THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 8

SPECIMEN PAPER

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 production manager of a soft drinks company arranged for a simple random sample of 500 cans of one of the major brands to be taken from the production line during the time that a large batch of cans was being produced.

(a) The main purposes of the sampling were to assess the accuracy of the can filling line and to check whether the regulation target mean value of 300 ml was being met satisfactorily. The mean content of a can in this sample was 302 ml and the sample standard deviation was 0.5 ml.

(i) Construct approximate 95% and 99% confidence intervals for the mean volume in cans from the production line.

(4)

(ii) Government regulations require the mean volume per can to be at least 300 ml. Using the results from part (i), state whether you would recommend any changes to the volume currently being produced by the filling equipment. Give reasons for your answer.

(3)

(iii) For routine use within the company, the production manager only requires to estimate the mean volume to ±0.05 ml. What sample size would be necessary to construct a 95% confidence interval that meets this requirement?

(4)

(b) A second measure from the 500 sampled cans was whether or not the ring pulls on the cans operated correctly. Of the 500 cans, 12 were noted as having defective ring-pulls.

Construct a 95% confidence interval for the proportion of all cans that have defective ring-pulls. Explain what this confidence interval shows, in language which the production manager (a non-statistician) would understand.

(6)

(c) The production manager would, for practical reasons, much prefer to use a systematic sample. Explain briefly what assumptions this would require, and what advice you might give him.

(3)

2. A government agency wishes to examine the claims for travelling expenses made by staff. The Chief Executive, under pressure to reduce expenses, has instructed that a monthly sample of expense claims should be scrutinised by the Finance Team. This will be carried out for three months.

The sample should have two purposes: (1) to inform the Finance Team of the nature of travelling expenses across the agency, and (2) to act as a deterrent for any employees considering making false or exaggerated claims.

The Head of the Finance Team is seeking your advice in developing a sampling methodology.

Some of the information below might be of assistance in developing the sampling scheme.

- The agency consists of 5 branches, some of which need to do more travelling than others.
- Each claim has a unique reference number. Details of claims (recipient, branch, date, purpose, amount etc) are stored on a central database.

(i) The Head of the Finance Team has heard the phrases "population", "sample", "sampling frame", "random sampling" and "stratified random sampling" but is not quite clear of their precise meanings. Outline what these terms mean in clear language, making reference to this particular sampling query.

(8)

(ii) Outline a possible survey methodology for the Finance Team to use. You should consider the requirements of the agency and the availability of data.

(9)

(iii) If it is decided to carry out a follow-up survey for one month later in the year, comment briefly on whether a ratio or regression method of estimation should be used, and why.

(3)

3. In 2005, university researchers sampled teachers working in state schools in a certain area to ascertain their views on a variety of educational issues. Teachers working in independent schools in the same area were also sampled.

One of the key questions asked was "Do you feel that your efforts are recognised by your Headteacher?" and the results for this question are summarised in the table.

Education sector	*Total number of teachers, N*	*Sample size, n*	*Number who answered "yes"*
State	5000	300	120
Independent	800	50	32

(i) Estimate the proportion of teachers from the state sector in the area who feel that their efforts are recognised by the Headteacher. Express your estimate in the form of an approximate 95% confidence interval.

(3)

(ii) The researchers were interested to see whether there was evidence of a difference between the proportions in the two sectors who answered "yes" to this question. Construct an approximate 95% confidence interval for this difference and state what conclusions you would draw from it.

(4)

(iii) Estimate the proportion of all teachers in the area who feel that their efforts are recognised by their Headteacher. Express your estimate in the form of an approximate 95% confidence interval.

(6)

(iv) Identify two key purposes of stratification. How do these apply in this survey?

(3)

(v) This survey is to be repeated next year with a sample of total size 250. Based upon the results from the 2005 survey, produce an optimal allocation of the sample between teachers in state schools and teachers in independent schools.

(4)

4. A medium sized electronics company recently carried out a survey of staff in its head office in order to ascertain their views on recent changes to their IT facilities.

Staff from three departments were questioned. One of the key study variables was the answer to the question "How many hours each week do you sit in front of a computer screen?" The results are shown below.

