	*mean (3)*
A	7335, 6882, 8353	7523	1336, 1693	1515	690, 735	712.5
B	3512, 4000, 4300	3937	977, 1152	1065	428, 460, 473	453.7
C	3822, 5558, 6910	5430	1516, 1607, 1650	1591	630, 675	652.5
D			1740, 1852	1796	447, 718, 935	700
E	4382, 6239, 8930	6517	1488, 1501, 1516	1502	674, 681, 781	712
F	4030, 4138, 4202	4123	1095, 1205, 1290	1197	465, 380, 395	413.3
G	5000, 5245, 5452	5232	1031, 1156, 1238	1142	380, 408, 454	414
H	2510, 2604, 2811, 2900, 2986	2762	738, 771, 864, 883	814	303, 325, 329	319
J	2247, 3800, 5267	3771	957, 1156, 1202	1105	225, 487	356

Two questions are of interest:

(1) Are there substantial differences between the measurements obtained at different laboratories?

(2) Can a model be found to predict the cycles to fatigue at strain levels other than those tested?

Write a preliminary report on your general conclusions about the questions posed in (1) and (2), based on the tabulated data, and supported by suitable graphical evidence. (You are advised to avoid spending excessive time on detailed statistical analysis or repetitive calculations.)

Express your report in a way that makes it accessible to non-technical readers.

(20)

2. (i) What do you understand by the term *simple random sampling*? Describe conditions under which it may not be a suitable sampling procedure or where it would be desirable to combine it with some other sampling method.

(6)

(ii) Choose three different types of *non-sampling error*, and briefly describe circumstances in surveys that give rise to these errors.

(9)

(iii) Outline the main disadvantages of telephone surveys.

(5)

3. Part of a printed questionnaire, to be sent to a sample of households in a Shoppers' Survey, is shown **on the next page**. It is intended that households should return the completed questionnaire by post in a (free) reply envelope.

Comment on the strengths and weaknesses of this questionnaire. Suggest alternative phrasing for those questions which you think could be improved.

(20)

The questionnaire for question 3 is shown on the next page

Questionnaire for question 3

NATIONAL SHOPPERS SURVEY

Instructions

1) This questionnaire should be completed by the main shopper in your household
2) Please give the answers that apply to you by putting an "X" in the appropriate boxes
3) Please return your completed survey in the reply-paid envelope

1. ABOUT YOU AND YOUR FAMILY

1. What is your date of birth? |__|__| |__|__| |__|__|__|__|
   Day Month Year

2. Are you? 1 □ Married 2 □ Single

3. How many persons are living at home? _____

4. Which group best describes your combined household income?

   1 □ Up to £5 000
   2 □ £5 000 – £10 000
   3 □ £10 000 – £20 000
   4 □ £20 000 – £30 000
   5 □ £30 000 – £50 000
   6 □ £50 000 plus

2. YOUR SHOPPING

When doing your main shopping:

1. Which stores do you use for food and grocery shopping? (please specify) _____

2. Why do you buy where you do?

   1 □ Distance
   2 □ Convenience
   3 □ Quality of Products
   4 □ Parking Facilities
   5 □ Prices
   6 □ Food Range

3. What do you spend on groceries?

   1 □ Under £15
   2 □ £15 – £24.99
   3 □ £25 – £34.99
   4 □ £35 – £44.99
   5 □ £45 – £59.99
   6 □ £60 – £74.99
   7 □ £75 – £89.99
   8 □ £90 – £104.99
   9 □ £105+

4. How often have you bought the following products in the past 3 months?

   Environmentally Friendly Products
   1 □ Very Often 2 □ Somewhat often 3 □ Not too often 4 □ Never
   Recycled Products
   1 □ Very Often 2 □ Somewhat often 3 □ Not too often 4 □ Never

5. How many packets of breakfast cereals do you buy?

   □ 1 □ 2 □ 3 4 □

6. What breakfast cereal brands do you buy regularly? (please tick all that apply)

   1 □ Kellogg
   2 □ Nabisco
   3 □ Jordan
   4 □ Shops Own
   5 □ Other (please list) _____

4. (i) Explain the difference which can exist between the *target population* and the *study population* in a survey. Define both of these terms and given an example of a situation where a difference is likely to exist and one where it should not exist.

