ROYAL STATISTICAL SOCIETY

ROYAL STATISTICAL SOCIETY EXAMINATIONS, 2012

REPORTS OF EXAMINERS

GENERAL COMMENTS

This report incorporates the comments made by examiners after marking the papers set in 2012 at all levels of the Society's examinations (Ordinary Certificate, Higher Certificate and Graduate Diploma). We would encourage all candidates intending to take the examinations in 2013 or subsequently to refer to the particular comments on the papers they expect to sit, as this is the primary means by which their examiners can communicate with them. We would also remind candidates that past papers (or specimen papers for new examinations) and reading lists are provided on the RSS website, and strongly suggest that all candidates will wish to make use of these vital resources as part of their preparation.

All levels of the Society's examinations are now fully modular. Candidates, and those advising them, should be aware of the benefits of a modular structure. Candidates do not need to sit all the modules at a particular level in the same year; indeed, we anticipate that only a small minority of candidates will do so. Candidates are most likely to be ultimately successful in passing at a particular level if they are realistic about the amount of time they have available for study and enter for an appropriate number of modules.

Most comments made by examiners refer to specific features of questions set in this year's papers. But every year examiners draw attention to general aspects of examination technique that could be improved. As we have noted in earlier reports, it is disappointing to see candidates losing marks unnecessarily. Several comments made by examiners in 2012 echo those made in other recent years. We therefore repeat here the advice to candidates given at the start of previous years' reports, revised so as to incorporate further general comments made by examiners following the 2011 papers.

The published syllabuses for the Higher Certificate and Graduate Diploma give details of mathematical topics with which candidates at those levels are expected to be familiar before embarking on study for the Society's examinations. You must make yourself aware of the necessary mathematics background for the examinations you intend to sit, and make every effort to master it. This year again, candidates lost many marks as a result of their poor skills in algebra and calculus.

Read any question you intend to answer slowly and carefully, and ensure you answer the question actually asked. Every year, some candidates reproduce bookwork that may have some relation to the topic but does not answer the question itself. Examiners award marks in accordance with detailed marking schemes, which assign marks for specific answers to each part of each question. There is therefore no point in writing down what you know about a different (if similar) topic, since the marking scheme will have no marks available for this.

On a related matter, be sure to carry out any specific instructions given in a question: e.g. round answers to three significant figures if that is what is asked; calculate the standard deviation, not just the variance, if that is what the question requires.

Take note of the number of marks allocated to each part of a question, as printed on the examination paper. It is a waste of your time writing a detailed two-page description of some topic, if this can only be awarded two marks.

When preparing for an examination, you will of course know that there will be certain details (definitions, formulae and the like) that you will be expected to have memorised. For any paper, candidates will be expected to know the definitions of all concepts relevant to the syllabus. As for formulae, it will be clear that (for example) a candidate who does not know the formula for a binomial probability function cannot fully understand the binomial distribution, so examiners may expect candidates to be able to quote that probability function when it is relevant to a syllabus. Similar examples can be given for other areas; formulae for sample variance and conditional probability (at Ordinary Certificate level) and sums of squares for appropriate analysis of variance models (at higher levels). In recent years, examiners have regularly pointed out that candidates had quoted key formulae incorrectly and therefore gone badly wrong from the start of a question.

Make sure you understand the difference between the instructions *explain* and *define*. An *explanation* of some concept requires one or more sentences; the concept concerned should be described in words and (if appropriate) the purpose or use should be outlined. In a mathematical examination, a *definition* is a short and precise statement, which may require the use of mathematical notation. If a definition is required, a rough description is likely to be awarded no marks.

Ensure that you include sufficient reasoning in your answers for the examiners to be sure about the basis for any conclusions you draw. For example, writing ‘the test statistic is greater than the value in tables’ without stating the value, the relevant sampling distribution or the degrees of freedom, will gain very few marks, if any.

In questions requiring calculations, it is understandable that errors will be made under examination conditions. When a candidate shows his or her working clearly, it is possible to give credit for use of a correct method even if there are errors in the numbers presented. However, when little or no working is shown, it is rarely possible to assess either the method being used or the source of the error. Candidates are therefore strongly advised to show sufficient working to make it quite clear which method is being used.

Be aware of the *RSS statistical tables* that are provided for candidates during all the examinations. This year, several examiners commented on the amount of unnecessary work that candidates had given themselves because they did not simply look up values in these tables. You may freely download a copy of the tables to use during your preparation for sitting the examinations, and you are strongly advised to get to know what is in these tables and how to use them efficiently.

In calculations with several steps, it is important not to round intermediate answers to too few significant figures. For example, if the final answer is to be quoted to three significant figures, then at least four significant figures will need to be retained for intermediate answers.

