ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

ROYAL STATISTICAL SOCIETY EXAMINATIONS, 2014

REPORTS OF EXAMINERS

GENERAL COMMENTS

This report incorporates the comments made by examiners after marking the papers set in 2014 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 2015 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 also 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 2014 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 2014 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 (albeit 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. This point is especially important in the two Ordinary Certificate modules, where there are disparate marks for each question.

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)

Question 1
The systematic nature of the sampling was not always recognised.

Question 2
Cluster sampling had obviously not been taught in some places, but most candidates seemed to have learnt something about questionnaires.

Question 3
The scenario was not understood at all by some candidates, who could not relate their answers to the normal periodic (annual) servicing. The idea that customers might want to wait until they at least had driven the car home before replying was not often picked up. Many of the personal details of owners would either be known to the garage already or would be totally redundant.

Question 4
Answers were reasonable, except when part (iii) was ignored.

Question 5
There was a wide range of quality of answers, some were very good but others only mentioned the legal restrictions that may exist.

Question 6
This was quite well answered on the whole.

Question 7
The doctor should not give away what she is looking for, but patients should be asked to keep a diary. A (very) few candidates spotted this.

Question 8
Surprisingly, this was not well answered; (b) is a simple situation and in (c) candidates seemed to be more concerned about votes being miscounted than whether the attendance was likely to be representative of the society.

Question 9
There was poor reading of the question; email addresses and easy access to computers imply that everyone can be contacted at work. The problem is to motivate them to items which Management feels might interest them.

Module 2 (Analysis and Presentation of Data)

Many candidates displayed a good grasp of the concepts and techniques required in this paper. There appeared to be fewer unprepared candidates than in recent years. A fairly common problem, however, was a failure to answer exactly the question asked, and this did result in the loss of marks.

Poor presentation continues to be an issue. Some candidates have very poor handwriting and they are likely to lose marks even with the most patient and persistent examiner. More commonly, calculations are scattered about the page in a way that makes them difficult to follow, and diagrams are, in many cases, poorly executed.

Question 1

Because this question dealt with proportions drawn from data, it was more popular than a probability question would have been, though of course it tests essentially the same concepts.

In part (ii), a comparison was required of the proportion of trains arriving on time and the proportion of trains arriving on time given that they departed late. Some candidates resorted to general descriptions in words, making a convincing argument much more difficult.

Question 2

This question was frequently well answered with the majority of candidates able to explain clearly the confusion in the news story. The argument required in the second part of the question was easy for many, but some (as in Question 1) were uncomfortable working with figures rather than words.

Question 3

The interpretations of the data in this question were often quite poor. In part (i), the point is that the lowest 10% of households have net debts of £9164 *or more*. Those two vital words were often omitted.

Then in part (ii), most candidates attempted to draw a graph taken directly from the data in the table, but this is not what the question is asking for. The required graph is a distribution curve for net household wealth: this will be a positively skewed distribution, and consequently the mean will be greater than the median.

In part (iv), we can be sure that there will be outliers at the upper end of the distribution. It isn't possible to say with absolute certainty whether or not there will be outliers at the lower end, but it is very likely that there are.

Question 4

Most candidates struggled with this. Only a very few were able to calculate the roots required to find the percentage growth rates in populations. As a consequence, the message in the data -- that absolute growth rates are rising but relative growth rates are almost stable -- was lost.

Question 5

Candidates clearly understood the data, but in some cases they lost marks through not addressing the precise requirements of the question. It was slightly more common to give additional unwanted analysis than to omit what was required. It is worth noting how the question is worded, too: the instruction 'Describe briefly' should suggest to candidates that a sentence or two will suffice. A lot of time can be wasted through writing a paragraph instead.

Question 6

Though again the data were generally understood, the answers to this question did have some common errors.

In part (i), the point to notice was the 7-figure cyclical pattern in the data. (Comments to the effect that 28 data points couldn't correspond to a 24 hour day, or that 28 data points couldn't be for a month unless it was February and not a leap year, did not receive credit.)

In part (ii), it was quite common for candidates not to calculate the last three values in the series of moving averages. This didn't appear to be a problem of not knowing the method; perhaps it was a case of not reading the question carefully.

In part (ii) a common error was to plot the moving averages incorrectly aligned with the raw data series. Each moving average should be plotted at the centre of the 7 data points on which it is based.

There were two common errors in part (iv). Many candidates did not talk about trend and variation in their statistical sense, using these terms instead in their everyday sense. So, for example, a candidate might comment that 'the trend is for there to be variation within the week'. The second error was a failure to relate the pattern in the data back to the original situation; that is, candidates were required to say something about the trend and variation in relation to hits on the blogger's website.

