ROYAL STATISTICAL SOCIETY EXAMINATIONS, 2003

REPORTS OF EXAMINERS

General comments

Most of the comments made by RSS examiners concern individual questions, or overall strengths or weaknesses of candidates on particular sections of syllabuses. These comments will be found in the later sections of this report. Every year, however, there are some comments on examination technique which are not specific to individual papers. This year, many examiners included general advice in their reports, writing in very similar terms. We feel it might therefore be helpful to candidates and to those preparing candidates for RSS examinations to collect comments of this sort into this separate section.

The most frequent comment concerns a matter which seems so obvious as hardly to need stating: **candidates should take great care when reading the questions**. We are sure that candidates are always told to do this, but every year we find that some candidates fail to follow instructions clearly given in the questions.
This year, one examiner wrote:

> ‘Some candidates did more than was requested in the question, and some answered a different question from that set. In an examination where there is a time constraint it is important not to spend time on extras for which no marks will be available, and a right answer to a wrong question receives no credit. It is also important to pay attention to the breakdown of marks given on the paper. Time spent writing a long answer to a part carrying a small number of marks is time wasted, and producing a very short answer, or no answer, to a part carrying a high number of marks is poor examination strategy.’

Another examiner complained that

> ‘a few candidates ignored some of the questions altogether and made up their own’.

It may help candidates if they always bear in mind that examiners allocate marks in detail to the different sections of a question. Marks are then awarded in accordance with this marking scheme, so that marks are available only for answering the question set in the paper. If a candidate (perhaps through reading a question too quickly) fails to notice a critical feature, the answer may well be irrelevant to the question set, and no marks will be available.

While examiners are keen to give candidates credit for relevant knowledge, they are disappointed when they see that candidates who clearly have some knowledge do not show that they can answer the question actually set.

Every year, a few candidates answer so few questions that they have no realistic chance of passing. While this can obviously result from inadequate preparation, we would emphasise to all candidates that it is usually good examination technique to attempt as many questions as the rubric allows. Clearly, once you find a question on a topic you like, it is sensible to persevere with it and try to answer it completely. But, in mathematical subjects, many examination questions start with relatively straightforward parts, with harder material later. So if, in an examination in which you may attempt 5 questions, you find yourself having attempted only 4 and have 15 minutes left, you may well do better by tackling one or two easy parts of a fifth question than by struggling with the final section of one of the first four.

Examiners realise, of course, that many candidates are under strain when taking examinations, and make allowances for this. But, every year, we find that some candidates simply fail to follow the instructions on the front covers of the answer books. We realise that candidates won't want to spend time during the examination reading the front cover, so we have produced a copy you can see at the URL

`http://www.rss.org.uk/exams/docs/examcover.jpg`

on the Society's website. You are strongly encouraged to look at this before the examination, and to ensure that you follow the instructions.

Ordinary Certificate Paper I

The examiner's intention has been to test the ability to apply statistical concepts at an appropriate level. This year, many candidates were able to do this quite well. What was important was to think about the statistical ideas and techniques in terms of the situation described in the question. A multi-stage sample, for example, would be pointless for a population whose names were all held on a Council database and who all worked at a single location.

On actual technical ideas, there is, perhaps, still some haziness on the basic difference between 'bias' and 'sampling error'. Bias is a *systematic* tendency to overestimate or underestimate some parameter, whereas *error* can be purely random. Thus, for example, stratified or simple random sampling are both unbiased, but stratification usually (on the assumption that strata tend to be more homogeneous) increases the 'precision' and decreases the variation between one sample estimate and the next. Candidates should also realise that if an examiner uses a term like 'random sample' (as in Question 2 (method B)) then he or she actually *means* a sample selected in a carefully controlled random manner (see standard textbooks for details). Though in everyday usage people may use 'random' when they really mean 'haphazard', statistics examiners will not. The other two technical terms often confused by candidates, reliability (obtaining a broadly similar estimate for repeat applications) and validity (measuring what we really claim to measure), featured less obviously in this year's paper.

Question 1

The question asked for a description of the method of data collection, types and uses of data, for either the UK decennial census or some other large 'official population survey'. It did not want a general discussion of survey techniques, nor a description of (say) an industrial survey. What *was* required was a clear statement of facts about the national census.

Question 2

(i) A 'target population' is the group of units (human or otherwise) for which we wish to obtain or infer information. It is not the same as a 'sampling frame'; this is, in essence, an available list of members of that population, which may have inadequacies.

(ii) The obvious possible target populations were the 3000 office workers and those who actually use the canteen. To split into 'salad bar users' and 'hot meal users' is artificial – many regular customers may buy either depending on mood or the weather (though some credit was given for this). Some credit was also given for 'council workers' and 'general public' on the grounds that the latter might be considered for admission. It is, in fact, improbable that any council would consider admitting the general public to a subsidised canteen, but perhaps this was not obvious to all candidates. Whatever populations were suggested, intelligent suggestions about the kinds of questions to ask were expected.

Question 3

(i) Bias is a systematic tendency to overestimate or underestimate some parameter, whereas error can be purely random. Selection bias is a tendency to overestimate or underestimate parameters because the method of selection tends to produce unrepresentative samples. Many candidates gave definitions of selection bias and response rate but did not really say why they were a problem.

(ii) A too common mistake here was to suggest that for scheme B selections made by interviewers could cause bias. The method actually specifies a 'random sample' (perhaps a simple random sample, though at least a random sample) – so this suggestion was simply mistaken. Another feature was those who said that a 'low response rate' was likely for all the methods (though did not tell us what we might do to get a high one). Another fault was the use of comparative language ('this will get a higher response rate') without saying what it was relative to. Some suggested that scheme C (sending emails) would be time consuming. However, given that it says each staff member has a works email address it seems obvious that these are somewhere centrally held – and mass emails with access to this would be very easy and quick. Note was taken in marking of those who made more creative comments; for example, that a follow-up could be made for scheme E, that a telephone survey might supplement some other type, or that two might be used together.

Question 4

(i) Almost everyone could do this – and felt comforted at a nice calculation which was straightforward.

(ii) Not so many did this part well. Some gave nonsense about stratification requiring less of a sampling frame (totally untrue), and others said it was 'less biased'. Neither simple random nor stratified random sampling are biased, and in most cases stratification improves precision. Others said that comparisons between strata were possible only with a stratified sample. This is not the case, although the comparisons are likely to be better if stratification is used, since the stratum sample sizes will then be more suitable. Some implausibly suggested that stratification might be expensive or time consuming, whereas if the strata are on a staff database this is unlikely to be so. Yet others suggested a Neyman allocation of sampling fractions. This was particularly unrealistic, since we are not estimating some single parameter on which to base the calculations, and it is very unlikely that we would have estimates of standard deviations anyway.

Question 5

(i) This was fairly open ended, but what was being looked for was some indication of a process of thinking about the scenario in real terms. A combination of methods might achieve what was sought, for example, in terms of 'feeling consulted' a total email or a foyer display could be combined with some other method.

(ii) This was looking for a basic description of multi-stage, followed by the conclusion that it was almost certainly pointless in this instance. Quite a lot of candidates did actually get both these points.

(iii) This again was open ended – to see if candidates could make a reasonable suggestion and defend it. By a ‘sampling method’ it was implied that the method of data collection (that is, postal or interview) would also be discussed or recommended.

Question 6

(i) This was fairly straightforward. The key point is that in a longitudinal survey it is the *same units* being surveyed over a time period.

(ii) Again this was fairly straightforward. A ‘panel’ need not imply actual meetings together.

Question 7

(i) This brought some amazing answers. Many candidates ignored the ‘specialised diet’ issue altogether, or framed very general questions such as ‘Why don’t you use the canteen?’ What was expected was clear distinction of open and closed questions (and a few candidates gave what were actually closed questions for both!), *and also* some sensible and relevant actual questions.

(ii) This should have been straightforward bookwork, launched from part (i).

(iii) Many of the suggested questions which supposedly could lead to bias were either just silly questions or perhaps ambiguous – which is a different issue. This part again was intended to be applied *to this survey*. A couple of candidates suggested there might be non-response bias if we asked a question about sexual experience. While this could well be true, such a question could not reasonably have been asked in a canteen survey, and something more relevant to the practical situation would have been more plausible.

