ROYAL STATISTICAL SOCIETY EXAMINATIONS, 2002

REPORTS OF EXAMINERS

Ordinary Certificate Paper I

As always in such examinations, candidates should take care that they answer the actual question as set. If there is a specific problem scenario then the answer should relate to it.

As far as possible candidates should relate their answers to realistic situations. Discussion of various problems and issues should show understanding of relative magnitudes of problems in a particular scenario. For example, the fact that some individuals in question 1 may be illiterate (uncommon in the UK) is dwarfed by the massive non-response of those who can read English but are apathetic.

Where problems can easily be overcome this should be indicated. Thus, for example, in question 6, computer crashes can cause loss of data - but fairly large data files can easily be backed up even on floppy disks. A problem easily avoided is not a problem but an indication of necessary practice.

Question 1
In a survey which estimates a parameter of a target population, *bias* is the difference between the true population parameter value and the mean figure which would be obtained were we to repeat the survey a large number of times using the same tools, procedures and analysis. Bias is a feature of the methodology used, not of individual respondents. Few candidates gave a good *definition* of bias, most just illustrated it.

Types of bias include:

- Selection Bias [arising from deficiencies in sampling frame, faulty sampling procedure or non-response]
- Questionnaire Bias [arising from leading or poorly worded questions]
- Interviewer Bias [either stimulated by the interviewer or in the recording of the answers]

In this survey there certainly was selection bias. Firstly, certain groups would be excluded – those not registered or those who had moved or could not read English. This, however, all pales into insignificance (and this was a real example taken from an actual published survey) compared with the massive non-response which saw only 52 responses out of about 43,342 distributed copies of the questionnaire. Those who did reply are likely to be atypical of the electorate; perhaps they have strong feelings on the subject, have time to write, are habitual complainers, etc. There may also have been questionnaire bias (actually there was) but we have no way to know that from the paper as set. There may have been interviewer bias for those who phoned in – we presume that those receiving such calls had no training on neutral prompting etc and had every motivation in wanting comments to be positive.

Some candidates did answer this well but common errors were:

- Failure to give a good general definition of ‘bias’ (rather than examples of it)
- Failure to list the possible types of bias and discuss all of them in context of this survey

- Failure to get a sense of proportion – the absolutely massive non-response is the central problem – and the overwhelming reason for it is likely to be apathy rather than illiteracy or blindness.
- Suggesting that the smallness of the sample actually obtained was a source of bias. A simple random sample of 52 would **not** have been biased, although its 'precision' would have been poor because of the sampling variation resulting from the small sample size.

Question 2
Surprisingly few candidates gave an excellent answer to this question.

The fact that the sample is unrepresentative does **not** in itself imply low reliability; it is quite possible to argue that the same procedure would consistently produce similar bias. Even strong bias does not imply unreliability in this sense. The issue is not whether exactly the same people would respond another time, but whether similar people would respond and produce a similar result.

The smallness of the sample obtained does raise a reliability issue. This is not because of non-response bias, but because a smaller sample results in higher sampling variability, which in turn implies 'unreliability'.

Whilst it might be relevant to mention that if a short-term change of time of survey radically altered results this would imply unreliability, the fact that over a long period of time opinions might change does not imply unreliability in any worthwhile sense. Obviously anything can change over sufficient time – this is not a mark of unreliability.

Question 3
For each of parts (i), (ii) and (iii) there were two parts – how to do it, and the advantages and disadvantages. Not all candidates answered both. A common fault again was a failure to get a sense of proportion. In general, taking a random sample is not particularly difficult, arduous, or lengthy, given a sampling frame. Some of the methods suggested (eg writing all the names of the 103,456 residents on pieces of paper and putting them in a hat) would have made them so, but only because this is unnecessarily cumbersome. Getting personal interviews for a good sized sample is inevitably time-consuming and expensive - selecting a random sample isn't.

For (i), issues could be raised of inadequacies of sampling frame, but with a postal survey the major issue is likely to be non-response bias. The main non-response problem is likely to relate to apathy, procrastination or concern for privacy. – people can't be bothered, don't get around to or simply don't want to participate.

For (ii) it may be noted:

- Stratification ensures representativeness and (usually) improves precision – but it does not remove selection bias as a simple random sample (assuming 100% response) is not biased.
- Personal interviews generally improve response rates. There are issues of training interviewers etc. Generally interviewers are instructed by professional polling groups to prompt strictly neutrally. They might use their own words, perhaps, to encourage cooperation in answering the survey, but should not usually 'explain the question' as this is very likely to introduce interviewer bias.

- Age is rarely available as a basis for stratification, since people's ages are not usually listed publicly.
- Telephone surveys are quick, cheap, and are increasingly used. Obviously there is bias since not everyone in the electorate is in the telephone directory. It is much easier in a telephone survey to enter responses directly into a suitable computer package – expensive equipment or laptops are not needed.

Question 4
Most candidates could do part (i) of this, few could do part (ii) and some obtained absurd answers to it. For part (iii), given that the information is available and the calculation takes about ten minutes, it is worth using the optimum allocation. The latter ensures maximum precision for given cost. Answers which relied on gut feelings or a partial look at some comparisons were not convincing.

Question 5
This was a basic bookwork question, inviting standard rote-learned bullet points. Some candidates duly complied and got good marks – though some concentrated on just one aspect (for example, the wording of questions on the questionnaire) about which they wrote at length.

Question 6
Answers to this should have been applied to the particular example, which specifically mentions 700 parents etc.

Parts (a) and (b) were fairly standard, though reference to a specific package (SPSS, Excel etc) was useful.

The answer to (c) primarily concerned automatic checks on data type on entry (is it a number, alphanumeric, how many digits etc). It would also concern some sampling for rechecking - though it seems unlikely that the whole lot could be entered twice.

The answer to (d) was intended to be along the lines of how missing observations could be dealt with (for example, in SPSS) by inserting a special code for 'Missing' distinct from 'Don't Know' or 'Inapplicable'. The package can then omit these from tables or list them separately. A large number of candidates suggested that the missing numbers could be 'imputed' or 'estimated' from other data. There are, of course, whole books written on how to deal with 'missing values'. But the picture such comments gave in the context of this question was (a) unrealistic (for example, it is hard to see how one could conclude that a boy actually played football on the grounds of demographic features) and (b) liable to lead to production of cross-tabulations some of which were actually made up or 'imputed'. It is hard to think of a question or context for such a survey by a Local Authority, in which descriptive tables should be produced indiscriminately mixing fact with invention (or 'imputation'). A third suggestion – recontacting the parents to ask about missing values – might just be possible in this instance, though in a more general survey of the public might well be unrealistic.

A number of people seemed to take this question to be about Optical Character Recognition rather than data analysis. I have not seen many such applications in which OCR is used – and it is unlikely to be available. No indication was given in the question that this was an issue, and for many of those who discussed OCR it was a distraction from the more central issues.

Part (ii) produced some interesting answers. A good computer package enables quick, accurate, and complex analysis - including production of tables/graphs, crosstabulations, and statistical analysis. It is hard to think of many disadvantages in computer analysing a survey of (hopefully) near 700 forms. Suggestions that data could be lost implies stupidity in not backing it up (large SPSS data files can fit even on a floppy disk), just as the 'virus threat' implies using an unprotected and at-risk machine. Expense of having a machine and package could, in some contexts, be an issue, though probably not in a UK Education Authority.

Question 7

The wording of this question could have been interpreted as asking for the methodology used to obtain weights (for example, in a Family Expenditure Survey), or for the actual categories and weights used. Likewise part (ii) could be answered either by writing a general piece on difficulties of collecting prices or a description of what is actually done to overcome the problems. In view of the ambiguity marking was of course lenient, but at least some actual details from a particular index and its connected surveys were expected.

Question 8

This was largely a bookwork question, but even then often had overly vague answers. The wording of the question was perhaps slightly misleading in that for a quota sample there is no explicit sampling frame, but what was required was clear enough.

On quota sampling, some answers assumed that the 'quota' was just the total sample required, and that a quota sample involved identifying the target group and getting responses from them up to a certain number or quota. The whole point of a quota sample is that quotas of different types (eg based on age, sex, social class) make up the total sample. Situations in which it is most useful are those requiring speed and economy, for example, getting quick public reaction to a particular recent political event.