Question 7

As with question 5, candidates had no difficulty understanding the data, but in many cases they did not address the exact requirements of the question. Part (ii) asks for a series of quite distinct comparisons and descriptions. The best answers were those which kept these comparisons and descriptions distinct. However, there many answers which gave a general narrative about the data, making the awarding of marks for specific points more difficult.

Question 8

This question was done well by many, though inevitably some became confused about precisely which *Price x Quantity* sums to calculate. The answers offered to part (iv) were suitably imaginative.

Question 9

The routine calculations in parts (i) and (ii) were usually done well, though in many cases there were several pages of working. Curiously, quite a few candidates simply omitted to calculate the mean and standard deviation for *y*. This may have been another instance of not reading the question carefully.

Part (iii) was found difficult by many. Some simply omitted it, some recalculated everything from scratch, some got confused about whether to multiply or divide by the conversion factors, but some did what was required with the intended minimum effort.

HIGHER CERTIFICATE IN STATISTICS

Module 1 (Data collection and interpretation)

Many candidates displayed a good grasp of relevant issues in collecting and interpreting data. However, many scripts were difficult to read, either because candidates' hand-writing is poor, or – more importantly – because the answer was poorly structured.

There was some tendency among the weaker candidates to ignore the specific requirements of the question in favour of writing as much as possible of a general nature, hoping thereby to pick up marks. Generally speaking, this is not a profitable approach.

Question 1

Though almost all candidates were able to pick out the main message in the data – namely that performance by the university's intake appears to be falling -- it was common to ignore much of the information given. So, for example, the data on the test scores in the whole population was not used to observe that the median score among the students has fallen from the 98th percentile nationally to the 96th percentile. It was also common to use the maximum and minimum scores for each cohort when these may well be unrepresentative; almost nobody realised that the 5th and 95th percentile scores would be more useful.

The critical comments offered in part (iii) were generally appropriate, though some of the remedies suggested (such as requiring all students in the cohort to take the test) were rather impractical.

Question 2

This question was done well by many. The calculations required were straightforward, and the comparisons based on them were broadly accurate.

The common weakness in answers, however, was the graphical representation. In part (i), a stacked bar chart for survivors and non-survivors in the four classes makes the message in the data plain. In part (ii), parallel bar charts for males and females in each of the four classes are the most appropriate.

In part (iii) it is important to present a conclusion supported by argument based on the data. Some candidates simply omitted to state their conclusion.

Question 3

Candidates handled the data well. The calculations required in part (i) were generally carried out accurately. Similarly, the standard errors in part (ii) were calculated correctly, though some were unable to explain clearly how a standard error should be used in interpreting data.

In part (iii) there was some evidence of candidates being on autopilot: having learned the concepts of stratified and cluster sampling, they gave textbook answers without any connection to the scenario in the question.

Part (iv) was generally answered well, showing a good understanding of what the data show.

Question 4

Most candidates were able to draw up a questionnaire without any difficulty. However, it was common for questions to be inappropriate in a survey of this type (e.g. date of birth when age band would be sufficient), or difficult to code (e.g. state you favourite radio station and say how often you listen to it).

Some requirements of the question were overlooked by the majority, such as the distinction between national and local news, sport, weather and traffic; e.g. use of the questionnaire to recruit more volunteers.

In part (ii) the modifications suggested were often few and perfunctory.

Part (ii), however, was generally well done.

Module 2 (Probability models)

The results were rather disappointing. A few candidates attempted only two questions.

Question 1

A few candidates answered this well but many did not. In considering the number of ways the various arrangements could arise it is best to consider first the possible values and then the number of positions those values could take. Some candidates thought a pair of equal values must be next to each other, this was incorrect.

Question 2

Some candidates received few marks because they could not evaluate the necessary integrals. Note also that the range of the integration is from θ to infinity. Candidates ought to know that a probability density function is non-negative. In (ii)(a) many candidates lost a mark by ignoring the question in the second sentence. The final part required equating the two expressions given and solving for *n*. Also remember *n* must be an integer.

Question 3

This question tested knowledge of the Poisson approximation to the binomial in (i) and (ii) and the Normal approximation to the binomial in (iii). It was the most popular question and the one answered best. Note that a continuity correction is necessary in (iii).