Question 8

(i) Answers to this seemed rather more realistic than last year. There were still some like ‘when the lights are off the information is rubbed off’ – but in general it seems to be recognised that it is no use citing as ‘disadvantages’ things one can do something about. Data should always be backed up (SPSS or Excel data files are not all that large), and virus protection should be in place. Given that, they are at little more risk than paper copies should the building burn down. Perhaps some overseas candidates might also note that in general in the UK councils have computer systems already in place (hence every worker having a works email) and do not suffer constant power cuts. There may, of course, be some expense, for example, in buying SPSS. There were also some unrealistic claims about large amounts of training needed for staff to enter data – sometimes in the same scripts which suggested that the same staff would be able to do statistical analysis by hand and spot any trends or patterns as they did so.

(ii) This was usually done satisfactorily – though some inevitably ignored the clear comment that optical character recognition (OCR) was not available. It really is not all that commonly available at present, and buying it is an unwarranted expense, but again it tends to be a ‘reflex’ word introduced by candidates who may never have seen an OCR device in their lives.

(iii) This was package-dependent, and a fairly wide allowance was made for this.

Ordinary Certificate Paper II

The overall standard was similar to that in previous years. There was a considerable increase in the overall number of candidates. A few candidates did not attempt sufficient questions to have a realistic chance of passing. There was little evidence that shortage of time was a contributing factor to this. It is gratifying that candidates have made a greater attempt to provide comments and interpretations when requested rather than concentrating solely on the more numerical parts of questions. Candidates seemed better prepared for the probability questions (apart from conditional probability) than in previous years. Candidates should be encouraged to leave probabilities in fractional form rather than converting to recurring decimals.

Question 1

This was a popular question that almost everyone attempted. In general the charts were better drawn than the tables. Candidates are reminded that a table should have a title and the source of the data should be indicated. It is helpful for the headings and categories to be separated by ruled lines. There was no need to show percentages in each category.

In (ii), it was important to realise that 43 letters were on ‘other’ or ‘unclassified’ subjects and a category for these letters had to be included on both the table and the chart. When drawing the chart, although the ‘other’ category had the largest frequency, it does not make logical sense for it to be placed first in the diagram as the reader needs to know what are the categories covered before he/she can comprehend the meaning of ‘other’.

Most candidates used graph paper for the charts, as was expected, and this made it easier to draw the bars of the appropriate length. Several candidates had great difficulty in plotting the large frequencies in (i) accurately although the same candidates coped with the smaller frequencies in (ii). In general, most candidates labelled the axes and categories appropriately and gave an appropriate title but it was only a minority who indicated the source of the data. A simple bar chart was expected and thus no shading to differentiate the categories was necessary or desirable. A handful of candidates drew a component bar chart that was marked as acceptable if drawn correctly.

Question 2

This was generally well answered. In part (ii) there were some nice comments about the company paying people to take the sofa away! Part (iii) caused the most problems, because although most candidates realised that you could not just add the percentage discounts, many obtained the incorrect answer of 77% by applying the further 10% discount to the wrong amount.

Question 3

Many candidates seemed unsure what was meant by a contingency table and many put the contestants’ letters in the boxes rather than the numbers of contestants. Both in this question and in question 8, candidates seemed not to understand the idea of conditional probability and just quoted the probability of both events. A large number of candidates found difficulty in finding the mean and standard deviation of a variable from a frequency table.

Question 4

The majority of candidates were not adequately prepared for this routine calculation. Several of those who attempted this question scored no marks; these were candidates who attempted to find the median of the frequencies or even of the number of classes. Those who correctly calculated which salary they were looking for out of the total of 72000 often failed to identify the relevant salary group, and simply put a £ sign in front of the frequency (for example, median = £36000). Some of those who did identify the salary group in which the median fell did not use linear interpolation to estimate where within this group the median was.

Part (ii) of the question was adequately done and in (iii) most knew that the few high salaries would distort the values of the mean and the standard deviation but hardly anyone mentioned the difficulty of estimating the mean and the standard deviation because of the open-ended classes.

Question 5

Most candidates were able to say that the correlation coefficient is a measure of the relationship between two variables but omitted the vital word *linear* . Many candidates did not read part (ii) of the question carefully, and therefore did not realise that they were being asked to plot graphs for r actually equal to $+1$, -1 and 0, not just approximately equal to these values. In part (iv), many candidates attempted to plot two graphs when what was required was a single graph. The calculations in part (v) were generally carried out well. There were a few errors in remembering the formula, some forgot to rank the data before applying the formula and some forgot to comment on their result. There were some answers greater than one with no comment that an error must have been made.

Question 6

The answers to this question were somewhat disappointing, giving the impression that most candidates were not secure in the meaning and interpretation of index numbers. Several candidates did not attempt the question and others had obviously left the question to be answered late in the examination when they were rushed for time.

It is important that candidates should realise that an index number is a percentage - several seemed to treat it as a sum of money. As a percentage, it is vital that both the time for which the index is calculated and the base period are clearly stated when the value of the index is calculated. Of course, in the interpretation stage, it is acceptable to discuss the values of an aggregate index number in terms of what it would cost currently to buy the same basket of fruit that cost £1 in the base period. Most candidates stated the advantage of using a weighted index in (i) but did not appreciate in (ii) that a Paasche index was to be preferred. Several stated that, because an index was being calculated relative to a base year, base-year weighting was appropriate!

The arithmetic of the calculation was generally satisfactorily done but some stated that it was impossible to calculate an index as some prices were given per kg and some per unit. A few others divided before summing so ending up with a quotient of two sums of unweighted price relatives. At the interpretation stage, most were able to give the percentage rise in prices as measured by the index. It was also important to say in this case that a Paasche index probably overestimated the price rises as the items that have gone up most have been consumed more by the family whereas for bananas, although the price has gone down, the family has consumed fewer of them. No one observed that although this was an index number of fruit prices, only four types of fruit had been taken as representative of all fruit.

Question 7

This was a popular question although a sizeable minority attempted only the first two parts. The explanation of terms was not well done on the whole. Many explained the term ‘seasonal’ by using the word ‘seasonal’ and assumed that seasonal components referred solely to seasons of the year. Even those who referred to variations about the trend did not always prefix the words ‘short-term’ and ‘regular’ although a few used the word ‘cyclical’ which implied they were not clear about the difference between seasonal and cyclical variation. In (c) and (d), many candidates were keen to say when each type of model should be used without actually stating what each model is.

The arithmetic in part (ii) and the interpretation were competently done in the majority of answers, although a few calculated a centred average by taking the four-point moving average of the four-point moving averages. In most answers the table of calculations was suitably laid out with headings. Amongst those who did attempt the last part, several went ahead with an additive model although a multiplicative model was requested. Others forgot to remove the trend before averaging but it was gratifying that a sizeable number could perform the desired calculation accurately. The interpretations were disappointing. In almost every instance the interpretation was in qualitative terms, and could have been made just by looking at the original data in the table. It was expected that the values of the seasonal indices be used in the interpretation, stating on average what percentage of the trend in passenger numbers could be expected in each quarter.

Question 8

Although many candidates were able to state that the probabilities would have been based on the results of repeated tossing, few commented on the large number of tosses required. 100 tosses was the number most frequently quoted whereas a figure considerably in excess of this is needed. The probability calculations were better than in earlier years, though again some candidates did not answer what is asked, namely to give the probability for each possible outcome in part (ii) before combining them in part (iii). There were problems with conditional probability again in part (iv), some candidates correctly worked out the probability of one head being on coin A when two heads were obtained but then did not divide by the probability that two heads were obtained in total.

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.

Eight of the 46 candidates had marks discounted for sixth or seventh questions whilst having five questions which scored at least as well.

The general standard was good. There was an overall pass rate of 31 out of 46 or 67%. The overall average mark was 60.3%. Twelve candidates obtained distinction standard (75% or more), four of these scoring over 90%.

As is usually the case, the more popular questions were associated with higher scores: the best were Question 4 (‘unseen’ continuous probability distribution) and Question 3 (Normal distribution), followed by Question 6 (Binomial + Bayes) and Question 5 (Poisson with asymptotic inference). Question 1 (permutations and combinations) and Question 2 (probabilities of sets) were popular but often poorly done, whilst Question 8 (correlation and regression, largely descriptive and interpretative) was slightly less popular; average scores of all three were around 9/20 to 11/20. Easily the least liked and worst question was no. 7 (Geometric distribution and probability generating function) with 11 attempts of which 10 were poor. However, it should be noted that even popular questions sometimes address poorly known topics whilst some less popular questions yield a few ’easy’ marks. More details will be given in the analysis by question.