When you complete a calculation, or finish answering a practical part of a question, try to check the plausibility of your result. For example, a variance cannot be negative, and a correlation coefficient cannot be outside the range -1 to +1. Similarly, a trend or regression line that does not pass through the main part of the data points is very unlikely to be correct.

If a rough sketch diagram is required, this can be provided in your answer book; there is no need to draw it accurately on graph paper. This might for example apply to a sketch of a probability density function. Of course, such sketches must always be sufficiently clear so that salient features stand out properly. However, when an accurate graph or chart is required, this should always be done on graph paper; and you should make sure you include a title and label the axes. This might for example apply to histograms.

Year after year, examiners comment that many candidates seem a lot more comfortable with calculations and graphs than with discussion or reports. Applied statisticians need to develop excellent communication skills, so the RSS examinations assess these as well as arithmetic and mathematical skills. You should practise answering discussion questions, possibly using past papers and solutions as a guide; you will find it helpful to talk through your answers with a knowledgeable person, a tutor or a statistician you work with.

It is important to follow the instructions on the front cover of the answer book. We realise that candidates will have little time to spend on reading the front cover during the examination itself, so we have produced a copy you can consult on the Society’s website. You are strongly encouraged to look at this before the examination, and to ensure that you follow the instructions. We draw your attention to the following instructions in particular:

1. Begin each answer on a new page. (You do **not** need to begin each **section** of an answer on a new page.)
2. Write the number of each question at the top of each page.
3. Graph paper should be attached opposite the answer to which it relates.
4. Enter in the space below (**not** in the side panel) the numbers of the questions attempted. (The question numbers should be written **in the order in which you answered the questions**. Note that the side panel is for the examiners' use only.)

It is also helpful to examiners, as well as simpler for candidates, when the answer to a question is written on consecutive pages of the answer book. We do realise that, in practice, candidates may sometimes need to return to a question later. If you do this, then it is helpful if you indicate this clearly on the page where the earlier attempt was made.

ORDINARY CERTIFICATE IN STATISTICS

Module 1 (Collection and Compilation of Data)

This examination was designed to test that candidates understood the basic concepts in the syllabus and were also able to apply them to real or realistic settings. Most of the questions were therefore set in the context of a survey. Answers should have been given in the context of the survey – not just as statements of generic textbook principles. It was pleasing to see candidates having the confidence to express their own opinions and advice about realistic practical issues – taking into account what was actually involved in the given scenario. For example, there are general issues about self-completion questionnaires compared with telephone interviews, but some of these are largely irrelevant when we are considering a survey of hotel clients. Candidates who scored well were able to describe or imagine the actual setting and the associated practical issues and then comment on the implications for a survey. An examiner may not always agree with a recommendation, but if it is well-argued then it gets marks. This was especially pertinent to question 7. Similarly, question 1 asked for information about a real survey. Some candidates gave very full descriptions of surveys and scored highly: others just repeated bookwork about surveys and scored nothing.

Not many candidates understood what was meant by ‘precision’ and many confused this with ‘accuracy’.

Question 8 was more numerical than the other questions. Candidates seemed to find this a hard question, and were unable to do the computations. Again, it is important to be able to use formulae not just to be able to quote them.

When answering a question it is wise to look at the number of marks allocated to it. If there are only 3 marks then you cannot score more than 3/3. On the other hand, if there are 8 marks then you really need to be making 8 points in your answer to score 8/8. This should guide you in the length of your answer.

In preparing for this examination, candidates are strongly advised to study real surveys and to relate the theory and guidance to practical situations.

Module 2 (Analysis and Presentation of Data)

Most candidates attempted all the questions. There were a few excellent scripts from candidates with an impressive grasp of the technical demands of the paper and an ability to offer perceptive comments on the data. At the other extreme there were some scripts from candidates who seemed completely unprepared for a paper at this level. About half the candidates scored marks in the range from 30% to 60%. In most cases they showed a commendable understanding of at least one topic, but also showed little or no understanding of topics of similar difficulty. Perhaps such candidates had not covered the whole syllabus.

This module is about both the *presentation* of data and the *analysis* of data: the former was often done much better than the latter. It is very important to realize that, without a proper context for understanding, data are just collections of meaningless numbers. When meeting a set of data, students should ask themselves what the story is: what the story is behind the collection of the data and what the story is that the data are able to tell. It is that level of analysis being tested in this paper.

Question 1
Though the terms examined here – median, quartiles, inter-quartile range – are elementary, this question proved difficult for many candidates. In particular, many were not able to calculate the quartiles correctly. (A common error was to take the 6^{th} and 18^{th} items from the 24 item data set.)

Many boxplots were of poor quality. Graph paper should be used and the scale should be accurate. A 'generic' boxplot with the 5 significant values marked on it but no attempt to have these values to scale is not acceptable.