Department	*Number in department, N*	*Sample size, n*	*Mean weekly number of hours,* $\bar{x}$	*Standard deviation of the number of hours, s*
Accounts	40	15	30	6
R&D	20	10	15	4
Marketing	10	3	20	3

(i) Explain when (a) the finite population correction, (b) the *t* statistic, should be used in the analysis of survey data. What would be the effect of not using them in making inferences from this survey?

(5)

(ii) For each of the three departments, calculate a 95% confidence interval for the mean number of hours that employees sit in front of a computer screen.

(5)

(iii) What sample size from the marketing department would have been necessary to estimate the mean number of hours that employees from that department sit in front of a computer screen to within ±3 hours at the 95% level of confidence?

(4)

(iv) Construct a 95% confidence interval for the mean hours sat in front of a computer screen for <u>all</u> the employees.

(6)

SOLUTIONS

Question 1

(a)(i) Let X (ml) be the volume contained in a can. The sample size (500) is very large, so the Central Limit Theorem assures us that, to a very good approximation, the underlying distribution of the sample mean can be taken as Normal (alternatively, it does not seem unreasonable to assume underlying Normality for X itself, though this is not stated in the question).

We have $\bar{x}$ = 302 and s = 0.5. Let μ be the true (long-term) mean of X. A 95% confidence interval for μ is given by $302 \pm (1.96 \times 0.5/\sqrt{500})$, i.e. 302 ± 0.044, so the interval is (301.96, 302.04). For a 99% interval replace 1.96 by 2.576 to give 302 ± 0.058, so the interval is (301.94, 302.06).

(a)(ii) The variability is very low, so the machinery is working well. Noting that the confidence intervals lie comfortably above the specified minimum of 300, and provided that the setting for the mean can be adjusted without affecting the variability, the setting could be altered to give a mean of 301 (or somewhat less) without danger of can contents going below the required amount.

(a)(iii) Let n be the sample size. It is required that $1.96 \times 0.5/\sqrt{n} < 0.05$, i.e.

$$\sqrt{n} > \frac{1.96 \times 0.5}{0.05} = 19.6, \quad \text{or } n > 384.16.$$

So take n = 385. [Note that n is still very large, so the Central Limit Theorem still comfortably applies. Had n turned out to be small, consideration would have had to be given to using a t value rather than 1.96 from N(0, 1) – and that would have required an assumption of Normality for the underlying distribution.]

(b) Let p be the true proportion of the population having defective ring pulls. The sample estimate is $\hat{p} = 12/500 = 0.024$. The variance of $\hat{p}$ is estimated as

$$\frac{\hat{p}(1-\hat{p})}{n} = \frac{0.024 \times 0.976}{500} = 0.00004685$$

so the standard deviation of $\hat{p}$ is estimated by $\sqrt{0.00004685} = 0.00684$.

Hence an approximate 95% confidence interval for p is given by $(0.024 \pm 1.96 \times 0.00684)$, i.e. it is (0.011, 0.037). With 95% probability of being right, we can say that the range 1.1% to 3.7% covers the true percentage of cans having defective ring pulls.

(c) Systematic sampling takes a can at regular intervals from the production line, say every 10th can. This is likely to be easy to do. [In contrast, simple random sampling first requires the cans to be numbered in some way, and then selects them "at random". Assuming this is done on the production line itself, there will sometimes be very small gaps between successive cans to be sampled, and sometimes large gaps. This may be very inconvenient and may lead to errors in taking the chosen cans as they come along the line.] However, if there are any trends in the volume being dispensed per can, it is possible for a systematic scheme to get in phase with the trend and so lead to bias in estimation. Since the variance found in the random sample was extremely small, trends seem unlikely. In this case, systematic sampling would be acceptable, and the samples would be likely to behave as though they were simple random samples. The results and formulae for simple random sampling would have to be used in any analysis but this would seem reasonable.

Question 2

(i) A population is a complete collection of items (people etc) to be studied. A sample is a selection from the population. A sampling frame is a list of all the items in the population. Here the population is all the claims made in a month, a sample consists of some of the claims, and the sampling frame is all the items in the database for that month. Random sampling is selecting from the population (using the sampling frame, with items numbered to identify them) in such a way that the probability of each particular item being sampled is known. [In simple random sampling each item has the same probability of appearing, but other random sampling methods are also possible.] Stratified random sampling is carried out when the items are first divided into groups or "strata", and a sample (often a simple random sample) is then taken from each group. The strata are formed on the basis of a characteristic which may differ from group to group, such as the branches of the agency or the sizes (broadly classified) of individual claims.