Predominant strengths and weaknesses listed above are similar to those noted in 2002. Many candidates are not confident with combinatorial analysis or probability calculations involving the intersections and unions of 3 events; there were also very few good attempts at the Bayesian argument required in Q6. In Q5, practical application of asymptotic inference in the Poisson context was generally weak, and sloppy presentation of the maximum likelihood argument was widespread. Candidates’ knowledge of generating functions (Q7) appears seriously deficient.

Question 1
Only the very simplest calculations [(i) a, b] were consistently well done; even $\binom{10}{6}$ for part (c) was problematic for many candidates, and very few realised that in part (ii) the palindromes formed from the digits 0, 1 and 2 could quickly be enumerated if no general method suggested itself. Numerous grossly incorrect arguments characterised many solutions to later parts of this question. Very few candidates realised that in part (iii) the palindrome is completely defined by its first three digits for which the number of possible sequences is simply $10 \times 10 \times 10$ or 1000.

Question 2
Part (i), in which all three events are independent, was satisfactory, but part (ii) based on pairwise independence was relatively weak. Very few candidates attempted to draw a Venn diagram in which the probabilities of the eight possible triple intersections $(A \cap B \cap C)$, $(A \cap B \cap \overline{C})$, ... , $(\overline{A} \cap \overline{B} \cap \overline{C})$ can easily be expressed in terms of x, although a few equivalent symbolic arguments were seen. Due probably to the absence of diagrams, no candidate correctly deduced the minimum and maximum possible values of x. In fact $P(\overline{A} \cap \overline{B} \cap C) = x - \frac{1}{24}$ and $P(\overline{A} \cap B \cap C) = \frac{1}{8} - x$ place the tightest constraints on x, from which it follows that $\frac{1}{24} \leq x \leq \frac{1}{8}$.

Question 3
This relatively straightforward exercise on the Normal distribution was popular and generally well done, with most candidates evidently well-drilled in the use of Normal tables. Most errors were therefore strategic, for example overlooking the contribution of the initial process in part (i). Similarly, some candidates failed to distinguish the situations in parts (ii) and (iii): in (ii) the comparison is between Y_A and Y_B whereas in (iii) it is between $(X_1 + Y_A)$ and $(X_2 + Y_B)$, where X_1 and X_2 are independent random variables distributed as is X. Not all candidates dealt correctly with the comparison of two independent mean (total) completion times in the final part.

Question 4
This analysis of the ‘unseen’ continuous [Beta (3, 2)] distribution was, in the main, competently done. Common minor errors in part (ii) included poor sketching of the graph (which is asymptotic to the x-axis at the origin) and failure to confirm that the turning point at $x = {}^2\!/_3$ is a maximum. In the final part, several candidates were content to define $F(x)$ only in the interval $[0, 1]$ and used the variance instead of square-rooting it to obtain the standard deviation of 0.2.

Question 5

Marks on this fairly popular question averaged about 12/20. The Poisson probabilities for λ were correctly found in most attempts, but marks were lost by candidates who felt constrained to draw a continuous curve through the points. Several candidates also wasted time by deriving the mean and variance rather than quoting these as asked. In part (ii), the standard of mathematics used in finding $\hat{\lambda}_{ML}$ was often slipshod, with Σ and Π notation frequently misused and the suffix of summation omitted; as in Q4, candidates often neglected to confirm the maximum by showing that the second derivative of the log-likelihood was negative. Too often the impression given was of bookwork imperfectly learned by rote but not understood. This idea was reinforced by the failure of all but a few candidates convincingly to deduce a large-sample asymptotic confidence interval for λ. In part (iii), several candidates correctly found the sample variance to be 25 and (relaxing the Poisson assumption) went on to find the correct (wider) confidence interval. Disappointingly, however, no candidate drew the inference that, because the variance of this large sample was several (four) times greater than the mean, the appropriateness of the Poisson model for these data was greatly in doubt.

Question 6

A very popular question with a satisfactory average score of 12.6/20. The standard results for the mean and variance of the Binomial distribution were well known, and as in Q5 a few proofs were given although not asked for. Most candidates correctly identified the correct Binomial and shifted Binomial distributions required in part (ii), although in subsection (c) several misinterpreted 'more than two questions' as 'at least two questions'. Part (iii) was less well done: several candidates failed to note that, for example, student A would necessarily get at least 27 questions right, so that a score of 29 would correspond to an outcome of 2 for a $B(48 - 27 = 21, 0.25)$ random variable, and by the same logic a score of 29 for C was impossible. Many candidates omitted to try the Bayesian calculation at the end of this part; perhaps suggesting that if greater emphasis had been placed on conditional probability in this question, the results would have been much less successful.

Question 7

Work on this question was extremely disappointing. There were astonishingly few attempts, all but one of which were poor. In part (i), vanishingly few candidates could articulate an argument based on the statistical independence of trials to derive the Geometric probability $q^x p$; graphs of this function against integer x were rather more successful, apart from the occasional intrusion of continuous curves. Derivation of the probability generating function in part (ii) was usually omitted and when attempted nearly always disastrous, witnessing again to ill-remembered rote-learned bookwork along with much confusion with the moment generating function despite being given the correct final formula. Several candidates showed further confusion by putting $s = 0$ rather than $s = 1$ after differentiating $G_X(s)$. In the final part, observing that $Y = X + 1$ and deducing that the p.g.f. of Y was $sG_X(s)$ was beyond nearly all candidates.

Question 8

Although largely descriptive and interpretative, this question was not popular; it produced an average score of 11.5/20 on 20 attempts. Many candidates correctly remembered the formula for r, but were at a loss to explain it. Scatter diagrams for strongly correlated and independent data were generally well done, but several candidates could not portray 'uncorrelated but not independent' data, e.g. by scatter about a roughly balanced-up-and-down (non-monotonic) trend. In the analysis of Minitab output, part (a) was generally good, but in (b) candidates often failed to make the simple calculation of $\sqrt{R^2}$ or to explain the effect of removing the possible outlier. In part (c), very few candidates appealed to the partial p-value of the constant term (with or without the outlier) to justify non-significance. In the final part, statistical critiques of the two analyses with and without the outlier usually failed to mention the significant residual noted in the (age, chol) output.

Higher Certificate Paper II – Statistical Methods

The main aim of the Statistical Methods paper is to examine the understanding of fundamental concepts of statistical analysis. This is achieved by asking candidates to solve standard problems of estimation and hypothesis testing with particular emphasis being placed upon assessing each candidate's ability to summarise and interpret the results obtained from statistical analyses.

There were only two questions in which the graphical presentation of data was necessary, and the general presentation of this was better than in previous years with most candidates using graph paper, remembering to label axes and including titles. Hopefully this improvement will continue in future years.

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. However, many candidates are poor at explaining the meaning and uses of statistical tests in general terms and some have difficulty in establishing which statistical procedure to perform if this is not stated in the question. Frequently candidates are unable to state and explain the assumptions required for tests to be valid 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 very vague or omitted entirely by some candidates.

As we have noted, candidates are generally good at calculating descriptive statistics such as mean, median and standard deviation. However, it is wise to keep in mind the general comment at the start of this document that

> 'when a calculator is used the **method of calculation should be stated in full**'.

It is only too easy to lose marks unnecessarily by just stating the numerical values of means, standard deviation etc. obtained from the statistical functions of their calculators without showing knowledge of the methodology by explaining how the value is obtained.

Several questions required the candidates to obtain confidence intervals, for example, for means or for the difference between means. When obtaining these, a large number of candidates failed to state the general expression or formula being used before inserting the calculated values for the means and standard deviations. Often this was done without making any attempt to explain what any of the values were or how they had been derived; presumably they were obtained directly from the statistical functions of their calculators.

When the general expression for a 95% confidence interval was quoted candidates frequently included just the numerical value obtained from statistical tables, say 1.96, but did not indicate that a particular percentage point of the Normal or t distribution was required. It is not clear whether candidates understand where this number comes from. Have candidates learnt the formula in this form? If so, would they be able to handle different confidence intervals such as 99% or 90% when this value would change?