Comment on the importance of obtaining a good sampling frame for a survey, and how it might be obtained in your two examples.

(9)

(ii) Using the computer output of descriptive statistics given below, explain how these can be used to summarise a set of data and how subsets of the data may be compared. Suggest other information, diagrams or tables that could have been produced from the raw data at the same time as the descriptive statistics and which would help interpretation of the data. Illustrate your answer with appropriate diagrams. The data are wages ($ per hour) in three different sectors (*A*, *B*, *C*) of employment in a region.

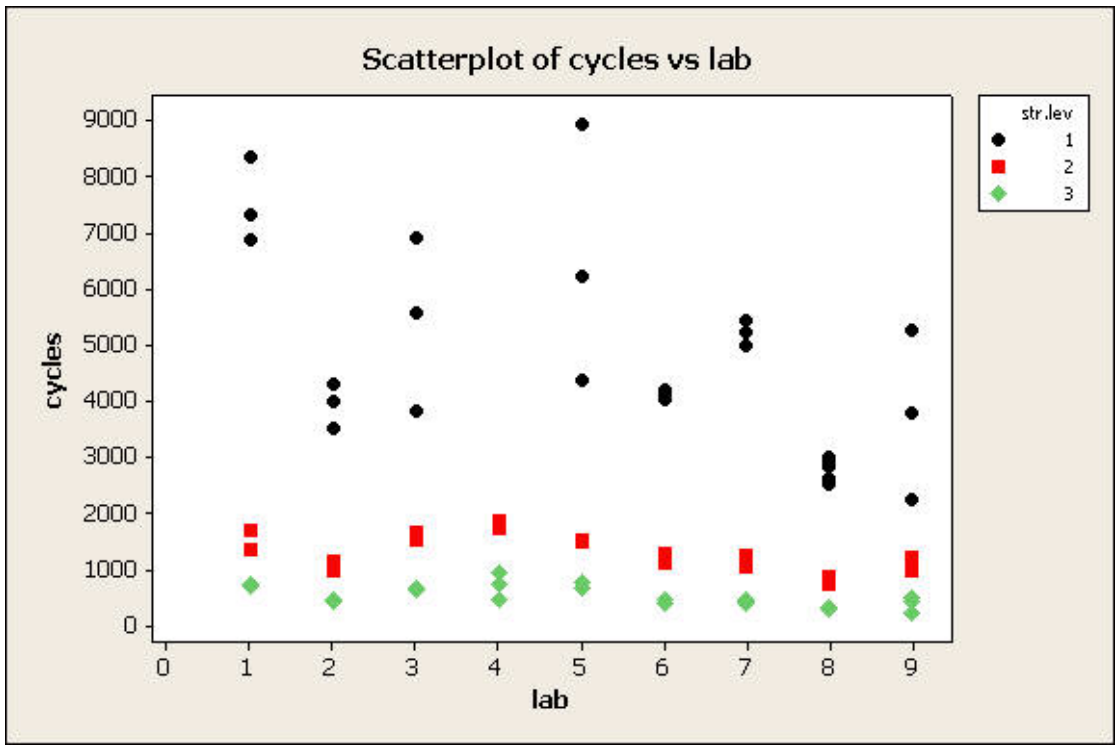
(11)

Sector	*N*	Mean	Median	StDev	SEMean	Min	Max	Q1	Q3
A	188	9.609	8.500	5.920	0.432	1.75	44.5	5.250	12.120
B	48	9.492	9.245	4.244	0.613	3.00	20.4	5.890	11.658
C	14	9.507	10.375	3.633	0.971	3.75	15.0	6.500	11.808

SOLUTIONS

Question 1

It is useful to have a graph showing the measurements made by each laboratory on each strain (laboratories *A* – *J* are relabelled 1 – 9). Individual responses are plotted.