Question 4

In part (i) a number of candidates forgot that, for example, the variance of 4X is 16Var(X) and so their answers were incorrect. In part (ii)(a) the independence of X and Y is crucial. Part (ii)(b) required solving $F_W = 0.25$ and F_W=0.75 using the normal tables. In part (iii) the probability that W is in the interval is 0.5 and K has a binomial distribution. A normal approximation, with continuity correction, was expected.

Module 3 (Basic statistical methods)

In general there were some very good solutions this year, and it was good to see candidates reporting their conclusions clearly and concisely in relation to the context of the questions.

Question 1

This was attempted by the majority of candidates. In part (i) the variance calculation was generally well done although a few candidates did not use the correct denominator for an unbiased estimator. A small number tried to use $\sum f_i(x_i - \bar{x})^2$ rather than the calculation formula and sometimes the denominator of $n-1$ was forgotten. It was pleasing to see that the equality of mean and variance for the Poisson distribution was well known. In part (ii) the expected frequencies were generally well found although the pooling of the two final categories (so that the expected frequencies are all at least 5) was not generally well done and this affected the final results and conclusions.

Question 2

Part (i) was a χ^2 test of association (although it could also have been done as a test of difference in two proportions). In general this was well done and Yates' correction was applied in most solutions. However, the null hypothesis required care here – this should make it clear that the test is one of association. Many candidates wrote 'H_0: there is no difference between the groups of patients', but did not state that the difference being investigated is in the two proportions who prefer/do not prefer the new packaging. In part (ii) McNemar's Test was well applied (again remembering the correction factor), the reasons for using it seeming to be well understood.

Question 3

Part (i) initially required some accurate figure work, both in finding the differences and then ranking them. There were elementary arithmetic errors in some solutions and candidates should be aware of the need for accuracy. It was good to see that the correct Wilcoxon test was applied, only one candidate performing the Mann-Whitney test instead. Care was required in looking up the comparison value from tables, remembering to compare the lower total with the tabulated 2.5% value for a two-tailed 5% test. Fewer candidates attempted the Sign Test in part (ii). The conclusions from the two tests were different. This was both because the Wilcoxon test takes into account the magnitudes of the differences **and** because the differences in one direction were much smaller than those in the other direction, resulting in considerable imbalance between the sums of positive and negative ranks.

Question 4

Part (i) was bookwork. The two types of error were well known. The meaning of the significance level is that it is the probability of making a Type I error. This was all that was required, but many solutions went on to explain how levels are generally fixed at particular values which was not relevant. Part (ii) was well done when it was attempted, although many solutions stopped after part (i). The standard deviation is given in the question, so normal values rather than t values should be used. Marks were also given for saying what probabilities were being found: $P(\bar{X} < 34)$ in both cases. In part (ii)(c), the quantity being found was the power of the test *in the particular case* where $\mu = 32.5$.

Module 4 (Linear models)

The paper was generally done very well, with the vast majority of students attempting 3 questions. Here are some remarks about individual questions.

Question 1
Graphs were drawn well and most people noticed the increasing trend. Few commented on the increasing scatter, which is important for criticising the use of the pmcc here.
Several candidates used an approximate t-test for the correlation coefficients, which is not necessary as the critical values are given in the Official tables.
It is not recommended to use the notation S_{x^2} for S_{xx}.

Question 2
This question was answered well. The full details of the model did not always contain the error term. Use simpler forms of calculations for S_{xx}.
In 2(b) some candidates did not understand what the standard error of the estimate meant.

Question 3
This question was answered by fewest candidates. In 3(iv) some candidates did not remember to convert the prediction back into the correct scale. Calculation of the half life was not done well.

Question 4
This question was done very well. However, most candidates did not allow for the number of replications to vary in each treatment group in stating the one-way model.

Module 5 (Further probability and inference)

Question 1 was the most popular question and the best answered, while Question 2 was the least popular and generally poorly answered.

Question 1
Part (iii): in finding the marginal distribution of *Y*, many candidates used 0 as the starting point of the range to be integrated instead of *y*.
Part (iii): the range where the marginal densities are non-zero (in this case 0 to 1 for both) should be given as well as the formulae for the densities in this range.

Question 2
Part (i): most candidates did not seem familiar with this way of using moment generating functions.
Part (ii): a common mistake was to assume that *U* and *V* are independent, which is only true if $a_1 b_1 = -a_2 b_2$.
Part (iii): many candidates did not seem to know how to deal with multiple solutions.

Question 3
Part (iii): many candidates did not recognise that *Y* has a Binomial distribution.