Several of the questions required the candidates to perform a hypothesis test. On the whole candidates have a good grasp of the basic requirements of hypothesis tests and can calculate the appropriate test statistics. However, many have difficulty in presenting the analysis clearly, 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 interpret what this means in relation to the problem posed in the question.
- Confusion between one- and two-sided tests. Some candidates state a two-sided alternative hypothesis and then proceed to perform a one-sided test and vice-versa.
- A common error in all two sided 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 sided test at the 0.05 significance level, a critical value from tables at the 0.025 significance level is 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. Alternatively some state whether the null hypothesis should be accepted or rejected without giving an explanation for their conclusion.
- The number of degrees of freedom is not always given and in some cases is incorrect.

Question 1
(i) Most candidates were able to perform a χ^2 test as instructed in the question, but demonstrated less competence when trying to interpret their findings, draw appropriate conclusions and make recommendations to the sports manufacturer. Although ‘correct’, many of the solutions offered were rather brief and failed to include all the necessary components to obtain full marks. For example, null and alternative hypotheses were omitted or the critical value from tables was not quoted or quoted without reference to the degrees of freedom or the level of significance used. Candidates should be reminded that they must include all the necessary working in their answers if they are to obtain full marks. Some candidates failed to include any recommendations to the sports manufacturer concerning the choice of commercial in their answers and many of those who did were incorrect or muddled.

(ii) Generally candidates were able to apply McNemar’s test to the data but a number had some difficulty in commenting upon the results obtained.

(iii) Most candidates found this part of the question especially difficult. Very few seemed to understand that McNemar’s test should be used on paired or matched data. Marks were also lost by candidates failing to give examples of situations in which each test would be preferred to the other.

Question 2

(i) It was pleasing to observe that, unlike previous years, all histograms were drawn on graph paper and the general level of presentation was better than in recent years. Only a few candidates did not include titles or failed to label axes. However, many candidates continue, incorrectly, to label the y-axes as ‘frequency’ or ‘Number of calls’ and use a scale that suggests that the frequency is represented by the height of each ‘bar’. This occurred on scripts from candidates who correctly understood that in a histogram it is the *area* of each ‘bar’, not the height, which represents the frequency of the group. Unfortunately, many candidates misunderstand this and continue to represent the frequency of each group by the height of the ‘bar’. There were some candidates who appreciated that the width of a bar should depend on the range of values covered, but unfortunately did not adjust the height of the bar so that the class frequency was represented by the area of the bar. Very few candidates indicated how the frequencies are represented on the histogram (for example, 1 cm^2 could represent 5 calls). Remembering to include this may help candidates to understand more fully that in a histogram frequency is associated with area and not height.

(ii) Most candidates have no difficulty in obtaining an estimate of the mean from grouped data; however many continue to lose marks by omitting the necessary working in their answers. Obtaining an estimate of the median proved to be more difficult. In general candidates were able to identify that the median would be in the interval ‘≥ 25 but < 30’ but many failed to explain fully how the median could be estimated using interpolation. Many gave a numerical expression without explaining what each of the values represented, which is insufficient. Descriptions of the distribution were generally correct but often rather vague and many candidates did not make reference to the shape of the distribution observed in the histogram, limiting their comments to the similarity in the values obtained for the mean and the median.

(iii) Not all the working was given, as required by the rubric. Some candidates lost marks by using the formula for a population variance rather than the sample variance. Many candidates did not quote the formula for the 95% confidence interval before substituting the relevant values into the expression. In some cases, although the values were correct, there was very little explanation as to what these values represented or how they were derived.

Question 3

Quite a number of the candidates who attempted this question performed lengthy yet inappropriate analyses that did not address the issues asked in the question. For example

- the difference in means between apparatus A and B was examined
- comparisons of the variability of apparatus A and B were made.

Neither of these addresses what was posed in the question.

Other candidates obtained the mean and standard deviation for each apparatus and without performing any statistical tests commented upon the magnitude of each compared to the laboratory standard. This is completely inadequate.

(i) A number of candidates stated the null hypothesis that the variance of each apparatus was equal to that of the laboratory standard and then performed t tests to compare the means of each apparatus with the laboratory standard.

(ii) This was generally better answered than part (i), although some candidates did not understand that bias related to the location of the means and compared the variances between apparatus A and B to examine bias.

For those candidates who did understand what was required, the solutions to parts (i) and (ii) were good although sometimes not all the necessary working was shown and hence marks were lost.

Candidates found particular difficulty in making correct comments on how the apparatus could be altered to improve the accuracy of the measurements. A surprising number of candidates erroneously stated that 'increasing the sample size would improve the accuracy of the readings'.

Question 4

(i) Most candidates were able to correctly identify that the appropriate analysis to perform was a sign test, but only a small number of candidates were entirely successful in its execution. Most candidates were able to calculate the test statistic and to make use of the binomial distribution with $n = 10$ and $p = {}^1\!/_2$, but many performed a two-sided test rather than a one-sided test. As the psychologist's claim was that visual memory was *more* effective than aural memory, a one-sided test was required. Some candidates stated a one-sided alternative hypothesis but performed a two-sided test, or vice versa. Other candidates made no reference to whether they were performing a one- or two-sided test.

A few candidates tried to perform a Mann-Whitney U test using the candidate number to rank the data for those scoring higher on A or V.

(a) Candidates had some difficulty in correctly explaining why a sign test would be inappropriate to analyse the data. Several incorrectly stated that 'the sign test could not be used as the data are matched'. A few candidates were unable to identify the correct test to perform, which is a Wilcoxon signed rank test.

Incorrect analyses included

- calculating the Spearman rank correlation coefficient between the before and after scores
- performing a Mann-Whitney U test.

Some candidates are confused between the procedures for a Mann-Whitney U test and a Wilcoxon signed rank test. A few correctly stated that a Wilcoxon signed rank test should be performed and then proceeded to execute a Mann-Whitney U test.

Candidates correctly attempting a Wilcoxon signed rank test were generally able to obtain the correct test statistic. A number of candidates, however, did not ignore the sign of the difference when calculating the ranks, with all negative differences incorrectly being given the lowest ranks. Statements of the null and alternative hypotheses, when given, were often imprecise or incorrect including statements relating to the equality or otherwise of the mean value in each group. Many candidates lost marks here as in part (i) by performing a two-sided rather than a one-sided test as was required.

(b) Most candidates could correctly identify that the appropriate parametric test would be a paired samples t test. Despite being told in the question that it was not required, some candidates went on to perform this analysis on the data. Only a few candidates stated that a necessary assumption for the analysis to be valid is that the distribution of the differences should be Normal. Even fewer noticed that this assumption did not seem to be valid because of two large differences or outliers.

Question 5

(i) In the main, the explanations of the Central Limit Theorem were poor with many candidates having difficulty in clearly expressing the key points. Although the question asked for an informal explanation, a number of candidates did quote a 'textbook' definition, and were unable to demonstrate that they understood its meaning or practical importance. Several candidates lost marks by not attempting to explain its practical importance at all.

(ii) Most candidates were able to correctly construct a 95% confidence interval for the difference between the means. Greater difficulty was encountered in interpreting the confidence interval, with only one or two candidates being successful. A common answer was merely to state that 'as the interval does not contain zero the drug is significantly better than placebo'. While this statement can indeed be justified, an interpretation of the interval must be based on what the interval tells us about the value of the parameter concerned. It is certainly better not to treat the concepts of confidence interval and hypothesis test as essentially identical.

(iii) This was generally well answered by candidates who knew the correct formula for the confidence interval.

Question 6

(i) The statement of the model was well done but many candidates failed to state the necessary assumptions for the analysis to be valid – that the terms in the model are additive and that the observations are sampled from Normal distributions with equal variances.

(ii) (a) Candidates can correctly construct a one-way ANOVA table and test for a difference between the groups. Many candidates conclude their analysis after completing the F test. 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 and used to examine differences between the mean yields. Many candidates lost marks unnecessarily by not including the null and alternative hypotheses, the numbers of degrees of freedom or the tabulated value obtained from F tables used to test the null hypothesis. In addition, candidates failing to include a report for the farmer also lost marks.