The substantial difference in mean levels between the three strains makes it hard to present the graphical results on a convenient scale. However, several points emerge.

(1) Some laboratories have consistently lower readings than others. For example, *H* has by some way the lowest mean throughout; *A*, *C*, *E* are high; *B*, *J* tend to be low. Variability within laboratories is also very different; this can be seen especially for strain 1, where *C*, *E*, *J* give very wide ranges while *F*, *G*, *H* do not. The basic material used does not appear to have been so variable, because not all within-laboratory variation is large; technical reasons in respect of resources of equipment or people is a more likely reason.

(2) Level 1 is the lowest strain level, and it shows much higher means and much more variation than levels 2 and 3. There is clearly an inverse relationship between cycles to fatigue and strain level. However, a model assuming constant variance would not be suitable; transformations of the *y* (cycles) variable could be explored, possibly log *y*. Prediction will be more accurate at higher strain levels, and should only be attempted within the range of levels already tested; extrapolation below level 1 or above level 3 would be unwise.

Overall, the laboratories lack consistency. If the aim is to have a laboratory-independent prediction, laboratory practice needs to be more consistent. If this cannot be achieved, the model used needs to incorporate a laboratory effect.

Question 2

(i) Simple random sampling, for samples of size n from a population of size N, is where every sample has the same probability of selection. (This probability is, of course, $1/\binom{N}{n}$.) A consequence of this is that every individual in the target population has the same probability of being selected for the sample.

If a population is not homogenous as a whole, but can be split into groups each of which is homogenous within itself, it will be better to select randomly within each group (i.e. stratified random sampling). This allows the different groups to be studied, as well as increasing precision of overall estimates. Also, when a very large population is to be sampled using, for example, a list of names, a systematic sample can be much easier to organise and may be treated as random provided any trends or cyclical patterns in the list are avoided.

(ii) (a) Errors in recording responses, due to poor training of enumerators or interviewers, and/or to carelessness or misunderstanding of subjects' answers. In a postal questionnaire, poor wording of questions may lead to respondents not answering the question intended.

(b) Transfer errors when data are taken from forms and entered into a processing system. Illegible answers could also occur on postal survey questionnaires.

(c) Non-response to postal surveys or refusal to co-operate/be interviewed. This may happen because of lack of interest in the topic being studied, objection to the wording of the questions or the approach of the interviewer, unwillingness to give time to answering, or simply being asked too often to take part in a survey.

(d) Failure to locate individuals/units chosen to take part in a survey. This may for example happen because of faulty lists, non-availability at the time an interviewer calls, premises being empty because people have moved, or different work and/or leisure habits so that individuals would need to be contacted at unusual times not planned for in the survey.

(iii) Telephone surveys only contact people available and willing to answer at the time of ringing, who have some interest in the topic under study, and whose numbers are not ex-directory (if a telephone directory is used as a sample frame). High rates of refusal to respond are likely from people who have been contacted frequently for such surveys. Further, in some countries by no means everyone has a telephone.

Question 3

Section 1 : Question 2 needs more boxes, for widowed, divorced/separated, living with partner.

Question 3 could be more specific and ask how many adults and how many children are living at home.

Question 4 needs to say annual combined household income, and it should be made clear where (e.g.) £10000 is to be entered by having "£5000 – £9999", "£10000 – £19999", etc.

Section 2 : Question 2 should have an instruction to put a mark in all relevant boxes.

Question 3 should say per week, to avoid relying on memory/ guesswork for longer periods. Also, the present question 3 could possibly be called "main shopping", with another question for "top-up shopping" with categories (say) "Under £10", "£10 – £19.99", "£20 plus".