On cluster sampling, some seemed to assume that the target population itself was actually some kind of small cluster. The central point here is really that clusters are sometimes randomly sampled from a population because they are in closer locational proximity and so cheaper and quicker to survey than an equivalent simple random sample which are more spread out geographically. The population is not 'put into clusters'(as some candidates wrote) but cluster sampling is used in a situation where the population is in fact already in clusters. In general a cluster sample is less precise than a similarly sized simple random sample, but it is much cheaper. It may be used to maximise precision for given expenditure.

Ordinary Certificate Paper II

Overall, the answers were clearly legible with relatively few spelling mistakes. Candidates should be reminded to follow the rubric by starting each question on a new page, listing all answers on the front of the booklet and tying in graph paper securely, preferably close to the answer to which it applies.

A few candidates used red pen in graphs and tables – they should be advised not to do so. Candidates need reminding that all tables should be neatly aligned with suitable headings. Axes on graphs should be ruled and clearly labelled, with units stated where appropriate. If there is more than one series plotted on a single graph, then each should be clearly identified.

Candidates should also be reminded to read the questions carefully. For example, in 7(i), the answer was to be rounded to 1 decimal place and in 8(i), the letters were to be marked near the plotted points.

Candidates should be advised that those who attempt all parts of every question stand a much better chance of passing. Candidates seemed reluctant to answer parts of questions requiring comments and, when they did, these comments were not always succinct and to the point. Only a few candidates indicated that they had run out of time.

Question 1

Although the majority of candidates noticed the unequal class intervals, a sizeable number did not. Having calculated the frequency density, some plotted the graph incorrectly as a series of separate bars of identical width. Others did not label what was being measured on the vertical axis. It is important that candidates realise that it is a continuous scale on the horizontal axis and that the scale markings should include the class boundaries or the class midpoints. As the variable was Age last birthday, the class boundaries of the 15-19 class, for example, were at 15 and 20 with a class mark of 17.5. Many candidates were half a year out in their plotting. Also, in calculating the frequency density, class widths of 4, 9 and 14 were incorrectly used instead of 5, 10 and 15. The best histograms gave a unit of area to indicate the scale.

Those who had plotted the graph correctly concluded that the organisation might be ageist by reference to the graph. A few calculated the median, although this was not expected. Some of those whose first language is not English were obviously uncertain as to what was meant by ‘ageist’, although the meaning of this term was explained in the question.

Part (iii) was generally well-answered, and even those who had failed to recognise the unequal class intervals in (i) realised their misleading consequences.

Question 2

Most people explained dispersion satisfactorily but then some defined the mean, median and mode as measures of dispersion! Three different measures were requested so standard deviation or variance counts as one measure. Similarly, only inter-quartile range or semi-interquartile range (quartile deviation) was accepted. The coefficient of variation was not accepted as it is a compound measure and again incorporates the standard deviation.

Question 3

Formulae for short-cut calculation of the standard deviation were often incorrectly remembered. Many could not obtain the frequency table correctly. Candidates should, at the very least, have checked that the total frequency was 100. Some did not realise that the class marks (i.e. midpoints) were at 99.5, 299.5 etc. In part (iv), many candidates did not compare the results.

Question 4

This was the least popular question. The Venn diagram was often incorrectly drawn as some assumed that everyone ticked at least one factor; by part (ii), candidates should probably have realised this and definitely realised it in part (iii). The conditional probabilities in (iii) were poorly understood by all but the best candidates.

Question 5

In the line graph, it is advisable for the points to be joined with ruled lines rather than freehand drawing. Some candidates were not used to finding an odd period moving average and attempted to centre the averages. The trend was sometimes plotted at the wrong times. All candidates should be advised to check that the trend looks as though it goes through the actual data series and if it does not they should recheck their work. Part (iii) was very poorly done with most assuming that it was the actual expenditure rather than the trend that was being predicted. Very few pointed out that the nature of the averaging process means that the moving average trend is not known at the most recent data point.

Question 6

This question was poorly answered with most candidates assuming more was shown on the graph than was actually there. Many answered only part (i) and ignored part (ii). In Diagram 1, most noticed the absence of a vertical axis with labels and units. In Diagram 2, only about half realised that the zero on the vertical scale had been suppressed and even though a handful noticed the lines were indistinguishable and crossed, they mostly assumed that it was Company A's Sales that dropped off and did not consider the possibility that it might be Company B. In Diagram 3, the vast majority assumed that the dotted lines were some sort of projection, although this was not stated. This was intended as a question on poor drawing of graphs but many got bogged down on the semantics of whether poor sales automatically meant disastrous results.

Question 7

(i) This was generally done satisfactorily apart from arithmetic slips and not rounding to the requested number of decimal places.

(ii) The chain-based method was obviously less familiar and some candidates gave a table giving all possible earnings relatives and expected the examiner to select the appropriate ones!

Most could explain the meaning of the fixed-base index but not the chain-based index.

Question 8

This question was well answered on the whole.

The question indicated which was the x and which the y variable and it was advisable that the candidates used the horizontal and vertical axes, respectively, for them.

In (i), most scatter diagrams were satisfactory, although the choice of scales could have been better and the letters were not always marked, as required.

In (ii), rank correlation was not wanted although some marks were given if it had been calculated correctly. The formulae for S_{xx} and S_{xy} were often incorrectly remembered. Some candidates forgot to comment on the value of the correlation coefficient.

In (iii), many gave themselves extra work by recalculating S_{xx} and S_{xy} instead of using the values found in (ii). Some rounded the value of the slope too early before calculating the intercept.

In (iv) two points are sufficient to fix the line. To draw the regression line accurately, it's usually best to calculate one point towards the left of the graph and one towards the right. It is usually best to check that $(\bar{x}, \bar{y})$ lies on the line.

Higher Certificate Paper I – Statistical Theory

The aim of this paper is to test the ability of candidates to understand and interpret basic statistical theory and to apply and adapt it to simple practical situations.

There were very few infringements of the rubric this year: only 2 of the 30 candidates provided answers to more than the requisite 5 questions.

The general standard was good. In all, 22 of the 30 candidates (73%) gained a pass mark on this paper, with an average mark of 64. It was pleasing to see 10 results at distinction level (75% or more) including three at over 90%.

Areas of the syllabus that continue to give cause for concern are combinatorial analysis and, in particular, conditional probability. Candidates and those who prepare them should be aware that the fundamental ideas of conditional probability are applied throughout statistics, and that an incorrectly calculated (or wrongly used) probability for a conditioning event makes catastrophic errors almost inevitable. In the present paper, weaknesses in this topic were perceived in questions 2(ii), 4(ii), 5 and 7 (albeit often accompanied by good work on other areas).

On the positive side, most candidates appeared to be competent in the more straightforward parts of the syllabus, including use of the Normal distribution and the central limit theorem, the Poisson and exponential distributions and correlation and simple linear regression calculations.

Question 1
This question on combinatorial analysis was both unpopular (only 12 attempts) and weak (average mark 8.3 out of 20). There were several derisory attempts which showed confusion in the $\binom{n}{r}$ formula, although most others had at least some notion of the combinatorial probabilities required. The final part, dealing with the probability of a 3-3-3-4 distribution of Clubs, was beyond the reach of almost all candidates, who seemed to be unaware of the method of successive selection of hands, noting at each stage how many cards are left as each successive hand is dealt with.

Question 2
A much more popular question with 25 attempts, but an average score of only 9.6 marks. The proportion more than 168 cm tall was generally well done, but surprisingly many candidates could not correctly interpret ‘within one standard deviation of the mean’, either omitting this part or finding the answer as $\Phi(1)$. Many candidates had difficulty obtaining the percentage points of a truncated Normal distribution and the correct proportion in part (ii), although these answers are simply obtained by means of conditional probabilities. The application of the central limit theorem in the final part (iii) was satisfactory in most cases.