(ii)(b) Candidates who attempted this part had a good general grasp of how the experiment could be re-designed. Marks were lost by not including all of the key points, especially that each fertiliser should be allocated at random within each block.

Question 7

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. Solutions could have contained line graphs depicting the trends over time for total expenditure and several commodities of particular importance or interest, or pie charts showing how the spending was allocated between the various commodities and services for particular years. Candidates who did attempt this question did not generally include any diagrams to support their discussion and hence did not obtain as many marks as might have been possible if they had done so.

Question 8
(i) Candidates were able to identify situations in which the F test is used, but lost marks by failing to illustrate their answers by the use of examples.

(ii) Several candidates failed to understand what was required and performed a range of incorrect analyses. Most candidates who were able to identify what was required gave good answers; however, some used incorrect numbers of degrees of freedom in the F test. Some candidates lost marks by failing to write a short report containing their recommendations to the manufacturer.

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 comprehensibly. 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.

Some candidates wrote that they *accepted* a null hypothesis. This is poor wording; it is more appropriate to conclude that they *do not reject* that hypothesis.

Question 1
This question on one-way ANOVA was very popular and was done by 46 of the 49 candidates. The average mark was 8.1.

In (i) some candidates did not state the assumptions although the question asked them to do so. Not all candidates who stated the assumptions did so correctly, for example candidates referred to the dial types as having Normal distributions; it is, of course, the variable being studied which has a Normal distribution. The assumptions in fact relate to *populations* from which random samples of measurements have been taken. Many candidates gave rather vague statements of the hypotheses, for example, stating as null hypothesis that there is no difference between the dial types, instead of stating the null in terms of equality of *means* of the populations from which samples of measurements have been taken.

Part (ii) asked for an estimate of the difference between two means and then for a confidence interval for the difference. Many candidates neglected to state the point estimate. As the experiment related to comparison of three dial types, candidates were expected to use the residual (or error) mean square which could be derived from the analysis of variance table given in the question, and some did this.

Many candidates found a confidence interval for the difference between two means using a pooled estimator of a common variance obtained from just the two samples and a t distribution. A few candidates used the degrees of freedom for the error mean square (18) when they used this second method and some used the degrees of freedom appropriate for this second method for the first. A few candidates based a confidence interval on the t distribution but found the variance as $(s_1^2/n_1) + (s_2^2/n_2)$. This sort of calculation was not required – indeed, this method is outside the scope of this paper. Not surprisingly, candidates who used it were generally not aware how to calculate the degrees of freedom appropriate for this. Had the samples been large, a Normal approximation could have been used without any necessity to assume a common variance. However since the assumption of a common variance was needed for part (i) it seems a little strange to relax this in part (ii). Explanation of the meaning of the confidence interval was in general poor.

Not all candidates attempted part (iii). Some of those who did stated the assumptions requested in (i) instead of or as well as saying how they could investigate them. A brief reference to assumptions to be investigated would have been sufficient.

Question 2

This was on two-way ANOVA with replication and was attempted by 37 of the candidates. The average mark was 8.7.

Part (i) asked for an explanation of interaction and on the whole was not well done with some very sloppy answers, including statements such as ‘Interaction occurs when variables interact with one another’.

Some candidates made a poor choice of scale for the plot in (ii). Some felt that the plot showed interaction because there was some crossing of the lines in the plot whereas it is more important to consider departure from parallelism (obviously non-parallel lines might not intersect in the plot). In this example the plot does not suggest severe interaction. Candidates were also asked to refer to the results of the analysis in order to comment on whether there appeared to be interaction and several of those who did this concluded incorrectly that the relatively large p-value meant that there was interaction, perhaps because this was what they had concluded from the plot. Not all candidates stated, or stated correctly, the hypotheses (the null here is that there is no interaction) and with a p-value of 0.154 we would not reject the null.

Although part (iii) asked for an explanation of how to interpret the p values in the output, some candidates made no reference to these in their answers. Some did not state the hypotheses which could be tested, and some did not refer to means of populations (see also the comments on Question 1). There were also some sloppy statements saying that a variable has an effect, without indicating the type of effect. Hardly any candidates stated the assumptions needed for their answer in (iii), although asked to do so.

Question 3

This question involved maximum likelihood estimation of the parameter of a Poisson distribution, a χ^2 goodness of fit test, and finding a confidence interval for a population mean. It was very popular, being attempted by 42 candidates. Most of these did it fairly well. The average mark was 11.6.

Some candidates did not set part (i) out well, equating the partial derivative (in terms of λ) to zero. The derivative is only equal to zero at a turning point and when equating it to zero $\hat{\lambda}$ should be used instead of λ. Many candidates had problems in dealing with the constant $x!$ both in writing down the likelihood and in finding the log of the likelihood. Few candidates checked that the turning point was indeed a maximum.

In (ii)(a), a few candidates appeared to have calculated the probabilities, although they could have read them from the printed tables. Some used the wrong degrees of freedom for the χ^2 test, some found an expected frequency for ‘exactly 5 claims’, whereas the expected value for ‘5 or more claims’ is needed; this ensures that the sum of the expected values is equal to the sum of the observed values. The expected frequency of the ‘5 or more’ group was rather small, at about 1.78, and it would be much better to combine the last two groups (4 and ‘5 or more’) to ensure a reasonable size for all expected values used in the χ^2 test.

Some candidates referred to the sample mean as $E(X)$. It is very important to distinguish clearly between *sample* and *population* quantities.

Question 4
This question was on time series and exponential smoothing. It was not very popular, being attempted by only 19 candidates. The average mark was 9.8.

In part (i) the explanations of how to calculate the forecasts were poor, though candidates were able to perform the calculations in (i) and (iv). Part (ii) was poorly done with hardly any candidates having any idea when it is appropriate to use a high value for the smoothing constant. There were not many attempts at part (iii) and some of the candidates who attempted this part appeared to be guessing. Knowing how to measure the accuracy of a model is important. The mean absolute deviation (MAD) is one method of doing so. In (v) the comments were poor and there was some confusion between negative and positive errors, with some candidates saying that the errors increased over time because the plot of errors resembled a scatter around a line of positive slope. In fact the change is from large negative to small negative to small positive to large positive errors.

Question 5
This question was on simple linear regression. It was very popular with 43 candidates attempting it and was reasonably well done. The average mark was 10.4.

In part (i), many candidates did a plot but made no comments. Some seemed rather confused as they used both ‘linear’ and ‘curved’ in describing the relation which is **not** linear.

In (iii), some did not understand what is meant by interpreting the coefficients, and some were confused as to the correct interpretation of the intercept. For the regression of HRS versus $\log \text{NBR}$, the intercept is the expected value of HRS when $\log \text{NBR} = 0$, that is, when $\text{NBR} = 1$. There was also some confusion regarding the units of time. The interpretation of the slope is that for each increase of 1 in $\log \text{NBR}$, i.e. an increase of 2.72 in NBR, the expected increase in the number of man-hours taken is 181,000 (not 181). Part (iv) was not attempted by many candidates, and some of those who did this part appeared to be guessing, or did not say which variable they would transform, or clearly did not understand what is meant by a transformation. A reasonable transformation might be to take the log of HRS.

Question 6
This question was on confidence intervals for a difference in population proportions and on estimation of the sample size needed for estimation of a population proportion. It was moderately popular, being attempted by 30 candidates. It was not well done, with an average mark of 8.0.

In (i), some candidates referred to significance although the question asked for a confidence interval. Few candidates had any idea why the method used in (i) was not appropriate, and there were a number of incorrect answers which appeared to be guesses. Basically, the reason why the method is not appropriate is that the proportions are not independent. However, in (ii) the proportions relate to two statements which are mutually exclusive, while in (iii), although the two statements relate to different situations, they are answered by members of the same sample. Both these situations are ones where we might want to compare proportions. In (iv), some candidates did not make clear that an estimate of the unknown population proportion is needed in order to estimate the sample size and some did some very strange and incorrect calculations.

Question 7
This question was on non-response and estimation of a sample size needed to obtain a required number of responses in a situation where it is thought not everyone will respond. This was not a popular question and was not well done. It was attempted by 22 candidates, and the average mark was 5.6.