(ii) Each claim has a unique reference number in the database, so there is a good sampling frame. The two purposes, information and deterrent, need to be met by constructing a number of questions (items of information), answers to which can be obtained from the data in the database. It may well be sensible for each branch of the agency to form a stratum. The sizes of claims could form further strata within each branch; probably claims can simply be "large" and "small", although because the need for travelling varies between branches it may be useful to specify more than two classes of claim. If there is particular interest in large claims, this stratum may be sampled more intensively (i.e. a larger fraction of its members chosen for sampling). Details collected could either be continuous measurements (e.g. time taken for a journey and the size of claim to nearest £, $, €) or some form of classification (e.g. the nature of the task for which the journey is made). Method of travel will clearly be important. Without knowing what sort of an agency this is, more details cannot be given.

A pilot to test the proposed method would be very useful in sharpening the suggested questions.

Analysis should first be done by the main stratum classification, i.e. branches. The sub-strata (claim size etc) can be studied within each stratum. It may perhaps be worthwhile to give results averaged over the whole study, but the differences between strata and within strata would probably be more useful. Conclusions should be relevant to the purposes of the study, although unexpected findings could also be mentioned.

(iii) After the initial three months there will be a lot of information on the general characteristics of claims, and this could be used as supplementary information in later studies. A regression method needs fewer assumptions than a ratio method; if it seems not to be giving useful results it can always be abandoned, so the preliminary analysis should include it.

Question 3

(i) Let p_S be the proportion answering "yes" for state teachers (and, later in the question, p_I similarly for independent teachers). The sample estimate of p_S is $\hat{p}_S = 120/300 = 0.4$ and its variance is estimated by $\hat{p}_S(1-\hat{p}_S)/300 = 0.4 \times 0.6/300 = 0.0008$. Thus an approximate 95% confidence interval for p_S has limits $0.4 \pm 1.96 \times \sqrt{0.0008}$, i.e. it is (0.345, 0.455). [See solution to question 1(b) for more details for this calculation.]

(ii) Similarly, $\hat{p}_I = 32/50 = 0.64$ and its variance is estimated by $0.64 \times 0.36/50 = 0.004608$. The difference $p_I - p_S$ is estimated by 0.24 and the variance of this difference by 0.0008 + 0.004608 = 0.005408. Thus an approximate 95% confidence interval for the difference has limits $0.24 \pm 1.96 \times \sqrt{0.005408}$, i.e. it is (0.096, 0.384).

The value 0 is not in the interval, so the difference between p_I and p_S is significant; there is 95% confidence that the interval from 9.6% to 38.4% covers the true percentage difference.

(iii) Let P be the proportion saying "yes" for all teachers in the area. P is a weighted average $W_I P_I + W_S P_S$, estimated as $\left(\frac{800}{5800} \times 0.64\right) + \left(\frac{5000}{5800} \times 0.4\right) = 0.433$. Note that the weights are the *total* numbers of teachers in the areas. The variance of this estimate is given by $W_I^2 \mathrm{Var}(\hat{p}_I) + W_S^2 \mathrm{Var}(\hat{p}_S)$

$$= \left(\frac{800}{5800}\right)^2 \times 0.004608 + \left(\frac{5000}{5800}\right)^2 \times 0.0008 = 0.0006822.$$

Thus an approximate 95% confidence interval for P has limits
$0.433 \pm 1.96 \times \sqrt{0.0006822}$, i.e. it is (0.382, 0.484).

(iv) Stratification subdivides a population into groups which it is anticipated might differ markedly from each other in terms of the characteristic(s) being studied. Good information may then be obtained on each group separately, and the whole-population estimates of parameters will be more precise than those from simple random sampling. In this study, the groups clearly differ markedly, as is seen from the significantly different estimates of p_I and p_S. Good information has obtained on each, and P has been better estimated than from simple random sampling.