Question 3
With 15 attempts this question was not popular, but it elicited several good answers and an average score of 13.1 marks. Most candidates understood that the set-up in (i) was binomial, although several did not bother to identify $p = 1/2$ and then left the mean and variance as general formulae. Convincing explanations of the given probability were rare, but most candidates computed the distribution correctly. Several found the mean and variance using decimals (which although slightly inaccurate were not penalised) rather than exact fractions. Only a few candidates realised that the total number of correct guesses was $2X$.

Question 4
Another popular question, with 25 attempts and a satisfactory average score of 13.2. Surprisingly, several candidates omitted to give a diagram of the Poisson (4) distribution, although most of the rest produced creditable bar charts (not histograms, which wrongly suggest the continuity of data between integer values). The required probabilities of events in this process happening over varying periods of time were generally well done. By contrast, only a few candidates were able to produce valid calculations of the conditional distribution of one component of a given sum of two independent Poisson variates. Of these, most found the required probability by calculating three Poisson probabilities instead of the easier binomial (20, 0.25) form to which they cancel.

Question 5
A deeply unpopular question (only 9 attempts), but most of these addressed at least the first two parts well and the average score was 11.9 out of 20. A serious weakness in the use of Bayes' theorem was exposed, in that very few candidates used the prior probabilities 0.1, 0.4 and 0.5 to arrive at the correct posterior probabilities of the events A, R and S. In the final part, many candidates substituted $p = 0.1$ to find the dominant posterior probability of A in this case (0.9468), but very few showed that this probability was a monotonic decreasing function of p, so that for all $p < 0.1$ the resulting probability was even greater.

Question 6
A popular and high-scoring question (21 attempts, average mark 14.1), probably due to the large proportion of bookwork, which was, in the main, competent or well-remembered. However, only a few candidates bothered to show that the stationary point of the (log-)likelihood was a maximum, and there were few sound derivations of the asymptotic Normal-theory confidence interval.

Question 7
This very popular question was generally well answered, with 24 attempts achieving the highest average score (14.5 marks), mainly due to good work on parts (i), (ii) and (iv). However, in part (iii) the conditional distribution of X caused much confusion, with many candidates failing to normalise their answers. Part (v) revealed that many candidates are still under the wrong impression that uncorrelated random variables are necessarily independent.

Question 8
This highly popular question was generally done well (23 attempts, average score 14.0). However, several inappropriately scaled graphs would have been improved by telescoping the distances from the origin to the data on both the x and y axes, and a few candidates were muddled in their use of formulae to calculate the product-moment correlation coefficient. The possible outlier arising from trainee I was noted by most, and generally sensible comments were made about the high leverage of this individual and his effect on the analysis.

Higher Certificate Paper II – Statistical Methods

The Statistical Methods paper aims to examine a candidate's understanding of the fundamental concepts of statistical analysis through questions involving estimation and hypothesis tests. Particular emphasis is placed upon assessing candidates' ability to summarise and interpret the results of statistical analyses.

Candidates are much better at obtaining the summarised results of analyses than at interpreting these results and at drawing appropriate conclusions. In general, candidates demonstrate an adequate grasp of the basic techniques required when performing a range of statistical tests and are good at calculating basic descriptive statistics. Candidates are poor at explaining the meaning and uses of statistical tests in general terms and have great difficulty in correctly interpreting the results of their statistical analyses. Often sections asking for results to be commented upon or reports written summarising findings are answered very vaguely – or are omitted entirely by some candidates.

As stated above, candidates are able to calculate descriptive statistics such as means, medians and standard deviations; however, the rubric for the examination states that ‘when a calculator is used the method of calculation should be stated in full’. As in previous years candidates continue to lose marks by just writing down the numerical values of means, standard deviation etc., presumably obtained from the statistical functions of their calculators without giving any associated working. Full marks can only be achieved when descriptive statistics are accompanied by the appropriate working such as the correct formula and/or an adequate explanation of how the value is obtained as appropriate.

Candidates should be encouraged to read the questions carefully. Although not as many as in recent years, some candidates continue to lose marks by not actually answering the question asked or waste time including additional information not asked for in the question. Also, some candidates omit sections of questions entirely.

Graphical presentation of data is generally untidy and poorly presented. Graph paper is not always used, axes not labelled and titles omitted.

On the whole candidates have a good grasp of the basic requirements of hypothesis tests and can calculate the appropriate test statistic needed. However, many have difficulty in presenting the analysis clearly, in drawing correct conclusions and in interpreting the results. Particular difficulties with hypothesis tests are as follows

- Selecting the appropriate statistical test to perform if this is not stated in the question.
- Listing and explaining the assumptions required for tests to be valid.
- A failure to state the null and alternative hypotheses. Many candidates conclude a question stating that the null hypothesis may be accepted or rejected without having stated what this is. In addition many candidates having accepted (or rejected) the null hypothesis fail to explain what this means in relation to the problem posed in the question.
- Confusion between one and two-tailed tests. Some candidates state a two-sided alternative hypothesis and then proceed to perform a one-tailed test and vice-versa.

A common error in all two-tailed hypothesis tests is to state that the significance level for the test is 0.05 and then obtain a critical value at the 0.05 significance level rather than the 0.025 level. Alternatively when performing a one-tailed test at the 0.05 significance level a critical value from tables at the 0.025 significance level is erroneously used.

Many candidates fail to give the values obtained from statistical tables. Some include statements such as ‘this test statistic is greater than (or less than) the value in the tables’ without stating precisely what the tabulated value is. (This may – just – be acceptable when the value of some well-known statistic is quite extreme or is very close to its expected value under the null hypothesis, but is otherwise bad practice!) Further, some candidates blandly state whether the null hypothesis should be accepted or rejected without giving an explanation for their conclusion.

Many candidates find it difficult to explain the conclusions that should be drawn when the null hypothesis is rejected or accepted in the context of the problem posed in the question.

Question 1

(i) Candidates appeared to have a good basic grasp of the different situations in which a two-sample t test and a paired sample t test might be used. However, candidates had more difficulty in explaining the reasons why one might choose to design an experiment using paired data; that is, that it reduces the amount of variability between the two sets of measurements. Some candidates lost marks by not 'using examples to illustrate' their answer as instructed in the question.

(ii) Most candidates understood that the appropriate analysis to perform was a two-sample t test, although not all stated the assumptions necessary for this test to be valid. The summary statistics for each group were generally correctly calculated although some candidates lost marks by obtaining standard deviations using the statistical functions of the calculator without showing any working or the correct formula. Not all candidates were able to calculate the pooled sample variance correctly and some attempted to perform a t test for the situation in which the two population variances cannot be assumed to be equal. However, they were unable to calculate the correct degrees of freedom. Note that the two-sample t test when equal variances cannot be assumed is not currently on the syllabus and candidates are not expected to have to perform this during the examination.

Other common errors included omitting the null and alternative hypotheses, using an incorrect number of degrees of freedom or stating that the test statistic was larger than the tabulated value without giving the value obtained from tables.

Question 2

(a) This was generally well answered, with the majority of candidates being able to state a definition for each of the four statistical terms. Some candidates did have difficulty in explaining the meaning of the definitions they had given especially for 'power' which caused the most difficulty for candidates. Many could give a textbook definition for power without clearly demonstrating that they understood what this meant. Some candidates erroneously stated that the Type I error was the probability of rejecting the null hypothesis when it is true. Similarly that the type II error was the probability of accepting the null hypothesis when it is actually false.

(b) Provided a candidate was able to identify that a one-sample t test was required to investigate the problem, both parts were fairly well answered. Candidates who lost marks did so because of a failure to include all the necessary details and explanations as listed in the general comments concerning hypothesis tests. Very few candidates were able to identify that the increased power due to the increased sample size in (ii) was responsible for the differing conclusions of the two tests

Question 3

Very few candidates attempted this question. It may have appeared unattractive as it is open-ended and requires candidates to explore the data and produce such statistics and diagrams that they consider to be appropriate to support their description of the main features of the data. These tasks cause problems for many candidates and many choose to avoid them.

A common error was to fail to comment upon the use of current prices, rather than the more helpful constant prices, in the table of data. Analyses influenced by this distortion were not acceptable.