Question 4
Part (a): a very basic result, but there were few good answers to this part.
Parts (b) (i) & (ii): integration by parts should be used to evaluate $E(Y)$ and $E(Y^2)$; it saves time to use the fact that the integral from 0 to ∞ of the probability density function $f(y)$ is 1.

Module 6 (Further applications of statistics)

The overall performance level was low, in particular on the discursive parts of the questions.

The best performance was usually on the pure number crunching questions, stepwise regression calculations, constructing an ANOVA table and to a lesser extent calculating rejection probabilities in quality control.

The worst performance was usually on the interpretation of what the calculations meant, comparing and contrasting of methods and experimental design.

This is common in many statistics exams, of course, but seemed even more extreme than usual here. To some extent this is a natural consequence of the fact that students can practice doing the calculations themselves, but they can't practice designing experiments themselves, so questions like that always end up involving memorising textbooks rather than learning by doing.

Module 7 (Time series and index numbers)

Question 1
Most candidates had a good theoretical understanding of the components of a time series, and how they combined to form additive and multiplicative series. Most candidates gave a practical explanation of the components of a time series. A better understanding of the impact of Easter, and how to adjust for it, would have helped some earn better marks.

Question 2
Nearly all candidates had a good theoretical understanding of simple exponential smoothing. Higher scores would have been obtained by a better knowledge of how to calculate moving average weights. Even so, candidates generally demonstrated the ability to apply the weights.

Question 3
The key to this question was noticing that the prices can be calculated by dividing values by quantities. Candidates who used this tended to score well.

Question 4

The key to this question was remembering the Lapseyres and Paasche formulae in terms of prices and quantities, and being able to substitute periods as described in the question. Candidates who then followed the instructions scored well.

Module 8 (Survey sampling and estimation)

This paper is designed to assess a wide range of topics associated with sample surveys and estimation. It is essential not only to understand the theory and be able to apply it to data but also be familiar with real-life surveys and the practical difficulties that might arise.

The overall standard of this paper in 2014 was good, with most candidates scoring marks in the range 30 to 49. All questions were attempted and done well. The most popular questions were questions 1, 2 and 4.

Candidates are reminded that they are expected to memorise certain formulae as indicated in the syllabus, and know how to apply these formulae to data.

Question 1

Most candidates were able to obtain point estimates and 95% confidence intervals for the population total in part (i) for stratified random sampling and part (ii) for simple random sampling. Some marks were lost in part (i) due to the variance calculations; although the formula was given, and candidates could define its terms, not everyone knew how to perform the computations. A common mistake in part (ii) (and question 3, part (ii)(b)) was to use the critical value of the Normal distribution in constructing confidence intervals, which is inappropriate with a sample size of 10 (i.e., the *t*-distribution should be used). A few candidates misread the question and gave estimates for the population mean, losing marks unnecessarily. Part (iii) was done well, similar numbers of candidates suggesting use of proportional or optimal allocation for a future survey. Either method would have worked.

Question 2

The concept of bias in part (i), and how it applied to this survey, was well understood by most candidates, i.e., “50 out of 103456 residents, self-selected, cannot be taken as a good representation of residents in the district”. Candidates’ answers to part (ii) were mixed. Some candidates misunderstood the question and instead described how they would select a random sample by each method in (a), (b) and (c). There were some good answers on the pros and cons of the 3 methods, but sometimes candidates did not give enough pertinent points to obtain full marks. Candidates are reminded that if there are 12 marks, there needs to be ~12 points to score 12/12.

Question 3

Few candidates attempted this question based on simple random sampling, and ratio estimators, but those who did, scored high marks. There was some confusion in part (i) over the standard deviation and the standard error (the latter was required). Nearly everyone gave the standard deviation of the measurements (and compared this quantity to the standard error in part (ii)(b)). Most candidates gained full marks in part (ii)(a) for the scatter plot, as well as noting the strong positive correlation, in commenting on the suitability of a ratio estimator, which was pleasing. Part (iii) was descriptive; some candidates wrote more generally, whereas the question required a more considered approach (i.e., how it applied to this survey).

Question 4

There were some good marks to all parts of question 4. Parts (a) and (c) were descriptive; there were some good answers, but sometimes candidates did not give enough pertinent points to obtain full marks. Candidates' discussion of the practical issues in "combining data from rural and urban areas" in part (a) was somewhat disappointing. Many candidates thought family income would differ between the rural and urban areas, but only a few went on to suggest use of stratification (or clustering) in combining information from the two areas. Part (b) on estimation of proportions, and determination of sample size, was done well, with most candidates scoring full marks.