In (i)(a), many candidates gave an account of non-response instead of explaining why it is a problem. In (i)(b), many made suggestions as to how to reduce non-response instead of confining their answers to follow-up procedures. Although it was clear that these candidates had a reasonable idea of what is meant by non-response and how to increase the response rate, good answers to wrong questions receive no credit. Answers to (ii)(a) tended to be too brief and few mentioned the effect of these strategies on getting information from those who might be non-respondents initially. Hardly anyone attempted (ii)(b).

Question 8
This question asked candidates to report on the main features shown in a table. It was attempted by only 8 candidates. The mean mark was 10.8.

A few candidates were confused as to what the figures in the table were. All figures were percentages except for the last line which gave the base in millions, but some candidates interpreted the percentages as numbers, and some the numbers as percentages, and other referred to some or both of these as rates. Some time series plots were described as line charts. Some candidates did not do any diagrams although the question specifically asks for diagrams. In answering questions of this nature, candidates should try to focus on the main points rather than on fine details.

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.

This year, all candidates found (at least) five questions they could attempt. One candidate attempted 6 questions rather than the required 5. There was a more even spread of attempts at the various questions than in some years, though almost every candidate attempted the questions on joint probability density functions, moment generating functions and simulation. As in other recent years, with the exception of 2002, the least popular question was the one on Markov Chains.

The overall standard of attempts was good. There were very good attempts at all the questions, several of them virtually flawless. In general, candidates seemed more comfortable reproducing standard proofs (for example, margins of the multinomial, Poisson moment generating function) than solving problems. Candidates should be encouraged to broaden their experience of attempting problems, since this paper is unlikely to be passed on standard proofs alone.

Question 1
This examined knowledge of the multinomial distribution, requiring candidates to reproduce standard proofs about the marginal and conditional distributions and then to use these results in an applied problem. About half the candidates attempted this question and they generally obtained very good marks for it.

Question 2
This tested candidates' knowledge of the Law of Total Probability and Bayes' Theorem. Most candidates attempted this question, but not very successfully in general. Many candidates had considerable difficulty with part (ii) of the question. They did not understand, first of all, that the woman described there had prior probability one-half of being a carrier of haemophilia (given the information about her parents). Secondly, they did not realise that the information about this woman's sons was vitally important in determining her condition and should, therefore, have been taken into account in order to obtain a posterior probability that she was a carrier.

Question 3
Candidates had to derive various moments of a bivariate distribution and then obtain the marginal probability density function of $X + Y$. Almost every candidate attempted this question. The standard of their answers was mixed. Part (iii) seemed to cause particular difficulties, with many candidates unable to determine the correct region for the required integration. In part (ii), some candidates caused themselves difficulties because they did not use the general formula for $E(X^r Y^s)$ to find moments such as $E(X)$ but insisted on obtaining the marginal distribution of X first. Candidates should be encouraged to look out for phrases such as '*hence* find ...', which can greatly simplify their work in an examination.

Question 4
Candidates were tested on their knowledge of transformations of two continuous random variables. About two-thirds of candidates attempted this question, but the general standard of their attempts was disappointing. As in 2002, some candidates seemed to be confused about the definition of the Jacobian of a transformation. In part (ii), almost no-one recognised that R and ϕ were independent, which would have saved work in integrating out ϕ, largely because they did not write down the joint range space of R and ϕ in part (i) (which should be a matter of routine in a question like this).

Question 5
This question examined moment generating functions and the Central Limit Theorem, in the context of the Poisson distribution. Virtually every candidate attempted this question and it was well done. A surprising number of candidates specified the wrong limits for the required summation, for example, from 0 to n or from 1 to n, rather than from 0 to ∞.

Question 6
This tested candidates' general ability to solve problems in probability, in the context of hazard functions for individual probability distributions and systems. About one-third of candidates attempted this question, but there were just two reasonable attempts at it. Generally, candidates were able to work through part (i) with no trouble but they seemed to experience considerable difficulty in framing any answer at all to parts (ii) and (iii).

Question 7
The question was about simulation using the inverse c.d.f. method, although part (ii) was set in a problem-solving context. Almost all candidates attempted this question, giving good answers for part (i) but with mixed results for part (ii).

Question 8
This tested work on Markov Chains. About one-quarter of the candidates attempted this question, with mixed success. Those who struggled with it did so largely because they did not write down the 2×2 transition matrix correctly at the outset.

Graduate Diploma Paper: Statistical Theory And Methods II

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

The questions that dealt with classical topics, 1-5 and 8, were each answered by more than half the candidates. Questions 6 and 7 were less popular. Questions 1, 3, 5 and 6 were answered well by at least two candidates. Questions 2 and 8 were tackled least successfully by candidates and appeared to highlight two specific areas of weakness (see the question by question comments).

One candidate attempted six questions; all the others attempted exactly five questions.

There was a large range of marks. Candidates who answered questions systematically tended to do better than those who jumped repeatedly from question to question, doing a fragment at a time.

Question 1
There were several good attempts at (i) but a disappointing number of candidates did not handle logs competently. Some were unaware of the functional invariance of MLEs. A small number of candidates appeared to be unable to use summation and product notation. There were a few good attempts at parts (ii) and (iii) but some candidates made little progress here. Several were unaware that different parameters typically lead to different Cramér-Rao lower bounds and hence they found the CRLB for the wrong parameter. There were very few serious attempts at (iv), none of them correct.

Question 2
This question was done poorly and indicated a lack of understanding of sufficiency. Most candidates were unable to say clearly what a sufficient statistic is. There were no good attempts at part (ii), which required use of the factorisation theorem in situations where the range of the density depends on the parameter. The standard trick here is to use indicator functions in the likelihood. Many candidates had the right idea for (iii) but common errors were: failure to differentiate the distribution function of Y correctly and omission of the range of the density. In (iv) some candidates could not write down the mean squared error of an estimator.

Question 3
There were some reasonable attempts at part (i). The likelihood ratio is a monotone function of the given statistic; some did not spot this and undertook much needless manipulation. Disappointingly, some candidates were unable to write down the Bernoulli likelihood correctly. Relatively few candidates could write down the large sample distribution of a sample proportion for part (ii), which meant that little progress could be made in the later parts. Some candidates did not know what power means.

Question 4
Part (i) and the first part of (ii) were generally done well. However, most candidates seemed to miss the second part of (ii) and made no attempt to find the sample information. This had a knock-on effect in part (iii) where some rather simple numerical calculations could not be completed because they depended on the sample information having been found.

Question 5
Part (i) and the first part of part (ii) were generally well done. There were some good attempts at finding the expected sample sizes but some candidates did not know the relevant formulae. In part (iii) there were some good attempts but the numerical results were often poorly explained.

Question 6
Only about a third of candidates attempted this question. Part (i) was done reasonably well but some candidates were rather imprecise. The majority of attempts at (iii) were successful. In parts (ii) and (iv) over half the attempts were good but the remainder got nowhere, mainly because they could not manipulate the given formulae/results accurately.

Question 7
Less than a quarter of candidates attempted this question. All attempts at part (i) were good. Those who knew what bias and risk mean also did well in (ii). There were no convincing attempts at (iii).

Question 8
This question was done poorly. Very few candidates seemed to know how to tackle this type of question. It is always a good idea to have some structure to the discussion. Also, ‘compare and contrast’ means that something more is needed than simply producing separate lists of parametric and non-parametric tests. Comparative statistical inference is an important part of the syllabus and deserves serious and careful consideration as a topic in its own right. The skills required are different from those required for much of the rest of the syllabus. A more reflective and wide-ranging approach is needed in order to appreciate the ‘big picture’.

Graduate Diploma Paper: Applied Statistics I

The main objective of the paper is to test an understanding of the theory underlying general linear models (and other related methods) and its application to practical problems. Knowledge of theory is, in itself, inadequate. Similarly, candidates who attempt to answer the practical parts without an understanding of the underlying theory will not be awarded many marks.

Most candidates showed weaknesses in both areas. Theory, which is largely bookwork, often included imprecise statements and errors, and in answering the more applied parts candidates often failed to relate the theory to the specific application described in the question.

It is very important to spend time reading the question, deciding which theory is relevant, and thinking about the application area or data provided. Follow the instructions; if the question requires you to sketch a graph then you are bound to lose marks if you do not do this! Also, information provided in the question is *usually* relevant, and should be used in the answer.