Question 4

(a) The statement of the model was well done but many candidates failed to state all the necessary assumptions for the model to be valid. Very few candidates stated that the effects in the model are additive and that the observations come from a Normal distribution.

(b) Most candidates can correctly construct a two-way ANOVA table and perform an F test for differences between treatments and blocks. Some candidates correctly performed two hypothesis tests from the ANOVA table (one for gravel and one for cement) but only provided one null and alternative hypothesis, generally for differences in cement. Many failed to state any null and alternative hypotheses at all. A common weakness was to obtain the test statistic from the ANOVA table and then proceed to state whether or not this gave rise to a statistically significant result without giving any explanation of how this conclusion was drawn. When comparing the value of a statistic to an F distribution, it is important to refer to critical values obtained from tables, the appropriate degrees of freedom, and the significance level; all of these are normally required for full marks to be obtained.

The question is open-ended, but many candidates concluded the analysis after completing the F test. However, to obtain full marks candidates were expected to explore the data further as statistically significant differences between the groups exist. For example the least significant difference could be obtained or one or more pairwise comparisons as appropriate. Many candidates lost marks unnecessarily by not including a report on the findings for the manager.

Question 5

(i) Histograms were often very untidy and graph paper was not always used. Full marks were not awarded to histograms that were not drawn on graph paper. As in previous years many candidates did not include titles and although axes were labelled these were often incorrect, especially on the y-axis, which many candidates labelled as ‘Number of trees’. This occurred on scripts from candidates who correctly understood that in a histogram the area of each ‘bar’ represents the frequency of the group and not the height. Some candidates continue to misunderstand this and continue to represent the frequency of each group by the height of the ‘bars’, although there seemed to be fewer candidates making this mistake than in previous years. There were some candidates who appreciated that the width of the bars should increase according to the range of values it covers, but unfortunately did not adjust the height of the bar so that the total area represented the frequency of the class. Very few candidates indicate how the frequencies are represented on the histogram (for example, 1 cm square represents 5 trees). Remembering to include this may help candidates to understand more fully that frequency is associated with area and not height.

Many candidates were uncertain how to deal with the open-ended category. Candidates were expected to choose a suitable end-point such as 70 and draw the histogram and calculate the summary statistics with reference to their chosen value. Any sensible value was accepted. Full marks were not awarded for histograms that did not include this class.

(ii) Most of the marks lost in this part of the question were either for careless arithmetic errors or for incorrect statements of the formula for the standard deviation. Some candidates were unsure of how to deal with the final open-ended class and erroneously ignored it entirely when calculating the mean and standard deviation.

(iii) Candidates had difficulty in explaining the differences between the estimated and actual values. Of course, the estimates are based on a presumption that the data are evenly distributed in each interval, which need not be the case. Some candidates erroneously stated that the median was affected by the open-ended category, which it would not be.

Question 6

(i) Generally candidates were able to make the link between the experimental situation and a geometric distribution, but were not always able to explain fully all the necessary assumptions and how they might be assumed to be satisfied here. In addition, some candidates lost marks by not explaining why the probability of a success in each trial, p, might reasonably be assumed to be $^1\!/_3$.

(ii) Most candidates understood that the appropriate analysis to perform was a chi-squared goodness of fit test and demonstrated a good basic grasp of this test. The calculated probabilities and expected values were generally correct, although some candidates lost marks by not explaining how the expected values were obtained from the calculated probabilities. When calculating the test statistic, many candidates mistakenly combined all the cells in which the observed count was less than 5 (rather than all the cells in which the expected count was less than 5). Marks were also lost by failing to give the null and alternative hypotheses, using the incorrect number of degrees of freedom, or misunderstanding when to reject or accept the null hypothesis.

Question 7

(i) This was generally well answered.

(ii) In both parts of the question many candidates were unable to establish whether a one- or a two-tailed test was required. When asked to test for an effect it is more usual to perform a two-tailed test since a significant increase or decrease in the outcome variable would constitute an effect. If one is only interested in, or asked to investigate, a change in one direction, a one-tailed test is used. If candidates chose to perform a one-tailed test and gave a good explanation why they had done so, full marks were awarded for correct solutions. (For example, the following argument was perfectly acceptable: ‘The purpose of the educational programme is to help overweight women to lose weight. It is only effective if it reduces weight, so a one-tailed test investigating whether the intervention leads to a significant reduction in weight is required.’)

(a) Only a small number of candidates were able to perform a sign test successfully. Most candidates were able to calculate the test statistic using a binomial distribution with $n = 14$ and $p = {}^1\!/_2$. However many were uncertain of the critical value and when to accept or reject the null hypothesis. Some candidates chose to use a Normal approximation to the binomial distribution to obtain the test statistic. Provided that candidates were able to explain why this approach was valid, full marks were awarded to correct solutions. In general, however, candidates using this approach were not very successful.

(b) Candidates were more successful in carrying out a Wilcoxon signed-rank test than a sign test and were usually able to obtain the correct test statistic. However, a number of candidates did not ignore the sign of the difference when calculating the ranks, with all negative differences incorrectly being given the lowest ranks. Many candidates lost marks by obtaining the incorrect critical value from tables for the Wilcoxon signed-rank test or by not understanding when to reject or accept the null hypothesis. Some candidates used a large sample Normal approximation for the test. When the sample size is small (less than 20, say) this method does not generally provide a good approximation and full marks were not awarded to candidates taking this approach. Most candidates who had drawn the correct conclusion from the two tests were able to identify that the differing conclusions were due to the increased power in the Wilcoxon signed-rank test as it uses more of the information in the data than the sign test.

Question 8

(i) The box-plots drawn were often untidy; not all included a scale on the axes, and some were not drawn on graph paper. Many candidates had difficulty in calculating the upper and lower quartiles as this required interpolation between two values. Surprisingly, many candidates correctly identified that the lower quartile would be represented by the 5.25th smallest data point and then concluded, without comment, that the lower quartile was given by 79,950, the value of the 5th smallest data point.

(ii) Most of the problems encountered with this question were due to the arithmetic rather than the statistical methods required. As the numbers involved are large, the number of digits involved in the calculations is more than can be displayed on some calculators. Candidates needed to perform the arithmetic either using exponential format or by expressing the values in more convenient units (say in units of £1000). Some candidates failed to identify that the appropriate test to perform was an F test to compare the equality of the variances of the selling price and erroneously performed a t test to compare the means.

Higher Certificate Paper III – Statistical Applications and Practice

The aim of the Statistical Applications and Practice syllabus is to develop skills in data analysis, using the theoretical concepts developed in the syllabuses for the Ordinary Certificate and Papers I and II of the Higher Certificate, to analyse real data sets and communicate the results comprehensively. The questions on the examination paper require candidates to select and carry out appropriate statistical procedures and to report the findings and conclusions clearly. Candidates are also expected to be able to interpret computer output from statistical packages. Detailed knowledge of specific packages is not required.

Candidates need to realise that it is not sufficient just to do the computations in the questions. They need also to state conclusions from their analyses and to attempt the discussion parts of questions. They should pay attention to the numbers of marks allotted to parts of questions to check that they have attempted a reasonable proportion of a question or of a part of a question. They should also avoid spending time calculating quantities which are given in the questions, unless instructed to do so.

Some candidates tend to round too early in their calculations which introduces inaccuracy.

In tests of hypotheses candidates should explain why something is or is not significant (for example, by quoting critical values). If they do not it is not possible to tell whether they have guessed. A possible exception is when a value is clearly not significant or highly significant, for example, $z = 0.2$ or $z = 50$; in such cases a statement such as 'clearly (not) significant' should be made. This comment refers to several questions.

Question 1

This question on a 2-way ANOVA with replication was very popular and was mostly well done. Some candidates lost marks by not giving enough details of the ANOVA, including clear statements of hypotheses and an indication of how conclusions were reached from the F values. Few candidates attempted part (ii) and some of those who did drew incorrect diagrams. The twelve means should have been stated as credit was given for these even if the diagram was incorrect. Some candidates thought that part (iii) was asking for comments on the diagram, whereas comments on the whole of the analysis were expected. (This was indicated clearly by the fact that this part was identified separately, rather than just being a further requirement of part (ii).) Attempts at part (iii) were rather poor and several candidates did not attempt this part. It should be noted that parts (ii) and (iii) were together worth 50% of the marks on this question.