GRADUATE DIPLOMA IN STATISTICS

Module 1 (Probability distributions)

In general, this paper was well done with some candidates achieving very high marks for it. In every question, the average mark was greater than 10/20. Few candidates appeared to struggle to find 5 questions to attempt.

Question 1
This question examined the Law of Total Probability and applied some results derived from it to sums of independent Normal random variables. About 80% of candidates attempted it, getting over half marks on average. In part (i), several candidates basically assumed the result to be proved about $E(X)$ in order to derive the result; this is a fundamental error in mathematics and gained no marks. For later parts of the question note that, if X, X_1, …, X_6 are independent $N(\mu,\sigma^2)$ random variables, then $6X \sim N(6\mu, 36\sigma^2)$ but $X_1 + \ldots + X_6 \sim N(6\mu, 6\sigma^2)$.

Question 2
This question required candidates to derive some important properties of a continuous random variable X, with a symmetric probability density function, and $Y = X^2$. About 80% of candidates attempted it and they scored just over half marks on average. In part (iii), a lot of candidates failed to realise that $E(XY) = E(X^3) = 0$ by the symmetry of the distribution of X. Also, a surprising number of candidates thought that, if the covariance between two random variables is zero then they must be independent. This is not true in general, though it is true that independent random variables are guaranteed to have zero correlation. For part (iv), it is important to realise that the direct method for finding the probability of $Y = g(X)$ using dy/dx is only valid for a strictly increasing or strictly decreasing function $g(.)$. The function $y = x^2$ is not strictly increasing or strictly decreasing when x can take both negative and positive values.

Question 3
This question concerned the joint probability density function of two continuous random variables, from which various parameter values had to be derived. Almost every candidate attempted it, and marks were generally high. In part (i), many candidates failed to write down the range spaces of the random variables and so failed to demonstrate fully that the distribution was exponential or uniform. Several candidates seemed to think that the joint distribution factorised so the random variables were independent, but this could not have been the case since their joint range space was not a rectangular region (or Cartesian product of sets).

Question 4
This question primarily examined the topic of simulation, though part (a) also required candidates to construct the joint probability distribution of two binary random variables from partial information. A minority of candidates attempted this question though those who did generally scored high marks.

Question 5
This question was based around properties of the multivariate normal distribution. Only a small number of candidates tried it; about half of them scored very high

marks, the others generally very low marks. Those who struggled with the question were clearly much less comfortable with part (b) than part (a).

Question 6
This question examined moment-generating functions and the Central Limit Theorem, in the context of gamma and (as a special case) exponential random variables. Almost all candidates attempted this question and they scored over half marks on average. In part (i), very few candidates made any effort to explain why the moment-generating function was only defined for $t < \theta$ (which was required for the relevant integral to converge). In part (ii), candidates were expected to deduce the moment-generating function of the exponential distribution from that of the gamma distribution, by setting $\alpha = 1$, but many chose to derive it from scratch, wasting a great deal of time in the process. In this part of the question, few candidates pointed out that the moment-generating function of the sum of the random variables was the product of the individual moment-generating functions only because the random variables were independent and few referred to the uniqueness property of moment-generating functions when identifying the sum as a gamma random variable. In part (iii), a common error was to try to use the Central Limit Theorem with gamma random variables (mean = μ/θ or n/θ) rather than exponential random variables (mean = $1/\theta$).

Question 7
This question required candidates to use general properties of sums of non-independent random variables to obtain results required when sampling without replacement. Fewer than half of all candidates attempted it, but they scored over half marks on average. One common error in part (ii) was to assume that the random variables were, in fact, independent and ignore the covariance between them.

Question 8
The final question examined the topic of bivariate transformations. About half the candidates attempted it, and their marks were almost all very high. In part (ii), several candidates would have benefited from realising that the joint probability density function of *R* and *Q* factorised on a rectangular range space, so the random variables were independent with marginal probability density functions that could be written down as factors of the joint probability density function (as in the Factorisation Theorem).

Module 2 (Statistical inference)

Questions were generally well-answered with the exception of Questions 6 and 8 which were not attempted by many candidates.

Question 1
Part (i): the words “Write down” and the small number of marks for this part should warn candidates that detailed calculations are not required.
Part (ii): in evaluating the Fisher information it is nearly always easier to find the expected value of the negative of the second derivative of the log likelihood rather than the expected value of its square.

Question 2
Part (i): unless told otherwise, when using calculus to find a maximum likelihood estimator, candidates should always check that the turning value of the log likelihood corresponds to a maximum.
Part (iii): it is important to follow the hint and not try using calculus for this part.