Question 2
The mean was generally calculated correctly, but there were many very strange reasons given for why this mean would under-estimate the patients' true mean age. It is nothing to do with skewed data, outliers, having two patients with same age or people mis-stating their ages. Quite simply, ages in years do not take into account the months since the last birthday. A better estimate would be obtained by adding 0.5 to the calculated mean.

Question 3
Though most candidates knew what the coefficient of variation was, almost nobody could say why it is not appropriate to use it with temperatures. The important thing here is that zero on a temperature does not represent the absence of anything: it is an arbitrary point. This is clear once a set of temperatures is converted from one scale to another. E.g. a set of temperatures with a mean of 32 and a standard deviation of 9 on the Fahrenheit scale would translate to a mean of 0 and a standard deviation of 5 in Celsius. Attempting to calculate the two coefficients of variation immediately shows up the problem.

The coefficient of variation (in suitable cases) is usually defined as the standard deviation divided by the mean, but sometimes a factor of 100 is applied. It is very important that candidates make it clear which definition they are using. Without such clarity they may lose marks.

Question 4
There was quite a lot of rather careless work in this question. Graphs were, in many cases, of poor quality; one common mistake was to set up scales which could not accommodate the data for the United States. It is *not* adequate to put a break in the scales or to indicate in some way that the data point for the United States is some way off the page. The point of a graph is to give a clear visual impression of the data.

The ratios were sometimes calculated 'upside down', and sometimes misinterpreted.

The rank correlation coefficient was often calculated correctly. Candidates sometimes seemed to be guessing whether the Pearson coefficient would be larger or smaller: the United States will be even more of an outlier when the data are not ranked.

Removing the United States data point gives a rank correlation of zero. The comment looked for was that an outlier can have a big effect on a correlation coefficient.

Question 5
The probability question was the least popular on the paper with some candidates omitting it and many making rather half-hearted attempts at it. Some, however, worked through it efficiently and with good understanding of probability formulae, so scoring high marks.

The comment part at the end was done well, even by some who made no progress with the calculations. This was a case in which many candidates did see the underlying story: it is likely that the test will be given mainly to those with symptoms or with a genetic pre-disposition to the condition.

Question 6
By contrast with the previous question, this was one in which many candidates failed to get to grips with the story.

The technical requirements of the question are not great – just simple arithmetic – and most of the marks were for commenting on what the data indicate. In part (i) the percentages indicate that females are more likely to be low risk than males *and* more likely to be high risk than males. Males are more likely to be medium risk than females.

In part (ii) candidates are asked to comment on how risk varies with age. The percentages of people who are low risk increase with age; the percentages of people who are high risk decrease with age; the percentages of people who are high risk fall then rise with age. Many candidates, perhaps the majority, completely failed to see this pattern because they looked at the absolute numbers in each cell of the table rather than the proportions these numbers represented.

In part (iii) the comment sought was that the data could be analysed by sex *and* age.

Question 7
Once again, many candidates failed to see the story here – or saw a story that owed much to their own imagination. So, for example, it was common to read that sales rose each year from Q1 to Q4: this is not true as Q4 sales are sometimes lower than Q3. It was common, too, to be told that there is a rising trend across the series. Again, not so: there is a rising trend for the first half of the series, but it then broadly levels off.

Many candidates struggled with the missing values, *a*, *b*, *c*. Many appeared not to see (or chose not to respond to) part (iii). A substantial minority thought that ‘a suitable graph’ for looking at the trend would be of the *un*-smoothed data. Those who hedged their bets by drawing the un-smoothed and the smoothed data were given credit.

Question 8
This was a straight-forward question for those who knew the definitions, and were accurate in their calculations. Unfortunately there were many who fell down on one or the other of these requirements.

Rather as in Question 3, the price relatives were often given with a factor of 100 built in. This is perfectly acceptable provide it is made clear, but any lack of clarity may be penalized.

HIGHER CERTIFICATE IN STATISTICS

Module 1 (Data collection and interpretation)

In general candidates need to pay more attention to answering the questions as set.

Question 1
This question was on finding summary measures for a grouped frequency distribution given a choice of different midpoints and how measures would change given two outliers. Not all candidates correctly chose Peter's choice of mid-points as the best method. No candidate transformed the midpoints to such as -2, -1, 0, 1, 2, 3 to find the mean and variance. In part (iii) not everyone wrote a short comment on the ages of patients seen by the doctor on one day.

Question 2
Part (a) of this question was on stratified and cluster sampling, how to choose a random sample using random number tables, and a comparison between a quota sampling method and a postal method. Part (b) asked for a plot and a comment on the plot. Many candidates omitted one of the parts of (a)(i). (a)(ii) was disappointing as very few candidates appeared to know how to use random number tables to draw a sample. In (b)(i) a handful of candidates drew incorrect diagrams instead of the time series diagram expected.