Question 2

This question on tests of hypotheses on the difference of means of two populations was very popular, but was not done particularly well. In part (a) the test could be based on the Normal distribution as the samples were large. The only assumptions needed are that the samples are large enough for the Normal approximation to hold and that the samples are independent. Some candidates mistakenly thought that the underlying populations had to be Normal and/or they had to have the same variance. Most candidates ignored the instruction in (a)(ii) to comment on further information which should be obtained to throw greater light on the impact of mobile telephones on student expenditure.

In (b) a Normality assumption is needed for a t test, and this should have been stated. Candidates appeared to be unaware that the pairing removed between-company variation. The commentator should have made the assumption that the samples are independent. Clearly they are not, so the method used was wrong.

More care needs to be taken in stating hypotheses. For example, 'H_0: $\mu_1 = \mu_2$' is meaningless unless μ_1 and μ_2 are defined. Equally 'H_0: there is no difference between males and females' is clearly untrue – mention of mean expenditure is needed. Some candidates appeared to be unaware of the difference between notation for samples and that for populations. Some found sample variances incorrectly. Using a natural notation, the sample variance is *defined* as $\frac{1}{n-1}\Sigma(x-\bar{x})^2$, but is generally most efficiently calculated as $\frac{1}{n-1}\{\Sigma x^2 - (\Sigma x)^2/n\}$.

Question 3

This question on simple regression was fairly popular and was mostly well done. In (i) some candidates were careless and used α and β for the estimators instead of $\hat{\alpha}$ and $\hat{\beta}$. In (iii), when interpreting r^2, the statement should point out that it measures the proportion of variation in GDP explained by a *linear* relation, in interpreting $\hat{\alpha}$ and $\hat{\beta}$ the units of measurement should be stated and $\hat{\alpha}$ should be related to the year. In (iv) an indication of how the position of the line was determined should have been given; for example, by calculation of two points on the line such as the values of y at $t = -5$ and $t = 5$. (Note that the value $t = 0$ is not a particularly good choice here, as it is in the middle of the given range of t.) Comments on the predictions in (v) were mostly poor.

Question 4

This question on time series was not popular and few candidates attempted all of it. In (i) many candidates did not correct the series for seasonal fluctuations as instructed, but just worked out the seasonal correction factors. Some of those who attempted this part apparently did not realise that the values in the corrected series have to be the same order of magnitude as the observed values as they made no comment when they were not. One or two candidates subtracted the correction to the mean quarterly values but should have added it as it is negative. Few candidates attempted parts (ii) and (iii). One or two suggested the exponential smoothing method in (iii) but this is not a method of seasonal correction.

Question 5

This question was based on computer output and involved drawing boxplots and writing a report. It was not popular. Some candidates calculated the quantities needed to draw the boxplots in (i). This was not necessary as these were given in the output. Outliers were not always clearly indicated in (i), and some candidates had some very strange ideas as to what might be an outlier. There was some confusion as to the direction of the skew for E&EEq and SS. As there is a long tail to the right, that is, the positive direction on a conventional graph, the skewness is positive. Attempts at (ii) tended to be poor or non-existent.

Question 6

There were hardly any attempts at this question on methods of sampling. Some candidates lost marks because it was not clear that they understood the method they were discussing. In some cases it was clear that they did not understand the method. Some candidates overlooked the fact that the questionnaires were to be sent by post.

Question 7

This question on estimation and a chi-square test was fairly popular, but was not done very well. In (i) and (iv), both of which involved finding a maximum, few candidates checked on the second order condition to confirm that they had found a maximum rather than a minimum. Insufficient care was taken with the setting out. For example, in (iv), when equating the derivative to zero to find the maximum likelihood estimator $\widehat{\lambda}$ of λ, the result needs to be expressed in the form $\widehat{\lambda} =$ ... rather than $\lambda =$ .... Similarly, in (iv), the expression for the ML estimator should be written as $\widehat{\lambda} = 2/\bar{x}$ and not as $\bar{x} = 2/\widehat{\lambda}$. In (ii), a sketch could be drawn by noting that $f(x) = 0$ when $x = 0$, and $f(x) \to 0$ as $x \to \infty$, and marking the position of the mode. An accurate diagram was not needed – this section was worth only 2 marks out of 20. Hardly anyone attempted (iii) or to find the moments estimator of λ in (iv). In (v) some candidates found the probabilities needed to calculate expected frequencies by integrating $f(x)$. This was not necessary as $F(x)$ was given in the question.

Question 8
This question on transformations and one-way ANOVA was fairly popular. Some candidates did reasonably well on it, but others scored low marks, often because they made only partial attempts at this question. Part (i) really required the statement of a model, either in words or symbols, as assumptions do not make sense *in vacuo*. In (ii) most candidates ignored the part asking if there was evidence of a better transformation. In (iii) the follow-on analysis to determine which herbicides differed from the others as regards their effect on the mean numbers of weeds was NOT required (though it was good to see that candidates were aware of this analysis). In (iv) a description of the plot under discussion was needed as there are several possible plots. A residual is not the same as the residual sum of squares. Some candidates appeared to be confused between a residual and the error term in the model.

Graduate Diploma Paper: Statistical Theory And Methods I

This paper examines probability theory – Bayes' Theorem, discrete and continuous random variables, univariate and bivariate distributions, transformations of random variables, simulation, order statistics, simple stochastic processes.

Overall, the standard of attempts at this year's paper was very good. There were two outstandingly good candidates, and almost all candidates seemed well prepared. One candidate answered just 4 questions, rather than the required 5, while another candidate attempted 6 questions. It was particularly pleasing to find that several candidates made excellent attempts at the question on Markov Chains (Question 8), a topic that has been largely ignored in the past two years' examinations.

Question 1
This question examined basic ideas about the cumulative distribution function, expected value and variance of a single continuous random variable, and transformations of a random variable, though the question was based around the concept of the survivor function. It was answered by only one-third of the candidates, and their attempts were generally poor. Several of those who answered this question attempted to find the expected values in parts (ii) and (iii) by direct integration of $x.f(x)$, as usual, in spite of the mathematical complexity of such an approach and the explicit instructions to the contrary which were written into the question.

Question 2
In this question, candidates had to derive the mean and variance of the beta distribution, then work with a bivariate distribution whose marginal distributions were also beta. Almost every candidate attempted this question. The general standard of their answers was very good.

Question 3

In this question, candidates were tested on their knowledge of transformations of two continuous random variables. About two-thirds of candidates attempted this question, and the standard of their attempts was good. Some candidates seemed to be confused about the definition of the Jacobian of a transformation.

Question 4
This question tested moment generating functions and the Central Limit theorem in the context of the exponential distribution. Virtually every candidate attempted this question. Their answers to part (i), in which they had to derive the m.g.f. of the exponential distribution and use it to find the mean and variance of the distribution, were generally good. In part (ii), however, several candidates appeared not to know how the m.g.f. of $(aX + b)$ can be found from the m.g.f. of X itself.

Question 5
This question examined order statistics. About two-thirds of the candidates attempted this question, with varying degrees of success. Several of those who made poorer attempts went wrong early on, by incorrectly remembering the general formulae for the p.d.f. of $U_{(1)}$ and the joint p.d.f. of $(U_{(1)}, U_{(2)})$. These candidates produced functions that they could not manipulate as required for the rest of the question.

Question 6
This question tested candidates' knowledge of the Law of Total Probability and Bayes' Theorem. Only two candidates attempted this question.

Question 7
This question was about simulation using the inverse c.d.f. method. Almost all candidates attempted this question, and the standard of their answers was generally good. Surprisingly, a few candidates did not seem to realise that the distributions defined in (ii) (a) and (b) were continuous and tried to simulate as though these distributions were discrete.

Question 8
This question tested work on Markov Chains. Half the candidates attempted this question; many of their answers were excellent, and the overall standard of their attempts was very good.

Graduate Diploma Paper: Statistical Theory And Methods II

The paper aims to test understanding of a range of statistical principles and methods, and their application in simple situations.