Question 3
Part (ii): A common mistake was to start with $P\left(F_1 \le \frac{\sigma_1^2}{\sigma_2^2} \le F_2\right) = 0.9$ (where F_1, F_2 are the 5% and 95% points of F_{n_1,n_2}) rather than with $P\left(F_1 \le U\frac{\sigma_2^2}{\sigma_1^2} \le F_2\right) = 0.9$.
Part (iii): another common mistake when finding the 5% point of F_{n_1,n_2} was to use the reciprocal of F_{n_1,n_2} rather than of F_{n_2,n_1}.

Question 4
Part (iv): when evaluating $E(\sum(X_i - 1)^2)$ and $var(\sum(X_i - 1)^2)$ it is beneficial to start with $E(X_i - 1)^2$ and $var(X_i - 1)^2$.

Question 5
Part (i): it makes the subsequent working much easier to notice immediately that the likelihood can be simplified to $constant \times p^{Y_1+3Y_2} \times (1-p)^{3Y_0+Y_1}$.
Parts (ii), (iii), (iv): a common mistake was just to describe the methods and not go on to make use of them for this particular example.

Question 6
Parts (a) and (b): for this type of "essay" question candidates should (i) address the question asked, and (ii) bear in mind the number of marks allocated. Long, rambling answers only peripherally relevant to the question asked do not attract many marks and waste time. As a rule of thumb, for candidates with around the average handwriting size the length of the answer should be approximately 2 lines per mark allocated.
Part (c): this was not well answered.

Question 7
Some of the definitions of sufficiency were rather vague; in particular, the Factorisation Theorem should not be used as a definition.

Module 3 (Stochastic processes and time series)

This year the paper had 5 questions on Stochastic Processes and 3 on Time Series. There were many excellent scripts which were a pleasure to mark and the majority of candidates were able to demonstrate their ability and knowledge to good effect.

Question 1
Although this was a fairly popular question on Branching Processes, a lot of the candidates attempted to prove $G_{n+1}(s) = G(G_n(s))$ thereby addressing the wrong question. The initial bookwork in part (i), requiring a random sums argument with pgfs, was poorly done as were parts (iii) – (v) in spite of the follow-through from the formula given in part (ii).

Question 2
This was a very accessible question which tested candidates' understanding of Markov Chains and stationary distributions at a basic level. It proved to be both the most popular question and the best answered. However part (iv), the single piece of theory on irreducibility and transience, was not well-done.

Question 3
Most candidates attempted this question on the M/M/1 queue and the marks covered almost the entire range. For many candidates the jump from the (given) equilibrium distribution of the number in the system to the waiting time distribution, was a step too far. Others were stumped by the optimal cost application in part (vi).

Question 4
This was the most mathematical question on the paper, involving differential equations and generating functions. There was an encouraging number of attempts, some of them excellent.

Question 5
The fifth stochastic processes problem was on M/G/1 queues and the application of the Pollaczek-Khintchine formula to automation. Perhaps due to lack of familiarity, it attracted few attempts.

Question 6
The most popular time series question on ARIMA models and the acf produced many good attempts. However there was also much confusion about whether stationarity and invertibility were related, and some naivety about what behaviour would be expected in a plot of a series from a given ARIMA model.

Question 7
Although less popular than I expected, this question on the Holt-Winters procedure allowed many candidates to demonstrate their understanding of the algorithm. The relevance of MAPE to the multiplicative form was a mystery to all bar one.

Question 8
This was the least popular time series question, perhaps because it was the last. The use of the Yule-Walker equations to estimate the pacf was not well grasped.

Module 4 (Modelling experimental data)

Good marks were generally achieved where candidates addressed all parts of a question. In design-focussed questions it is generally important to be able to construct the sets of treatments (in blocks) to be considered, and important to check that, when constructed, these sets do have the required properties. For questions focussed on the output from analyses, the key element to the question is usually the interpretation of the presented (or calculated) results, and this interpretation needs to go beyond a simple description of what has been presented, ideally providing inferences and conclusions in a non-technical language.

Question 1

This question covered possibly the kind of interpretation that should be standard work for the applied (consultant) statistician. Most answers provided a clear interpretation of the ANOVA table, but many then failed to describe the observed effects indicated for the significant terms. There were some good descriptions of the assumptions underlying this analysis, though a number of answers were quite confused about the assumption of independence. Generally poor descriptions of the construction of the residual plots used to assess these assumptions, possibly considering this construction to be trivial as provided by computer packages, and many answers then failed to indicate the information that could be gleaned from each of these plots. A few answers identified the binomial form of the data, with more suggesting the use of a generalized linear model as an alternative analysis approach. Some sensible suggestions of applying transformations prior to analysis, but often with an inappropriate choice of transformation without clear thought about the form of the data.