Question 4 might refer to the previous month only (again to avoid relying on memory/guesswork) and should give some numbers such as "More than 5 times", "3 to 5 times", "Once or twice" and "Never (or hardly ever)".

Question 5 should say per week and perhaps be explicitly restricted to the last week. It should include a box for 0 (zero) and possibly one for "More than 4".

Question 4

(i) An essential part of planning a survey is to decide exactly what population the results should apply to – this is the *target population* – and what resources will be needed to achieve this. If some parts of the target population are particularly difficult or expensive to cover, resources of time, money and personnel may not be sufficient to carry out a survey of adequate size. When a part of the target population is omitted for practical reasons, and the survey is carried out on the remainder, this remainder is the *study population* and is not the same as the target population.

If the omitted part of the target population is likely to be rather different from the study population, this must be made clear in a report; but if there is no clear reason to expect that there will be a difference, the results could be used as representing the whole population.

An agricultural survey of a crop grown by large-scale farmers and also by smallholders in a region is made more expensive by the need to travel to each chosen smallholding. But the two types of grower are likely to give different results and so, even if smallholders only provide a relatively small part of the crop yield, both groups must be studied.

On the other hand, if a crop is grown in a large region over which the climate is reasonably uniform, it may be possible to set up administration for sampling in only a few centres in the region and concentrate work near them.

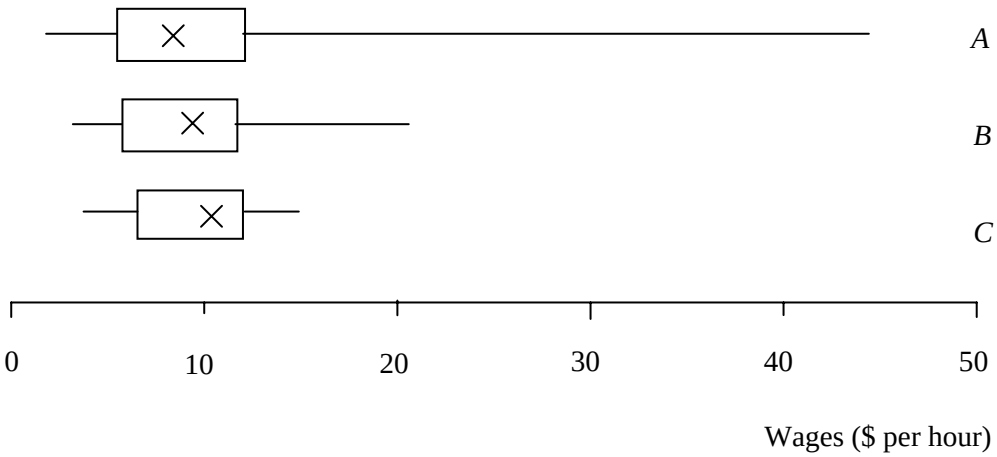
A sampling frame needs to be up-to-date and complete so far as it possibly can. Duplications or omissions should be avoided. In the first situation above, a good list of smallholders for the whole region is required, as well as of major growers. This could add to the basic costs. In the second situation, the lists are only required for limited areas through the whole region.

(ii) Given only the descriptive statistics summary, not all the useful diagrams can be drawn. (However the computer can be asked for dot-plots and stem and leaf diagrams as part of the output.)

The most useful diagrams, without doing any statistical tests, are boxplots.

[Histograms could be asked for as part of the output, but *B* and *C* do not have enough observations to make histograms very useful.]

Solution continued on next page



C is skew to the left, *A* is somewhat skew to the right but appears to have at least one very large outlier. All groups have similar central values, and *A* is more disperse than *B* or *C*. Apart from the longer right hand whisker, *B* is fairly symmetrical. *C* is a small group, but there seems to be some concentration of data just above the median.