Question 3
Part (a) of this question asked candidates to construct questions for a specified mail survey. Part (b) asked for the main differences between questionnaires designed for mail surveys and those for use with an interviewer. The questions were well designed for the most part but some candidates did not pay sufficient attention to the questions requested. Part (b) was reasonably well done.

Question 4
This question was on non-response bias and response error. In (i) of this question too, insufficient attention was paid to the wording. Not everyone had the right idea in (iii).

Module 2 (Probability models)

Candidates generally coped satisfactorily with this paper, which may have been slightly easier this year. A good proportion of distinctions was seen (30%), but there continues to be a undesirably high proportion of mathematically weak candidates. Several candidates showed very uneven quality of coverage, i.e. very good and derisory answers to different questions on the same script. Questions 1 and 2 (discrete probability and combinations; Normal distribution) were satisfactorily done (scores averaging about 13/20); questions 3 and 4 (Poisson and exponential) were weaker, averaging about 10/20.

Question 1
A fairly popular question, attempted by 76% of candidates, average score 13/20. Part i) was usually good, but statements of the probability mass function were sometimes inaccurate or else omitted. There were many good answers to part ii), but several weaker students did not understand how to set up the probability distribution of $Z = XY$. The notion of mode was usually well understood. Part iv) was less good: very few candidates articulated a clear argument for Var(Z) using the independence of X and Y, and many attempts to work from the distribution of Z went wrong due to the heavier arithmetic involved.

Question 2
The most popular question, tried by over 90% of candidates, average score 13/20. Parts i), ii) and iii) were generally well done, albeit with a sprinkling of '1 –', sign or tail errors or other arithmetical slips. However, there was significant confusion over the calculation of var(W) in ii) and var(V) in iii). The sample size calculation in part iv) was less well done, often apparently because of inability to deal with the necessary algebraic manipulation.

Question 3
The least popular question, attempted by 61% of candidates, averaging just under 10/20. Parts i) and ii) yielded many good answers, although errors in dealing with 'at least 2' as the complement of '0 or 1' were common. Many students struggled with part iii): binomial distributions were sometimes mentioned, the mean and/or variance of N_{20} were often wrong, the continuity correction for the normal approximation was often omitted, and very few appreciated the failure of the normal approximation to represent adequately the right tail probabilities of the (positively skew) Poisson distribution.

Question 4
This question on the exponential distribution and conditional probability was tried by 67% of candidates, who achieved an average score of just over 10/20. Part i) revealed an often poor command of integration, and of the manipulation required to deduce the median of X. Part ii)a) was usually good, but in parts ii)b) and c) there were surprisingly many errors. Several students reverted to treating X and Y as integer-valued, possibly confusing the exponential with the Poisson distribution. There were few correct answers to the conditional probability calculation in part iv).

Module 3 (Basic statistical methods)

Examiners' reports, necessarily, focus on what candidates have done wrong. It is appropriate, therefore to preface that list by reporting that candidates for this paper did rather well.

Statistics examinations provide a stern test, since they require a mix of modelling, technical and interpretative skills. Candidates are generally well prepared for this, but it can be understandably difficult for them to tease out exactly what is required under the pressure of a timed, written examination. Thus there is often a kneejerk reaction to a few of the words ..."Ah! I need to do some interpretation here" ... which can lead to rather aimless waffle.

There can be no better example of this than in question 3, where the final part of the question required candidates to explore the relationship between a goodness of fit test, and the associated physical scenario. Only one candidate really got to grips with this. The rest made unfocused comments on the physical situation only.

So, not unusually, the most important advice to future candidates is to **read the question**!

Question 1

(i) Far too many candidates wasted time by computing 90% and 99% confidence intervals, instead of interpreting the output.

(ii) Many candidates neglected to mention that the sampling needed to have been random. Of those that did note that a t-distribution had been used, many failed to specify which t-distribution, and few were able to explain why.

(iii) One might have thought that candidates who referred to repeatedly sampling might have understood the issue. In fact many who got that far failed to convince by referring specifically and only to the given interval.

(iv) Most candidates answered this correctly.

(v) Quite a lot of candidates amused themselves by computing s. Most of them successfully arrived at 13.904, as had been given, 3 times, on the printout! However, most candidates had been well drilled in this piece of bookwork, and had good enough algebraic skills for it.

Question 2

(a)(i) Few candidates argued whether a one-tailed test might be appropriate.

(ii) Most gave good answers on types of error.

(iii) Very few candidates managed to get their words right, often making marking difficult. There were many cases where it was clear that they had not said what they meant to say. For instance, "With $\alpha = 0.05$ a type 1 error is more likely."