Question 2

This is a common type of design problem, requiring the statistician to balance various different constraints in constructing a suitable design, but was not attempted by many candidates. Those that did were generally unclear on the distinction between confounded and fractional factorial designs, and none identified how to combine the two ideas. Most answers failed to consider the construction in the two steps that are needed – first to select a suitable fraction of the full factorial treatment set, and then to divide this fraction into two groups based on an appropriate confounding contrast. The outline ANOVA tables tended to forget that only 16 observations were included in the first part of the question, therefore indicating far too many degrees of freedom both for estimating effects and for the residual. There were no approaches presented for the last part of the question, considering extensions to use all 48 available observations.

Question 3

Answers generally provided good descriptions of a BIBD though some explanations of the relationships between the parameters were unclear about the basis for each equality, and many answers failed to identify that parameters must all be integer values (most importantly for λ). Many answers failed to show that the solution to part (ii) was the minimum size of experiment – just showing that the relationships hold for the presented parameter values is not enough – and some failed to calculate the overall size of the experiment for part (iii). There were some good solutions for the treatment allocation, though a number failed to correctly permute pairs of treatments, and then clearly failed to check the pattern of within-block joint occurrences. Some of the suggested randomisation processes were overly complex, unnecessarily keeping sets of blocks together in consecutive days, and most failed to consider the random allocation of codes to treatments. There were a few good and very complete answers.

Question 4

This was quite a popular question with some good answers. Calculations of sums of squares were generally good, though some answers failed to include the block term, and some made careless errors in subtracting the block sum of squares from the overall treatment sum of squares when obtaining the interaction sum of squares.

Calculation of degrees of freedom, mean squares and variance ratios were generally good, and identification of significant terms was good. Many answers were not entirely clear on what is meant by a contrast, or on the value of using orthogonal contrasts, though almost all answers identified the conditions for orthogonality of two contrasts. Some contrast definitions were good, though often failed to identify contrasts associated with the interaction between concentration and chemical. Many sets of contrast coefficients failed to properly take account of the standard treatment, and many answers suggested non-orthogonal contrasts. The values of the contrast coefficients were generally not clearly presented, though calculations of contrast sums of squares were usually correctly calculated and interpreted.

Question 5
Not a very popular question. Many answers were quite vague on the components of a GLM, and, in particular, about the components of a log-linear model for analysis of data in a contingency table – no answers mentioned that the data were from a multinomial distribution. Many answers failed to identify the saturated model, often identifying the most complex model included in the list of models. The description of what the baseline model represented was generally poor, though there was usually a better appreciation of what the more complex models represented. There were some good descriptions of the backward elimination approach to identify the parsimonious model, though most answers failed to compare the saturated model to the next most complex, and a couple of answers tried to use a forward selection approach (which tends to result in too complex a model) or an incorrect backward elimination approach (based on removing significant terms rather than non-significant terms). The interpretation of the best model was generally missing from the submitted answers, certainly in terms that a non-statistician would understand.

Question 6
This was a fairly popular question, generally with good answers to the first part about the Gauss-Markov theorem, though some answers failed to note that the least squares estimator was unbiased. Some answers presented useful theory about the impact of complete multicollinearity, but often there appeared to be limited understanding of the practical impact of partial multicollinearity. There were some clear descriptions of the impact of changing the threshold, though this part was missing in a number of answers. The explanation of Step 5 was often based on the calculation of an incorrect test statistic (the ratio of the residual mean squares for the best two options), so reached the correct conclusions for the wrong reason. Many answers just provided a description of the presented stepwise process rather than an interpretation of the fitted model parameters, and most answers needed considerably more thought about the different variables (and particularly the construction of shape and surface area). Answers were generally good on assumptions, though there were some erroneous suggestions that all explanatory variables needed to be uncorrelated (which was clearly not the case here!).

Question 7
There were relatively few answers for this question, and most were rather vague on the differences between linear and non-linear models – clearer identification of examples would probably have helped. Most answers were clear on the principle behind Newton-Raphson, though usually expressed algebraically for a single parameter. There was some confusion about the interpretation of the different

parameters, and almost all answers failed to note that it would be reasonable to set parameter *a* to zero. Most answers understood the principle of separating the residual into pure error and lack-of-fit, but needed more explanation about how the pure error sum of squares could be obtained from the data. Most answers tried a forward selection rather than backward elimination approach to exploring the combined model, producing an unconvincing model with parameter *a* allowed to vary. Most answers lacked any real interpretation of the best combined model in terms of the original parameter values.