Questions usually have a ‘theme’, and the various parts lead you through. Even if you cannot answer one part you may need information from it to answer a later part. In this paper it is *not* sufficient merely to repeat things that you have learnt from books; you need to be able to *apply* the theory to the problems presented.

Question 1
(i) Very few candidates answered this well. There was evidence of a basic grasp of the underlying theory, but a general inability to relate this to the data provided. It is important to know what ACFs and PACFs look like for real data.

(ii) Although this question was relatively straightforward, answers were riddled with errors, and there was inadequate *justification* of the working. There is usually a time series question on this paper, and preparation should not be difficult. You need to learn the basic theory and also study simple worked examples from the texts.

Question 2
Tasks like this are an important part of the applied statistician’s role. There were very few convincing answers. The answers tended to be self-contradictory or confusing. Although (a) might require some thought, the principles underlying (b) are commonly discussed in text books on multiple regression. You need to be clear about how to interpret output from packages, and about the relative merits of different modelling approaches. This means describing the output in terms of the variables and application areas.

Question 3
Very few candidates tackled this. MANOVA is clearly on the syllabus. The material was largely bookwork and routine. Remember that **any** topic in the syllabus can be examined, even if it has not appeared in recent years.

Question 4
This was a popular question, and reasonably well answered. Descriptions of the purpose of Principal Component Analysis and interpretation of the principal components were quite well done. However, candidates generally failed to describe the implications of the two problems with the data: the fact that the original data were ordinal, and the large number of missing values. Also, the components should be interpreted in terms of the original variables, in this case the actual questions on the questionnaire (not just Q1, Q2 etc).

Question 5
The bookwork was, on the whole, imprecisely presented. Few candidates saw the relevance of the data about numbers of pupils. The theme of the question was weighted least squares. Without grasping why this *might* be considered here, it was difficult to tackle the later parts of the question.

Graphs were of a disappointing standard. There is little value in a graph which has ‘school number’ on the x-axis.

Question 6
(i) Quite well done, although work could have been more precise.

(ii) Generally well done apart from a failure to state the hypotheses.

(iii) There were many arithmetic errors here in what should have been standard work.

Question 7
(i) Few candidates provided an example, as requested.

(ii) Not well done; this is routine bookwork.

(iii) This question asked candidates to use the result of (ii). Even if you were unable to do (ii), you should have used the stated results. Merely quoting results without proof gained **no** marks in this section.

(iv) Few candidates followed the instructions to produce summary statistics and to sketch a graph. Indeed, few were able to relate this part of the question to the earlier parts.

Question 8
There is usually a question asking for completion of an ANOVA table, which requires accurate arithmetic. Although part (i) was generally well done, work was riddled with errors. In some cases this made the later parts of the question more complicated.

(ii) Few candidates did a complete analysis.

(iii) Explanations were *far* too technical, and in many cases either inaccurate or self-contradictory.

Graduate Diploma Paper: Applied Statistics II

This paper aims to examine candidates' understanding of the fundamental concepts of designed experiments and sample surveys, and their ability to apply these to the analysis of data.

All candidates followed the rubric, although a few candidates who submitted four good answers did better than those who did five sketchy answers. As in previous years, some candidates continued to lose marks by not answering the question asked or omitting sections of the question entirely. Graphical presentation of data was generally untidy and poorly presented.

The general standard was disappointing. In all, 10 of the 22 candidates (45%) gained fewer than 40 marks on this paper. Only two candidates performed exceptionally well, gaining above 60 marks. The average mark was below 10 out of 20 for all questions except the demography question (average mark 14.4).

Candidates were much better at obtaining the summarised results of analyses i.e. constructing the analysis of variance, than at interpreting these results or making appropriate conclusions. Their knowledge of standard textbook methodology was poor e.g. LS estimation, properties of estimators etc. Questions which were more descriptive, or which required essays, were preferred (Questions 5 and 7). As in previous years, the response surface question was not popular.

Question 1 (17 attempts)
This question involved summarising the results for a Latin square design, but also understanding the reasons for its use, and how the randomisation is performed.

Most candidates were able to construct the analysis of variance for a Latin square design, but were less clear on the advantages and disadvantages of such designs. Only a few candidates mentioned the small degrees of freedom available for residual. Others thought Latin squares were computationally more difficult to analyse, and this would be a disadvantage.

In part (iii), candidates needed to state the underlying principles; selecting one of the four 4×4 squares, and randomising rows and columns and allocating letters to treatments at random. Many candidates did not appear to understand what the question was asking.

Attempts at part (iv) were rather poor and some candidates omitted this part. Most candidates seemed to be unable to write down a set of linear contrasts among the four treatments, even though these were the main effects and interaction of a 2^2 factorial treatment structure.

Question 2 (13 attempts)
This question required candidates to construct the analysis of variance for a 4×3 factorial design replicated 3 times, and to perform follow-up analyses, writing a report of their findings.

Parts (i) and (ii), which involved summarising the data using analysis of variance, were mostly well done. Attempts at parts (iii), (iv) and (v) were rather poor. Most candidates did not know how to partition sums of squares into linear, quadratic and cubic single degree of freedom components. Some candidates drew incorrect diagrams of the means, with the qualitative factor 'gas type' on the x-axis. Plots were often untidy, and not all included a scale.

Question 3 (13 attempts)
This question required knowledge of randomised block designs and the application of least squares when all comparisons are against a control treatment.

Part (i) was more theoretical and not done well. Most candidates correctly specified the model for a randomised block design, although a few included a term representing the interaction between treatments and blocks. Only a few candidates attempted parts (b) and (c). One or two suggested minimising the residual sum of squares to obtain a least squares estimator for the difference between the effect of a control treatment and a new treatment, but were not sure how to do this.

Part (ii), which involved summarising the data using analysis of variance, was well done. It was not relevant to perform an overall F test to assess treatment differences in this situation, because all comparisons were against the control treatment, although all candidates did. A few candidates constructed a 95% confidence interval using the critical values of the Normal distribution. There was some confusion over the interpretation of the 95% confidence intervals, with some candidates concluding no treatment differences against the control treatment, even though the confidence intervals did not contain zero.

Only a few candidates answered part (b). There were some good responses, mainly relating to the small percentages, and the use of transformations. All candidates overlooked the issues of multiple comparisons.

Question 4 (3 attempts)
This was not popular. One candidate did manage to complete all parts, and did reasonably well.

Part (i) required candidates to comment on the weakness of a design comprising 3 points of a 2^2 factorial layout for determining operating conditions that maximise yield. All candidates overlooked the possibility of an interaction between the two factors. All four combinations are needed in order to discover the shape of the response surface in the experimental region.

Part (ii) involved testing the lack of fit of a model fitted to 4 replicates of a 2^2 factorial design, and constructing the path of steepest ascent. Candidates assumed the lack of fit of the model could not be tested because the design did not include runs at the design centre. The 4 replicate runs at each design point will provide adequate degrees of freedom to test the fit of the response model. Had centre runs been included, a test for curvature would be possible also.

Question 5 (14 attempts)
Answers to part (i) were rather vague and rambled on. Candidates tended to assume that the term *random sampling* implied that each population member had the same chance of being included ion the sample (or EPSEM - equal probability of selection method), and some gave stratified random sampling as an example of EPSEM. With random sampling, units are selected by a probability mechanism, but not all sampling methods will give every item in the population an equal chance of selection. A few candidates gave quota sampling as an example of non-random sampling.

Part (ii) was not done well, even though this was a standard bookwork question. Most candidates could not write down the formula for an unbiased estimate of the variance of a large population based on a simple random sample. Only a few candidates attempted the sample size calculations for the pilot survey but provided incorrect answers. The sample size should be calculated for each objective separately, and the maximum sample size used. There were some rather odd answers.

In part (iii), some candidates confused *two-stage* sampling with *two-phase* sampling. Some candidates provided interesting examples of surveys in their countries that used both stratification and clustering in sample design.

Question 6 (13 attempts)
This question involved standard bookwork on the properties of ratio estimators, and its application to estimating the population total based on a simple random sample.

Part (i) was not well done, even though this was a standard bookwork question. Candidates did not know how to derive the bias of the ratio estimator although a few realised that this involved expected values. A few candidates managed to obtain an alternative expression for the estimate of the variance of the ratio estimator in terms of s_y^2 , s_x^2 and ρ.