(b)(i) The majority of candidates ignored the fact that the standard deviation was known, computed the sample standard deviation, and performed a t-test.

(ii) Most identified that the test would be one-sided. Only half of those reasoned that the null hypothesis would certainly not be rejected.

(iii) It was pleasing to see how many candidates could successfully manipulate the algebra in this part question.

Question 3

(i) Some candidates failed to "confirm". For instance, some wrote down correct expressions involving binomial coefficients, but then jumped straight to the given answer without showing any computations.

(ii) Candidates were told, quite explicitly, to perform a goodness-of-fit test. Their hypotheses more often than not demonstrated that they had not read the question carefully enough, very often referring to the MD's bias or lack of it. The majority failed to amalgamate classes appropriately.

(iii) The comments in the preamble are repeated here ... the final part of the question required candidates to explore the relationship between a goodness of fit test, and the associated physical scenario. Only one candidate really got to grips with this. The rest made unfocused comments on the physical situation only.
Of course, candidates who had not formulated their hypotheses correctly in part (ii) were not in a position to proceed sensibly with this part of the question.

Question 4

(a)(i) Few referred to median, and many specified a two-tailed test.

(ii) Candidates scored well on this.

(b)(i) Very few candidates were able to formulate hypotheses which were appropriate for this paired data scenario.

(ii) Again, candidates generally scored well on this part question.

Module 4 (Linear models)

Candidates generally coped well with this paper, which may have been slightly easier this year. A good proportion of distinctions was seen (31%). A very widespread shortcoming in answers to this ‘applied’ paper is the lack of contextualised conclusions given by candidates in the light of their analyses of the problems set. Question 1 (balanced ANOVA) was the most popular, being attempted by 83% of candidates with a high average score of 14.5/20. Question 4 (simple regression and correlation) was almost as popular (tried by 81% of candidates, albeit with less success). Question 2 (multiple regression, interpretation of computer output) was attempted by 69% of candidates, with a satisfactory average score of 13/20. Question 3 (simple linear regression and linearization) was least popular (tried by 62% of candidates) and less well done (average score 11/20).

Question 1

This most popular question was tried by 83% of candidates, who achieved a high average score of 14.5/20. In part i) statements of the model and standard assumptions were usually good, but the terms *balanced*, *randomised* and *replication* were often not clearly explained. The relatively simple calculations in part ii)a) were usually well done, although a context-based conclusion was often lacking. Part ii)b) was less good, with many students being unsure how to calculate the *t*-interval asked for.

Question 2

This question was tried by 69% of candidates, with average score 13/20. Part i) on the model statement for a regression on two variables was usually good. In part ii)a) some attempts wrongly used a two-tailed test for the global test of the full regression. The (two-tailed) partial *t*-tests were usually correct, albeit with occasional confusion as to degrees of freedom; contextual conclusions were rare. The calculations in part ii)b) were usually sound, and generally sensible comparisons of the models were made in part ii)c).

Question 3

61% of candidates attempted this less popular question, achieving an average score of just over 11/20. In part i) the graph plot was usually satisfactory, but few candidates recognised possible nonlinearity or increasing scatter. Several candidates found it difficult to express correctly the log transformation required in part ii), and in calculating the transformed regression in part iii) errors were common, perhaps due to carelessness in dealing with the 'difficult' numbers involved. Comparison of the untransformed and transformed regressions in part iv) tended not to focus on the nonlinearity of the original model if this had not been noted earlier.

Question 4
This question on association (regression, product-moment and rank correlations) was tried by 81% of candidates but with a modest average score of 11.4%, who achieved an average score of just over 10/20. Part i) revealed widespread confusion about dependent and independent (regressor) variables and about the stochastic and structural assumptions made in the four models considered. In parts iii) and iv), many students coped sensibly with the calculation of the product-moment and rank correlations, and with the tests, but several used misremembered formulae and/or two-tailed tests. A few candidates wasted time by calculating the *t*-test for correlation rather than using the RSS table. Comparisons of r and r_s in part v) tended to ignore the curvature of the scatter diagram if this was not noted earlier.

Module 5 (Further probability and inference)

Question 1
A generally poorly answered question, owing mainly to weak calculus skills, especially with regard to integration by parts. Also, some candidates were careless with the upper limits of integrals with values like "x", "y" and "1" appearing instead of "∞". In part (iii), it is not correct to claim that P(W=V=1) = P(W=1) P(V=1) since W and V are not independent.

Question 2
The best answered question, though again poor calculus skills were sometimes exhibited when differentiating in part (b) (ii). Some candidates did not recall that $\pi_W(t) = \prod_i \pi_{Y_i}(t)$. The last part, in which the pgf is used to find P(W=k) was not well answered.