(v) The optimal allocation takes $n_h \propto N_h \sqrt{p_h(1-p_h)}$, where h indexes the strata. Sample estimates of the p_h are used, and we have $N_I \sqrt{p_I(1-p_I)} = 800\sqrt{0.64 \times 0.36} = 384.0$ and $N_S \sqrt{p_S(1-p_S)} = 5000\sqrt{0.4 \times 0.6} = 2449.5$. Now, the required total sample size is 250, so we divide 250 in the proportions 384.0 : 2449.5 to get 33.88 and 216.12 respectively. So take 216 from the state sector and 34 from the independent sector.

Question 4

(i) (a) In sampling from a finite population of N items, whose variance is S^2, let n be the size of the sample; then $f = n/N$ is the "sampling fraction" and $(1 - f)$ is the "finite population correction". To illustrate its use, let the mean of X, a continuous variable measured on each of the n units, be $\bar{x}$, then a standard result is $\text{Var}(\bar{x}) = (1 - f)S^2/n$; whereas for an infinite population, using the same notation, we would have $\text{Var}(\bar{x}) = S^2/n$. Hence the name "finite population correction" (fpc). When f is very small (in practice say <5%), use of it makes very little numerical difference to calculations and it can be neglected in theoretical work with little adverse effect. But for larger f, the fpc should always be used.

(b) When S^2 is not known, an estimate s^2 from the sample must be used in calculating confidence intervals for $\bar{x}$ (or in any formal hypothesis tests). Hence a t statistic, *not* a N(0, 1) statistic, is required. The "infinite population" formula for a confidence interval using a critical point from the N(0, 1) distribution (typically 1.96 for a 95% interval) is replaced by the corresponding formula using the appropriate point from the t_{n-1} distribution. As a guideline, this should be used for $n < 30$. For larger values of n, it would be satisfactory to use N(0, 1).

In (a), failure to use the fpc make variance estimates too large. In (b), failure to use t makes confidence intervals too narrow.

(ii) Accounts: we have $f = 15/40 = 0.375$. The two-tail 5% point of t_{14} is 2.145. So the interval is given by $30 \pm 2.145\sqrt{(1-0.375)36/15} = 30 \pm 2.73$, i.e. it is (27.27, 32.73).

R & D: we have $f = 10/20 = 0.5$. The two-tail 5% point of t_9 is 2.262. So the interval is given by $15 \pm 2.262\sqrt{(1-0.5)16/10} = 15 \pm 2.02$, i.e. it is (12.98, 17.02).

Marketing: we have $f = 3/10 = 0.3$. The two-tail 5% point of t_2 is 4.303. So the interval is given by $20 \pm 4.303\sqrt{(1-0.3)9/3} = 20 \pm 6.24$, i.e. it is (13.76, 26.24).

(iii) It is assumed that the sample standard deviation remains the same. So a larger sample will be needed in order that t should have more degrees of freedom, thus reducing the 5% critical point that determines the width of the confidence interval. In part (ii), the value ±6.24 was found for the case $n = 3$. Trying $n = 4$ gives a t critical point of 3.182 (from t_3) leading to ±3.70. Trying $n = 5$ gives a t critical point of 2.776 (from t_4) leading to ±2.63. So use $n = 5$.

Solution continued on next page

(iv) The overall mean is estimated by the usual stratified sample mean $\bar{x}_{st}$ given by

$$\bar{x}_{st} = \sum_i \frac{N_i}{N}\bar{x}_i = \left(\frac{40}{70}\times 30\right)+\left(\frac{20}{70}\times 15\right)+\left(\frac{10}{70}\times 20\right) = 24.29$$

(note that the weighting is by the *population* stratum sizes). The variance of this is

$$\mathrm{Var}\left(\bar{x}_{st}\right) = \sum\left(\frac{N_i}{N}\right)^2\left(1-f_i\right)\frac{S_i^2}{n_i} \qquad \text{[estimate the } S_i^2 \text{ by the sample values]}$$

$$= \left(\frac{40}{70}\right)^2\left(1-0.375\right)\frac{36}{15} + \left(\frac{20}{70}\right)^2\left(1-0.5\right)\frac{16}{10} + \left(\frac{10}{70}\right)^2\left(1-0.3\right)\frac{9}{3} = 0.59796.$$

This "pooled estimate" of the variance has 25 df.

The two-tail 5% point of t_{25} is 2.060. So the 95% confidence interval is given by $24.29 \pm 2.060\sqrt{0.59796} = 24.29 \pm 1.59$, i.e. it is (22.70, 25.88).