Questions 1–6 and 8 were the most popular, with these being answered by at least one-half of the candidates. Of these, Questions 3, 4 and 8 were answered well by at least two candidates. Question 7 was not popular.

Question 1
Only a few candidates could write down the correct likelihood function in part (i). No candidate gave a completely correct answer to part (ii). Several candidates did not calculate the relevant expected values in part (iii), and only a few gave the appropriate degrees of freedom.

Question 2
Few candidates could define a loss function and the Bayes risk, but parts (i) and (ii) were generally well done. No candidate was able to minimise the Bayes risk in part (iii) and there was only one correct solution to the uniform prior calculation.

Question 3

Only one candidate could relate the two correlation coefficients in part (i). In part (ii), only two candidates could properly derive the p-value. In contrast part (iii) was generally well done.

Question 4

Not many candidates could explain how sample size can be determined using the power. In part (i), few candidates obtained the correct critical region. Similarly, there were only a few correct answers to part (ii). Both parts (iii) and (iv) were poorly answered, with there being just two correct solutions.

Question 5

No candidate stated the key result that $\widehat{\theta} \stackrel{.}{\sim} N(\theta, 1/I(\theta))$ for large n. Part (i) was correctly answered by only a few candidates, with many confusing $E(\overline{X}^2)$ with $E\left(\sum_{i=1}^{n} X_i^2\right)/n$. In part (ii), several candidates did not write down the likelihood function in terms of θ. No candidate could derive $E(\widehat{\theta}^2)$ in part (iii), but there were some reasonable attempts at the other two calculations.

Question 6

There were few good answers to the bookwork. Only one candidate could obtain the distribution function in part (i). In part (ii), no candidate could explain why $W|\theta$ is a pivotal quantity. There were a few reasonable answers to part (iii), but none were completely correct.

Question 7

This was not a popular question. There were only three poor attempts.

Question 8

Part (i) was only answered well by two candidates. Few candidates gave completely correct answers to part (ii). Part (iii) was generally well done, but only three candidates gave correct solutions to part (iv).

Graduate Diploma Paper: Applied Statistics I

This paper is designed to test whether candidates can carry out a range of important statistical techniques (linear model, multivariate analyses and simple time series) in practical situations. This requires a grasp of basic theory and the ability to apply it and to interpret computer output.

All candidates followed the rubric, although some candidates who submitted four good answers did much better than those who did five sketchy answers. Apart from careless mistakes the main problem was a failure to answer the question that was asked. Most of the questions were based on a case study. The applied statistician needs to use information about the nature and background of data when modelling: automatic mathematical-type approaches are not sufficient. Candidates tended to repeat bookwork without relating it to the specific question being asked, and so attracted few marks. Even relatively simple questions which asked for descriptions of graphs or simple summary statistics were not done well. In contrast, those candidates who did well submitted answers that showed a good grasp of both basic theory and of the practicalities of data analysis, they were able critically to discuss the relevance of methods for a particular set of data, and produced thorough descriptions of data and comprehensive interpretations of output.

Question 1
This was very popular – most candidates attempted it. It was generally reasonably well-done, although the working was often messy and careless. There is usually a question on time series, so it is important to be confident of basic definitions and simple manipulations of series. However, working must be accurate.

Question 2
This was not popular, although candidates who attempted it made a reasonable attempt at the parts they answered. The study described is observational and not a true experiment, so there are potential confounding factors for the feedback method, especially prior ability. ANCOVA is often suggested as a way of dealing with such studies. It is important to appreciate what ANCOVA does, and how it differs from a t test. ANCOVA does not necessarily address all the problems, and it can be difficult to interpret if the slopes of the regression lines are different for the two groups. The forms of the different linear models demonstrate this, and the question is based on the relative merits of different types of analyses for these data. The aim of the question was to explore the theoretical basis of ANCOVA and its practical application.

Question 3
This was quite popular and reasonably well done. Most candidates interpreted the plots correctly and were able to discuss how they might tackle problems revealed by such plots. The major weaknesses were that candidates gave sketchy descriptions of the plots – you should describe all that you see – and later answers were general as opposed to tailored to the specific case studies described. Good answers were those that used general knowledge (for example, about geography) to make suggestions, and not merely repeat bookwork about diagnostic plots.
It is important to answer the question as asked.

Question 4
This was not popular, although it was fairly straightforward. It is based on a real consultancy problem posed to a statistician, although the application area is disguised. There is no evidence of a relation between the initial grading and the success rate. This should be fairly clear in the graph, and the formal analysis confirms this. Do not be afraid to conclude that there is no evidence of a relationship – this is what you often see in practice. Some candidates over-interpreted the graph, and others forgot that the null model **is** a model to be considered. In this case the null model is the best, which confirms the impression from the graph.

Statisticians might disagree about whether a formal analysis is appropriate given that the result is ‘obvious’ from the graph. The last part of the question asked for a comment about this. Any reasoned answer would have been accepted. In contrast, the statement for the manager must be non-technical and certainly **not** use terms like ‘scaled deviance’ or ‘forward selection’.

Question 5
This was popular, but few candidates grasped the point of the question. Principal component analysis is often advocated as a way of reducing the number of predictor variables and of removing multicollinearity. If the principal component analysis is performed on the predictor variables only, the selected components may not be strongly correlated with the response variable. This is what happens in this question which contains some odd results worthy of comment.

The problem as posed is very difficult: there are many more potential predictor variables than observations. While this should not happen it does! Neither forward selection nor use of principal components yields a well-fitting model. However, there is a set of variables that gives a perfect fit. This particular combination of variables is not encountered in forward selection and is not suggested by the principal component analysis.

You are not told what the variables are, and this makes interpretation of the principal components (which has a degree of subjectivity) and choice of model difficult. Description of the correlation structure in the data was reasonably well done, but many people failed to try to interpret the principal components. Most lost the point of the question and amazingly some candidates were suggesting in the last part that principal component analysis might help to identify predictor variables even though it was perfectly obvious that this method does not help here.

A question like this tests that you know what happens in practice – bookwork is not sufficient.

Question 6
This was a mixture of practice and theory. It was popular, but few candidates did well in both parts of the question. In contrast to Question 5 you are told here about a set of variables for regression. In practice, automatic variable selection is dangerous, and one should take into account the nature and quality of the variables as well. The first part of the question asked about this and tested whether candidates could show a true modelling approach combining background knowledge about variables with interpretation of formal analyses.

The second part of the question was bookwork. Those who could replicate the bookwork did well here.

Question 7
Few candidates attempted this question.

Of those who did there was a tendency to reproduce theory without relating it to the case study described. For example, the question does not ask what discriminant analysis is, but how it might be used in this problem. General answers unrelated to the case study attracted few marks.

Summary statistics were given, and the question asked for descriptions of them. This was not well-done in general. The applied statistician **must** be able to describe what he or she sees in the data. This should not be difficult.

Discriminant analysis is a multivariate method, which by definition uses more than one variable. It can therefore be more powerful than univariate approaches. In this example there is one variable that appears to model the group membership as effectively as a multivariate model. This is the point of the question. The question required an understanding of the purpose of discriminant analysis, and the ability to describe and then interpret fairly simple output. Few candidates could do this.

Question 8
This was very popular, but not always well done. You need to be clear about the hypotheses being tested, and how to interpret results. The correct form of the model is decided during the design stage and not based on the data. In fact the conclusions are somewhat different for the two scenarios.

Working was often untidy and careless, which did not help candidates.

Graduate Diploma Paper: Applied Statistics II

The Applied Statistics Paper II syllabus covers the application of statistical methods to censuses, surveys and designed experiments, and some elementary topics in demography. A total of 31 candidates registered for and sat the paper.

Overall, the performance of candidates sitting AS Paper II was disappointing. Nine candidates obtained fewer than 30 marks, and it is hard to view them as having been adequately prepared to sit the paper.

General strengths of candidates included knowledge of how to obtain the analysis of variance for different types of experimental design, for example block designs and factorial layouts.