Question 3
Part (a) was surprisingly poorly answered for a descriptive question on an important topic. Note that a rough guide to the length of this type of question is to write approximately two lines per mark. Part (b) was again often poorly answered; candidates realised that the Binomial distribution was involved but did not read the question carefully enough.

Question 4
Quite an unpopular question despite it being quite standard, with some very good answers. Note that in part (iii), it is only necessary to show that $E(\hat{\theta}Y) = E(\hat{\theta})E(Y)$.

Module 6 (Further applications of statistics)

Question 1
There were several quite reasonable attempts at this question but in part (ii) quite a few candidates lost marks for forgetting that the second sample may not be needed. So the contribution from the second sample is only included conditional on its actually taking place. (Tree diagrams can help here).

Question 2
Again there were several good answers, and one place where several marks were lost. The analysis of variance must include a row for the sum of squares for AB, calculated as SS Treatments minus contributions for A and B. The degrees of freedom for AB are the product of those for A and those for B. If that is omitted the degrees of freedom and the sums of squares are wrong, so the residual and its d.f. are also wrong. And when there is an interaction it is wrong to make comments about main effects; the information lies wholly in the interaction. This is where a graph is valuable as well as the numerical analysis.

Question 3
A surprising number of candidates could not derive the OLS estimates of the parameters in a linear regression model. Some found the equations and did not solve them, a few tried to include the residual term e in the analysis, instead of minimising the sum of squares of the e's. Once the first model was fitted the second one was done very carefully. When a candidate knew about computer analysis the answers to the final part were good, but quite a few gave no evidence of being familiar with computers.

Question 4
Dummy (Indicator) Variables are obviously not always understood, but marks could be obtained for the graphs although quite a few people drew graphs that did not meet the requirements specified. Linear models are <u>linear in the parameters</u>, as the best candidates knew but many others did not. And several did know about transformations, in this case log to base e. Several did not, however, know how to linearise power models and how this is done. In part (b)(ii) only the second one is non-linear.

Module 7 (Time series and index numbers)

The answers to questions 1 and 2 leave the impression that the candidates struggled more with subjective questions than objective questions. In general, the more guidance there is in a question, the easier it should be to answer. However, there is clearly a balance to be struck as too much guidance is tantamount to spoon-feeding and the exam then becomes less a test of understanding than of the ability to follow instructions.

In question 4, a number of candidates focussed on relatively few aspects of setting-up and producing the new index, instead of giving an overview of the wider range of tasks involved. The question goes into some detail about available resources.

Candidates should generally assume that all information given in a question is of relevance to their answer.

Module 8 (Survey sampling and estimation)

All candidates attempted 3 questions. The most popular questions, attempted by nearly all candidates, were question 1 on simple random sampling and question 2 on stratified random sampling. A similar number attempted questions 3 and 4, as their 3rd question. The paper, in general, was satisfactorily done (with a few exceptional candidates). Higher marks were obtained on questions 1 and 2.

As in previous years, candidates lost marks on the more descriptive parts by not appreciating the practical implications of aspects of the situations described in the question. Too often, answers were much longer than necessary. Answers need to be more targeted, and aligned with the marking scheme.

Candidates are reminded that they are expected to memorise certain formulae as indicated in the syllabus.

Question 1

This question was done well. Most candidates were able to construct confidence intervals for a true proportion and a mean under simple random sampling. A few candidates calculated the variance of the sample proportion, p, wrongly as $(1/n)\ p(1-p)/n$ [analogous to $(1/n)\ s^2$, for the sample mean] rather than $p(1-p)/n$. Part (i)(b) was often omitted, suggesting candidates did not know much about sample size or how to rearrange the formula for a 95% confidence interval to obtain an expression for n.

In part (iii), the use of systematic sampling would require the list of athletes to be arranged in random order, but hardly anyone mentioned this. Nevertheless, there were some good suggestions on ordering the list, which was pleasing, e.g., by registration entry number. Some candidates misread the question, and wrote about the process for selecting a systematic sample, and lost marks.

Question 2

This question was done well generally. Most candidates were able to calculate an estimate of the population mean under stratified sampling, but some seemed unsure about the calculation of the standard error. Other candidates did not read the question properly, and constructed a 95% confidence interval for each stratum separately, and lost marks.

Answers to part (ii) were disappointing. The samples sizes in the strata under proportional allocation were generally calculated correctly. However, marks were lost on the more descriptive parts because candidates often wrote more generally on these topics, rather than looking at the data, and discussing this. The very different sample means and standard deviations in the strata should have alerted candidates to the benefits of using stratified sampling, compared to simple random sampling.