Question 8
This was a relatively popular question, with many answers providing good explanations for the use of a transformation, but few being particularly clear on the use of a weighted analysis. Most answers provided a good description of the observed response, noting both the change in variance and the possible curved nature of the response, but many did not consider the benefit of applying a log transformation to produce a more generalizable model. There were a few sensible comments about the need to obtain more data, particularly for older drivers, to avoid extrapolation. Some good exploration of different aspects of the modelling to identify the best model to use, though some conclusions appeared to be drawn based on just one aspect (e.g. percentage variance accounted for or Normal plot), and some incorrectly being concerned about the relative sizes of the standard errors.

Module 5 (Topics in applied statistics)

In general, this paper was reasonably well done, though a few candidates were unable to answer five questions. The paper always consists of two questions on each of the four main topics from the syllabus, so future candidates might be more successful if they aim to learn three topics in depth rather than attempt to cover all four topics in less detail. Many candidates seemed more comfortable carrying out mathematical or arithmetical tasks than engaging with a practical context or expressing their ideas in written form. Candidates in future years are strongly advised to practise writing short notes on important topics from the course and writing out brief conclusions from analyses they carry out during their preparation for the exam.

Question 1
This question was about principal component analysis, based on results from a study of brain activity in sufferers from Parkinson’s disease. Every candidate attempted it and they achieved just over half marks on average. Many candidates answered part (iii) with no mention of what a scree plot displays or description of how eigenvalues relate to variance explained. For part (iv), it is important to note from Table 2 that all the loadings have very similar magnitudes; for example, the small difference between 0.406 and 0.412 is unlikely to affect how we interpret one of the loadings for PC1. Several candidates appeared to have the mistaken impression that negative loadings should be ignored.

Question 2
This question examined discriminant analysis and cluster analysis. The majority of candidates attempted it but their marks were rather poor in general. The main

reason for this was not usually that their answers were wrong but that they made only some of the relevant points so attracted only partial credit.

Question 3
This question required candidates to calculate the Kaplan-Meier survivor function from a small data set, obtain a confidence interval for survival at one time point, and interpret the log rank test comparing survival in two treatment conditions. About three-quarters of the candidates answered this question and many of their attempts were very good. Several candidates failed to produce the correct Kaplan-Meier survivor function in part (i) because they handled censored data incorrectly; the patient with the censored survival time of 10^*, for example, should have been counted as still 'at risk' at 10 weeks. In part (ii), Greenwood's formula was often stated incorrectly; elements of $\mathrm{var}(\hat{S}(t))$ were often confused with the corresponding elements of $\mathrm{se}(\hat{S}(t))$.

Question 4
This question was also about survival analysis, in this case fitting a Cox proportional hazards model to a dataset with a mixture of binary and continuous explanatory variables. As for Question 3, about three-quarters of the candidates answered this question and many of their attempts were very good. In part (i), few candidates explained that the 'baseline' hazard is the hazard for subjects on no methadone, who are treated in Clinic 1, with no prison record (or some similar combination of values of the explanatory variables). In part (ii), many candidates discussed the statistical significance of each of the parameters but did not mention the direction of the effect (for example, increased doses of methadone were generally associated with a lower hazard, i.e. a longer time on treatment). For part (iv), note that $\mathrm{se}(20b_1) = 20\,\mathrm{se}(b_1)$.

Question 5
Question 5 examined odds ratios and the Mantel-Haenszel procedure. About half the candidates answered this question, many of them achieving very high marks for it. In part (i), some candidates forgot that the usual confidence interval is for the log of the odds ratio (rather than the odds ratio itself).

Question 6
This question tested candidates' knowledge of direct and indirect standardisation, in the context of deaths from coronary heart disease. About half of the candidates attempted it and it had the highest average mark of any question on the paper. Most candidates were able to explain the key terms clearly and standardise the death rates correctly.

Question 7
This question was about simple and stratified random sampling. Very few candidates attempted it and they obtained very few marks for their attempts. The theoretical result derived in part (i) should have helped candidates answer part (ii).

Question 8
The last question concentrated on cluster sampling. About half the candidates answered it but very few of them got good marks. In part (a), a few candidates appeared to confuse cluster sampling with cluster analysis. There also appeared to

be widespread confusion between clusters and strata. In part (b), many of the candidates struggled to make use of the hint and plug the correct sample values into the formula provided.