Weaknesses included a lack of knowledge of basic principles of experimental design (for example, confounding, bias, randomisation, blocking); use of incorrect standard errors for treatment comparisons (for example, in the presence of missing data or for incomplete block designs); confusion over the appropriateness of using critical values of the t distribution or Normal distribution.

As in previous years, just over one third of the candidates attempted the question on response surface design and analysis. This is one of the most poorly understood areas of the syllabus. Fewer candidates answered the sample survey questions. Candidates were not familiar with cluster sampling, and the different estimators that are widely used. Standard formulae for estimating the population proportion based on simple or stratified random sampling were not known. Most candidates managed the more practical parts of questions on considerations for designing a survey and/or constructing a questionnaire.

Candidates answered the correct number of questions, and adhered to the rubric.

Question 1 (25 attempts)
Candidates were asked to comment on the weakness of an experimental layout in which the effects of treatment and litter were confounded. Most candidates noted that the better layout would be to regard litters as blocks, and for each mouse within a litter to receive a different diet – but could not always explain why.

Very few candidates correctly demonstrated how treatments would be randomly allocated in a randomised block design. There was also some confusion over the purpose of randomisation and blocking in designed experiments.

Section (iii) required candidates to obtain estimates of two items of missing data by finding those values (of m and n) which made the corresponding residuals zero. Only one candidate attempted this, giving a correct expression for the residuals. All other candidates deviated from the question and showed the required outcome using an iterative procedure, losing a few marks.

Most candidates knew how to construct confidence intervals but lost marks due to using incorrect standard errors or an incorrect number of degrees of freedom for the residual error (i.e. 2 df lost due to missing data).

Question 2 (26 attempts)
Part (a) asked candidates to obtain the analysis of variance for an experimental layout consisting of four replicates of a 2^3 factorial design. The effect estimates were given as part of the question.

Most candidates did not use the information provided to calculate the sums of squares for the main effects and interactions. The sum of squares could be obtained by squaring each effect estimate, and multiplying by $2r$. Some candidates divided by $2r$. Others lost time by re-analysing the data using Yates' method. Often the sum of squares for differences between greenhouses was omitted from the analysis of variance. Not all candidates were familiar with the concept of confounding, and seemed to confuse this with fractional factorial designs. Only a few candidates suggested an appropriate design (using complete or partial confounding) for four replicates of a 2^3 factorial design arranged in 8 blocks of 4 units.

Question 3 (20 attempts)
In part (ii), candidates correctly stated the conditions necessary for a balanced incomplete block design to exist but were less familiar with how to apply these conditions. A necessary condition for a BIBD to exist is that λ, the number of times each pair of treatments appears in the same block, is an integer. Thus, a BIBD does not exist for 10 treatments in blocks of 4 units, using not more than 90 units in all.

Candidates lost marks in part (iv). Standard formulae for the variance of the difference between two treatment means ($k\sigma^2/t\lambda$) and the residual degrees of freedom ($rt - t - b + 1$) were not known, and thus candidates had difficulty in comparing the three BIBD designs.

Question 4 (12 attempts)
This question required knowledge of 2^2 factorial designs and their application as first-order designs in response surface methodology.

There was some confusion over the design in part (i), with answers including: 2^3 factorial design, composite design, single factor experiment. Most candidates could explain how the steepest ascent procedure worked but could not apply the procedure to real data in part (iii).

More candidates than in previous years were able to calculate the coefficients of a first-order model. The easiest approach is to analyse the data as a 2^2 factorial design (with coded factors ± 1), and use the fact that the coefficients of the first-order model are one half of the corresponding factor effect estimates.

None of the candidates attempted part (iv).

Question 5 (20 attempts)
Some candidates thought systematic sampling and/or quota sampling were equal probability selection methods. Neither sampling method is random (i.e. element of subjective choice).

Candidates were able to distinguish between ordinary stratification and post-hoc stratification but were not aware of the consequences of post-stratification on the precision of estimators.

Part (iv) was often omitted or poorly answered. Candidates were not familiar with how to determine sample size to satisfy (i) a given standard error and (ii) a range of values for P, the population proportion. Often, the formula for the variance of the estimator of P was not known.

Question 6 (14 attempts)
This question required a basic knowledge of cluster sampling and how to calculate different estimators of the population mean, and their standard errors. Most candidates were not familiar with how to obtain estimators based on the cluster sample ratio or cluster sample total.

Question 7 (18 attempts)
Most candidates found part (a) difficult. Many candidates could not write down an expression for a stratified estimator of the population proportion. Standard formulae for the variance of estimators were often incorrect in part (ii).

Some candidates lost marks due to misreading part (ii), and thought values of p_m and p_f in (a) referred to Design A and those in (b) referred to Design B.

Question 8 (16 attempts)
Candidates were asked to compare and comment on the crude and direct / indirect rates. Often comparisons were made between direct and indirect rates, without reference to crude rates and the impact of standardisation.

Most candidates could define and apply direct and indirect methods of standardisation but failed to state the assumptions underlying its use.

Graduate Diploma Option: Statistics for Economics

Many candidates seemed inadequately prepared for this Option, with marks so low that they had little chance of passing the paper as a whole.

Candidates were too ready to go through familiar routines without pausing to consider whether they were applicable or valid in the particular contexts of the questions.

They were not willing or able to consider the effect of the economic contexts of the questions, treating the problems as mere numerical exercises.

Question A1
The first regression is unsatisfactory for two obvious statistical reasons and one economic reason. The standard errors of the coefficients are very large, pointing to probable multicollinearity among the E variables. The DW value is utterly unacceptable – there is surely some strong autocorrelation. The signs of the E coefficients make no sense economically.

Two of the Es are far from being significant, and are dropped in moving to the second regression, which makes good sense. The standard errors have dropped to acceptable levels, but DW is even worse.

The third regression is a re-estimation of the model in (b), using first differences. This has proved very successful in terms of the standard errors and of DW, and makes good sense. But the possibility of an exogenous change in $GFCF$ is not considered, so (d) is tried, adding a constant. A constant in a model for ΔF is equivalent to a linear time trend in F, and a negative coefficient is no surprise. It is not statistically significant, and a case can be made out for either (c) or (d) as the preferred regression.

The total sum of squares $\Sigma(\Delta F - \overline{\Delta F})^2$ can be inferred from (d), and together with $rss = 14034007$ from (c) the R^2 for (c) can be obtained.

The quarterly dummies, or their first differences, cannot be tested (meaningfully) one at a time. One has to re-estimate the model omitting all three and do an F test on the two residual sums of squares.

Question A2

Candidates should be well practised in some standard routine for doing the first three parts of this question, getting

$$\begin{array}{llllll} C & = & \underset{(0.3340)}{0.2537} & + & \underset{(0.0448)}{0.9178Y}, & r^2 = 0.9588, \quad s = 0.2287 \end{array}$$

$\widehat{C}(Y = 9.40) = 8.881$

95% Prediction Interval is 8.661 to 9.101.

95% Confidence Interval is 8.353 to 9.409.

Given $Y = 9.40$ and $C = 9.20$, one would therefore conclude that the previous, estimated, relation held, and that consumption was unusually high for unexplained (stochastic) reasons.

It is standard practice to denote income by Y, even when it is acting as a predetermined (right-hand-side) variable. Over hasty candidates sometimes used wrong formulae out of force of habit.

The last part of the question was badly done. It does not suffice to confirm that 0.2537 is not significantly different from zero and 0.9178 is not significantly different from 1. A joint test is required. Since $\Sigma(C-Y)^2 = \Sigma C^2 + \Sigma Y^2 - 2\Sigma CY = 3.5883$, the F statistic is

$$\frac{(3.5883 - 0.941451)/2}{0.941451/18} = 25.303.$$

Under the null hypothesis, this has the $F_{(2,18)}$ distribution. The result is very highly significant, and the null hypothesis must be rejected.

(It is readily found that the residual sum of squares about the fitted regression, given by the expression $(1-r^2)\Sigma(c-\overline{c})^2$, is 0.941451.)

Question A3

(a) Verbal explanations could have been fuller and more practical, and especially more related to estimation of economic data in particular.