Part (iii) was a straightforward application of simple random sampling. Some candidates lost marks by not using a finite population correction in the calculation of the variance, i.e., the sampling fraction ~12%. Others used the critical value of the Normal distribution which was inappropriate with a sample size of 11.

Question 3

Answers to this question were rather disappointing.

Part (i) required candidates to calculate a standard deviation from the raw sums and sums of squares of y, i.e., $(\sum y^2 - (\sum y)^2/n)/(n\text{-}1)$. Most candidates made mistakes.

In part (ii), only a few candidates mentioned the likely correlation between the number of cattle and the total area under cattle for each farm, as a reason for considering either a ratio or regression estimator to be appropriate for these data. Those who attempted part (ii)(b) used various approaches to comment on the relative standard errors, including relative efficiency, which was pleasing.

Answers to part (iii) were often disappointing. Candidates wrote about stratification and clustering in general, but not necessarily on how and why it might be useful for such a survey, and what issues it might overcome, e.g. in rural areas it is unlikely that maps or sampling frames would be available.

Question 4

This question, in general, was satisfactorily done.

Candidates were able to explain the difference between a cross-sectional sample survey and a longitudinal study, although some did not realise how these differed in population and use, i.e., looking at changes over time. In discussing the advantages and disadvantages of using a longitudinal study, many candidates wrote more generally, rather than specific to the proposed survey. They did not, for example, comment on the ease of selecting a sample from last year's graduates, as an advantage, nor the study taking 7 years before full results are available, as a disadvantage.

There were some good answers to part (ii)(b). To gain full marks, candidates were expected to cover the various aspects of survey methodology, as outlined in the question. Often, candidates wrote in detail about one or two aspects, and lost marks.

GRADUATE DIPLOMA IN STATISTICS

Module 1 (Probability distributions)

Of 33 candidates who submitted scripts, 12 passed and 21 failed. Three candidates scored very high marks, but the overall standard was disappointing, with a number of scripts very weak indeed. Candidates did not fail because they ran out of time: they failed because their knowledge of the syllabus was inadequate. Some were careless in reading the questions, by giving answers related to the subject matter, but not following the actual instructions – see, for example, comments on questions 2, 4 and 6 below. Questions 3, 4 and 6 were most popular, Q8 was largely ignored.

Question 1
This question tests the ideas of independence. Despite part (ii) asking for a demonstration that events A, B and C were *not* independent, several candidates answered part (i) by making the assumption that those three events *were* independent. Many answers were sloppily presented – for example, writing as though an event, and its probability, are the same thing. Those who understood what part (b) was about generally made good progress, but many candidates wrote irrelevant material.

Question 2
Here mathematical manipulation let down several attempts to compute the skewness of the Poisson distribution. Asked to find E(1/(X+1)), some made the basic error of assuming it was the same as 1/(E(X)+1). For the final part, few candidates looked at the ratio of successive probabilities (for a fixed λ), choosing instead to see how the probability of outcome k changes as λ changes. This is not answering the question set.

Question 3
Finding the mean and variance of a continuous Uniform distribution should be simple at this level, but many attempts contained slips, both in logic and in mathematical manipulation. Those who correctly used the given hint for the second part generally did well, but several candidates obtained a negative value for a variance, blithely continuing as though this were correct.

Question 4
This question asks for the pgf of a geometric distribution, and to *calculate* a mean and a variance. Omitting any mention of a pgf, and simply *stating* the values of those quantities will not accrue marks. Some candidates thought the phrases “Bernoulli trials” and “Bernoulli distribution” are identical in meaning.

Question 5
A key point is that, when U has the continuous uniform distribution over [0, 1], then the variable 1-U has the same distribution; few candidates specifically noted this fact. The use of antithetic variates entered the syllabus under the new arrangements, and several candidates showed good preparation for this type of question.

Question 6
Part (i) instructs candidates to use an mgf to obtain answers: finding these answers by *different* means does not carry credit. Manipulation of the mgf in parts (ii)/(iii) caused problems, and it was surprising to find that, at the very end, about half the candidates who gave answers scaled their variable by dividing by the variance, not the standard deviation.

Question 7
The material of this question also specifically entered the syllabus recently, and about half the candidates made some attempt at an answer. In part (ii), demonstrating that a turning point was indeed a maximum was often omitted.

Question 8
Most attempts scored zero marks, making no progress on any of the three parts; but one candidate (maybe knowledgeable enough to recognize this as “Bertrand’s Paradox”) gave an excellent answer.

Module 2 (Statistical inference)

Question 1
A popular question with some very good answers. In part (ii), candidates should be aware that the CRLB for $e^{-\lambda}$ can be deduced from the CRLB for λ.

Question 2
A popular, generally well answered question, though the derivative with respect to *x* of 2^x caused some candidates problems.