Part (iii) was done quite well. Most candidates estimated the total yield using a ratio estimator but the reasons for their choice were often vague. Most candidates had difficulties calculating the standard error, even though the formula for the ratio estimator was given in (b) of part (i). There was confusion over units (since measurements on each unit were in gm but the total weight of the consignment in kg); and obtaining an estimate of the total yield when N, the total population size, is unknown.

Only a few candidates attempted part (c), on calculating sample size, but the answers were not correct.

Question 7 (14 attempts)
Part (a) required candidates to discuss considerations that affect the choice of wording in questionnaire design. There were some interesting examples especially for a question that could bias the response because of its strong wording. A few candidates confused ‘strong wording’ with technical jargon. Some candidates lost marks because they did not provide examples.

In part (b), some candidates had concerns over the use of the term ‘peers’ even though this was defined in the question. Most candidates agreed that respondents who failed to answer the question were likely to give rise to bias in the results, but could not explain why i.e. because non-responders will tend towards being users rather than non-users.

Question 8 (16 attempts)
This question on fertility and construction of elementary fetal death rates was popular, and well done. There was confusion over period and cohort analysis of fertility. Some candidates lost marks on the interpretation of the death rates, and thought the question required them to explain the differences in the construction of crude and age-adjusted death rates (rather than differences in interpretation).

Graduate Diploma Option: Statistics for Economics

The standard of work was, as in previous years, very poor, with few candidates displaying an ability in statistics at the Graduate Diploma level.

The two questions relating to comparatively elementary topics were especially poorly tackled.

Question A1
(a) There was widespread confusion between trend and seasonally-corrected series (the difference is the random element).

Practical examples of the use of seasonally adjusted rather than original data were largely lacking.

(b) Column C7 represents a centred four-quarter moving average of the logarithms of the original data, so C8 is a combination of seasonal and random effects. Few candidates showed that

$$\frac{2.21920}{8}+\frac{2.28238}{4}+\frac{2.49321}{4}+\frac{2.33214}{4}+\frac{2.35138}{8}=2.348255$$

and that $2.49321 - 2.348255 = 0.144955 \approx 0.1450$, despite being asked to do so in the question.

The seasonally-corrected data for the first two quarters of 2001 were seldom found.

The data show a strong upward trend, making multiplicative corrections more useful than additive corrections, though this could usefully have been considered having regard to the details of the calculation.

Question A2
(i) Interest rates have a major influence on the costs of holding raw material stocks and finished goods, so it is reasonable to expect that interest rates may affect inventory levels. The growing use of Operational Research methodology over the period should have resulted in greater efficiency in stock control, and a time trend can act as proxy for the use of OR methods. Both these variables can be expected to have negative coefficients, though other influences might confuse the issue.

The partial correlation coefficient is 0.3026. The test statistic is 1.347, distributed as t_{18}, and is not significant. Candidates sometimes tested R^2, or used an incorrect number of degrees of freedom. For part (c), the t statistic for the regression coefficient for x is $\frac{0.05589}{0.04162} = 1.343$, the same test statistic except for rounding errors.

Each test is effectively of whether x and y are linearly connected when allowance is made for their linear time trends.

The question asked for the importance of DW in the *seven* regressions, including the first one. For this first regression DW = 0.82, which leads to its rejection as a viable economic model and confirms the importance of the lagged values of GDP.

Note that (d) and (f) are different forms of the same relation. Comparing (d) and (e) gives a statistically insignificant result, whereas (c) is significantly preferable to (e) and is perhaps the best model of those fitted.

Question A3
(i) P-e ratios take no account of expectations – of growth of earnings, of safety, etc.

(ii) 0.3529 is distributed $F_{2,57}$, and is not significant. On the assumptions of analysis of variance, the null hypothesis that $\mu_1 = \mu_2 = \mu_3$ is sustained.

(iii) Normality and homoscedasticity are assumed. But the data are obviously positively skewed and are truncated at zero. Estimated variances are 30.41, 81.86 and 116.59, which do not immediately confirm the assumption of homoscedasticity.

(iv) The result of the analysis of variance was so conclusive as to lead one to take the test as being sufficiently robust.

(v) Take logarithms (or perhaps square roots). (It can be found that taking logarithms resolves these possible problems, though candidates were not expected actually to do so.)

(vi) Many answers were utterly and obviously wrong, so much so that candidates should have rejected them as absurd. The usual error was in confusing the estimated standard deviation of the population, $\sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$, with the estimated standard error of the sample mean, $\sqrt{\frac{\Sigma(x-\overline{x})^2}{n(n-1)}}$. Candidates' confidence intervals were thus far too wide. In fact $15.1 < \mu < 21.5$ is correct.

(vii) The Kruskal-Wallis procedure tests the hypothesis that all observations come from the same distribution.

Question A4
This apparently elementary question revealed widespread ignorance of the standard formula for estimating a median, and an inability to draw histograms for grouped data with unequal class intervals.

Note that, for example, the age group recorded as 20–44 is of width 25 years, and has mid-point 32.5 years. The length of the Scotland male 0-9 bar of the pyramid is proportional to $\frac{313}{(2487+2633)} \div 10$, or 0.006113, and the 20-44 bar is proportional to $\frac{933}{(2487+2633)} \div 25 = 0.007289$.

Graduate Diploma Option: Econometrics

As the aim is to assess the candidates' knowledge of the basic concepts and techniques of econometrics, the questions were set to cover three important 'classical' topics: simultaneous equations (question B1), random regressor models and instrumental variable estimation (question B2), and simple heteroscedasticity, bias and efficiency (question B3). More recent topics in time series analysis were covered by the last question.

Understandably, the vast majority of candidates chose question B4. This question had more options; candidates who are not highly technical or are not confident about answering other questions with more specific requirements are generally well advised to tackle this type of question. The next most popular question was question B3. Again, the apparent ease of the question was probably behind its selection. However, only 9 out of 20 marks were on relatively easy material, and only a handful of candidates were successful in the last two parts of the question. Question B2 appeared more difficult but was in fact easier, especially since the instrumental estimator formula for the slope was given (and hence that of the OLS could be easily 'guessed'). Only one candidate attempted this question. The first two parts of question B1 were straightforward and easy. Only part (iii) was challenging. While other questions did not appear to cause misunderstanding, a few candidates attempting to answer part (ii) of question B1 misunderstood what was required (to show that the parameters can be expressed in different ways) and performed identification tests.

Out of 15 candidates who attempted these questions, two were above average to excellent, and four were below average.

Question B1
This type of question is of the 'either you know it or you don't' type. Most candidates got very low marks. The question needed to be read particularly carefully, to avoid misunderstanding part (ii).

Because the table of cross-products was given, one can reasonably 'guess' that the estimation required would not involve lengthy matrix inversion. All candidates needed to remember was the formula for an OLS estimator.

Question B2
Only one candidate answered this question, which required familiarity with instrumental variable estimation and probability limits. The other derivations and computations required were straightforward. To help the candidates, the formulae were given for the instrumental case, and a candidate should not find it difficult to convert these to the simpler Ordinary Least Squares case. The last part required substituting the values given as cross products.

Question B3
This appeared to be the easiest question. Indeed, parts (i) to (iii) were easy, although not all candidates did well in them.

Part (iii) required candidates to show that the estimator is a GLS estimator. This involved transforming variables so as to obtain an error term with constant variance. So one needed to find the required transformation, write down the new (transformed) equation, and then show that the error terms have a constant variance, so that an OLS estimator would be suitable. Alternatively, one can obtain a matrix of weights, and apply the matrix formula for GLS to obtain the estimator suggested in the question.

Many candidates misunderstood what was required in part (v). What was needed was to describe a real-life situation; for example, the stock market (variance changes from low in calm periods to high during crises and crashes), or seasonal changes in consumption (Christmas, summer/winter), or household income (higher income households may have a larger variance than low income households).

Question B4
This question had the highest average score. Most parts were understood and answered quite satisfactorily. However, the answers to parts (b) and (c) were generally poorer.

Graduate Diploma Option: Operational Research
Graduate Diploma Option: Medical Statistics
Graduate Diploma Option: Biometry
Graduate Diploma Option: Statistics for Industry and Quality Improvement

The numbers of candidates for these components of the paper were very small, and it is therefore not possible to give detailed reports.