(b) For a uniform sampling fraction with $n_1 = 100$, $n_2 = 60$, $n_3 = 40$, $\widehat{\mu} = 0.5\widehat{\mu}_1 + 0.3\widehat{\mu}_2 + 0.2\widehat{\mu}_3$. Since the sampling in the various strata is independent,

$$\begin{aligned} \text{var}(\widehat{\mu}) &= 0.25\ \text{var}(\widehat{\mu_1}) + 0.09\ \text{var}\ (\widehat{\mu_2}) + 0.04\ \text{var}(\widehat{\mu_3}) \\ &= 0.00492329 + 0.00374755 + 0.00291743 \\ &= 0.01158827 \\ \text{so } SE(\widehat{\mu}) &= 0.1076. \end{aligned}$$

This is smaller than 0.1112 for simple random sampling, but not greatly.

The three strata do not differ enormously in these means, so getting the ‘right’ proportions from each of them is not all that helpful. If the means had been 1, 21 or 41 (say), with the same variances, stratification would have been more helpful.

It was not necessary to remember any complicated specialist formulae. An understanding of basic principles was sufficient to answer this question. The practical benefits of stratification are often exaggerated, as this example demonstrates.

Question A4
Investment trusts can be expected to have lower means and much lower variances for their yields than do other companies. Lower means because risk-averse investors welcome the safety and security of investment trusts as compared with other companies and because of the administrative costs of running trusts. Lower variances because of the averaging nature of holding wide varieties of shares in portfolios. A little thought would therefore have suggested that one-tailed tests would be appropriate, but many candidates used two-tailed tests without explanation, out of sheer force of habit.
The first test is a simple F test of the two variances, and gives a significant answer.

The second test can be done as a t test or, equivalently, as a one-way analysis of variance. Both methods assume Normality and equal variances. The data are necessarily truncated at zero, and show positive skewness, so Normality is implausible. The F test rejected the hypothesis of equal variances.

The usual non-parametric test would be a Mann-Whitney (or Wilcoxon) test. This is sometimes thought of as a test of the hypothesis that the medians are equal, but formally it tests the hypothesis that the distributions are identical in every respect.

Graduate Diploma Option: Econometrics

The overall standard was not high, but about a third of the candidates (7 out of 20) did well on this Option. One major general remark is that many candidates did not attempt all parts of some questions.

Question B1 (7 attempts)
This was the least popular question but with the second highest average mark. Despite its appearance, this is not a difficult question and most candidates did well. Some candidates, however, did not distinguish between the distribution and the density functions.

Question B2 (16 attempts)
This was the most popular question. Although it was meant to be a relatively easy question, and was rightly detected as such by most candidates, the attempts to answer the question were disappointingly poor (average mark 7 out of 20). Although most candidates did answer the first part satisfactorily, in the second part only a couple of candidates pointed out that the parameters were elasticities. Only one candidate answered the last two parts of the question correctly. Other candidates failed to use the only possible formula available from the information in the question. (The restricted and unrestricted residual sums of squares should be used; these are obtained from the variance and not from either of the R^2 values,)

Question B3 (10 attempts)
This question had the lowest average of the paper. A handful of candidates answered the first two parts satisfactorily. No candidate answered the last part correctly. This is mainly due to the fact that the majority of candidates did not know or remember the 2SLS formula in matrix form, and none could correctly translate the data given in the question into 2SLS estimates.

Question B4 (15 attempts)
This second most popular question had the highest average. This is not surprising given its flexibility. Almost all candidates did well in this question. However, the ‘instrumental variables’ and ‘selecting appropriate set of regressors’ parts were not properly understood, especially the latter.

Graduate Diploma Option: Operational Research

The performances of the five candidates for this Option were very mixed. Three candidates did pretty well. They seemed well prepared and, by and large, to have understood and applied the correct methods, despite one or two arithmetical errors. However, the other two candidates performed very badly, even on the purely numerical questions. One candidate barely wrote 3 sides and did not seem to be at all prepared or to have any real understanding.

Question C1
Critical path and network analysis: the network part (a) was done well by both candidates, although there were some small arithmetical errors in calculating the ESTs and LETs. (Candidates were not penalised heavily for arithmetical slips.) They also tended to use an excessive number of dummy activities. However one candidate did not appear to have read the question properly for the second part (b) and the other candidate made a totally unjustified (and incorrect) assumption about the critical path not changing, so this part was not answered well by either candidate.

Question C2
Linear programming: surprisingly, no candidates did part (a) particularly well, given that it was a very standard problem. Mostly, they did not use enough artificial variables and some people did not realise they needed to use a 2-phase approach. Marking was generous in respect of the arithmetical slips which can easily occur with this type of question, but all candidates made some fairly fundamental errors. Part (b), the transportation problem, was answered well by just one of the three who attempted it. Another candidate attempted to formulate it as an LP, which clearly showed a total lack of understanding of the method.

Question C3
EOQ question: both candidates made rather heavy weather of part (a), but mostly the basic ideas seemed well understood and applied. The second part (b) was also quite well answered although one candidate worked it out from first principles instead of using the standard formula, and thus unfortunately wasted time and was unable to complete Question C4.

Question C4
Simulation question: This question was the most popular and was attempted by all five candidates for this Option. One candidate gained full marks, while two answered it fairly well. However, the answers of two of the candidates were extremely poor. Part (b) required some repetitive numerical calculations and again the examiner was prepared to ignore minor errors, as long as the candidates clearly showed they knew what they were doing.

Graduate Diploma Option: Medical Statistics

The number of candidates was small, and a meaningful report could not be compiled.

Graduate Diploma Option: Biometry

Question E1
The question of how to deal with missing values in experiments has of course developed rapidly with the use of computer packages. The syllabus aims to test knowledge of the basic methods which can

be applied, some of which are still perfectly accessible by hand. The formula method is old and well known; it can be extended to two or more missing values by an iterative method, using a first guess at one of the values and with the aid of this estimating the other. The next step is to put this estimate into the data and use the formula again to find a better figure to replace the initial guess. These steps are repeated until convergence is approached. Covariance using a dummy (0, 1) variate is described in various books, and can be used for as many missing values as required, given a standard multiple regression program.

Question E2
The major problem using a model of the form specified is that least-squares equations require knowledge of some of the parameters before they can be solved for the others. However, after a graph has been drawn carefully, and a smooth curve fitted as well as possible, two estimates can be found from asymptotes as x approaches zero and infinity. Thus two parameters can be first-guessed and this forms the basis for estimates of the other two, so the process can be repeated. If a logistic model is suggested as an alternative, it is certainly a possibility but is less satisfactory if the curve is not symmetrical. The method in (a) needs a transformation of the original equation into a form that can more easily be solved, and so the usual assumption of Normally distributed residual terms in the model is no longer made. Any statistical tests of fit will only be approximate. The ‘Message’ points out the least well-fitting data item, but it is in the middle part of the curve and there is no other evidence of difficulty with the fitted model.

Question E3
Fieller’s Theorem is a standard technique for dealing with ratios of Normally distributed random variates. Bioassay texts, in particular, explain it. But it is not limited to bioassay and the second part of the question applies it to the ratio of two linear functions of Normally distributed random variates which arises in finding a change-point.

Question E4
This question gave the opportunity for anyone with practical experience to show this in two chosen areas of agriculture or biometry. As emphasised in the question (and in the syllabus), general comments at a non-specialist level are not given marks unless they are also applied in a suitable context.

Graduate Diploma Option: Statistics for Industry and Quality Improvement

Question F1
The median of the absolute difference between X_1 and X_2 corresponds to the upper quartile of the distribution of $X_1 - X_2$. The shifts follow one another immediately, but the chart of the 21 consecutive observations reveals a jump in mean at shift-changes. Instructions for finding the estimate of σ are in the question, and its value is used in a standard type of chart.

Question F2
The standard Shewhart Chart is required in (a), and in (b) a Poisson process is appropriate.

Question F3
This is standard Queueing Theory.

Question F4

Knowledge of experimental design is required for this question, with some basic ideas of response surfaces. The design is Resolution 3. Aliases of G are CD, AF, BE. Because of the orthogonal design, G is estimated independently of the other factors. Half-Normal plots are explained in various books.