Question 3
Parts (ii) and (iii) were generally well answered. Surprisingly, the likelihood plot in part (i) was incorrectly drawn by most candidates and some candidates incorrectly tried to use a Normal approximation in part (iv).

Question 4
A popular, but not generally well answered question. Several candidates incorrectly wrote the likelihood ratio as $L(\alpha)/L(\beta)$. Part (iii) is an example of the use of the relationship between hypothesis testing and confidence intervals.

Question 5
The three shorter question format did not prove popular. Part (b)(i) was answered well but not so parts (a) and (c) despite both being very standard questions. Even candidates who otherwise answered part (a) correctly, did not realise that the maximum difference was 0.3, occurring just prior to 0.6.

Question 6
A moderately popular, generally well answered question. Note that in part (ii) it is necessary to evaluate the expectation of the probability with respect to the posterior distribution.

Question 7
Few candidates attempted this question, suggesting that this topic is often overlooked.

Question 8
A popular question; answers were usually good but rather too short. As well as its direct uses (such as the Normal approximation to the Binomial), the Central Limit Theorem is at the root of several asymptotic results, such as the Normality of the maximum likelihood estimator for large samples under regularity conditions.

Module 3 (Stochastic processes and time series)

This year there were 5 questions on stochastic processes and 3 on time series. As usual, the time series questions were the most popular. Indeed, all candidates attempted question 6 on the infinite moving average representation of an AR(2) process. Most candidates were able to verify stationarity through examination of the roots of the autoregressive characteristic equation, but many candidates were not able to successfully atttempt the rest of the question. Many candidates made a reasonable attempt at the more practical questions 7 and 8, on ARIMA modelling and exponential smoothing, respectively, but it was disappointing how many in question 7 were not able to discuss clearly the use of the acf and pacf to identify appropriate models to fit to the data.

The most popular of the stochastic processes questions was question 2 on the properties of a 4-state Markov chain. Many candidates were able to calculate successfully the 2-step probabilites in part (ii), but it was disappointing how few had a sufficient understanding of the concepts of transient and recurrent states required in part (i). The least popular was the other question on a Markov chain, question 3, but those candidates who attempted it gave some very good answers, including the evaluation of the stationary distribution.

What was a fairly routine question on a branching process, question 1, was attempted by fewer candidates than might have been expected. Question 4, on the properties of a continuous time Markov chain model, was very well done by a few of the candidates, but question 5, which involved concepts that arise in the discussion of the embedded Markov chain for an M/G/1 queue, was, perhaps unsurprisingly, the one that produced the fewest good solutions.

Module 4 (Modelling experimental data)

Question 5
Surprisingly few candidates knew about hierarchical sampling schemes. So the Analysis of Variance table was wrong, degrees of freedom were wrong, and therefore (ii)(b) could not really be answered.

Question 6
In part (ii) candidates were asked to derive the two basic equations, namely to explain how they arise and not just to write them down from memory. Not many people could actually write down a design, given the values of the parameters – use all possible pairs for (iii)(a) and omit one from each block in (iii)(c). These two types are very standard designs. And in the final part a practical comment could be made on the choice of design, mentioning the sensitivity of a taster's palate after several samples have been tasted.

Question 7
The defining contrast is I = ABCE = BCDF = ADEF, the third being the interaction of the first two independent interactions specified. There are three usable sets for residual. In (vi) we could argue that so many F-values less than 1 suggests that the estimate of error is wrong. Removing the large item does make a difference.

Question 8
Several candidates did follow the hint at the end of the first paragraph to highlight the difference in means for M. In the analysis of variance table, sites has 1 d.f. and blocks within sites has 4 d.f., its sum of squares being calculated as the difference between SS blocks (5 d.f.) and SS sites. Adding a treatments by sites interaction uses up another 4 d.f. Comments could include the removal of X from any future experiment (as well as having trained all sites).

Module 5 (Topics in applied statistics)

Question 5
As usual for this topic, most of the candidates who attempted it gave good answers.

Question 6
Although a fairly standard topic, this was not as popular as usual. Again, most of the answers were good.

Question 7
This question attracted only one good answer. It is standard bookwork and similar questions previously have usually produced better answers than this. There is the temptation, when asked to prove a result which is given, that an element of fudging will appear.

Question 8
Cluster sampling does not require such a detailed sampling frame as simple random sampling or stratification, which is one good reason for using it. Of course there has to be a sensible way of choosing clusters, but often there is. Some candidates did explain the possible difficulties with systematic sampling, but a surprising number seemed unaware of them. No one seemed to know that it can be analysed by a method based on cluster sampling. Comments on whether sampling was worth doing at all were rather weak. A look at the pattern of hourly landings would have strongly suggested the answer No; few people seemed to have spotted that.