STATISTICS

Paper 4040/01

Paper 1

General comments

The overall standard of work submitted was very similar to last year, except for the fact that there were fewer exceptionally high marks, and considerably fewer very low marks.

Fewer candidates than previously, lost marks through not giving their final answers to the levels of accuracy required by questions, when this was stated. It continues to be the case, however, that in parts of questions requiring comments, many candidates produce general comments that have obviously been learned by rote from a textbook or similar source, when what is clearly asked for is a specific comment in the context of the question. Almost all candidates lost 2 or 3 marks for this reason in **Question 6**. One possible reason for these 'errors' is that candidates are not reading questions sufficiently carefully, and Centres should be advising them of the need to ensure they answer exactly what a question asks.

There is a further point of which Centres would do well to advise candidates to assist them in answering questions in this examination. Examination questions are prepared with great care, and all information given in them is included for a particular purpose. There are two ways in which candidates should use this fact to their advantage. Firstly candidates should ask themselves what information is given in a question which is relevant to being able to give an answer; **parts (iii)** and **(vi)** of **Question 9** provide excellent examples of this. Secondly, particularly in relation to later parts of a question, candidates should ascertain which, if any, parts of the information provided in a question have not yet been used.

Comments on specific questions

Section A

Question 1

For only a very small number of candidates did answers to this question appear to be the result of complete guesswork. Among those who clearly had quite a good understanding of the different methods of sampling, the most common error was confusion between quota and stratified sampling.

Question 2

Many candidates scored full marks on this question, and others would have done so too had they heeded the instruction in the question to give answers to two decimal places. The most common method error was, not surprisingly, failure to use the squares of the radii in **(ii)**.

Answers: **(i)** $1448.12 ($1448.13 also allowed) **(ii)** $5213.25 **(iii)** $1839.12

Question 3

(i) and **(iii)** were almost always answered correctly. However, in **(ii)**, an error of omission, identical to that in similar questions on both the previous two years' papers, was a frequent cause of loss of the mark for this part. A comment on what x represented needed to refer not only to the ownership of a motorbike and bicycle, but also to the fact that a car was not owned. Very few candidates interpreted correctly what was required in **(iv)**. The candidate's value of '21+y' was required as a percentage of '63+y'.

Answers: **(i)** 3 **(iii)** 12 **(iv)** 44%

Question 4

Given that a similar question had appeared in last year's paper, and specific reference was made in the report on that paper to the need for the cumulative frequency graph of a set of discrete data to be in the form of a step polygon, it was disappointing to find again that fewer than five per cent of candidates presented such a graph. As with the question last year it was, however, still possible, even if a step polygon was not drawn, to score four of the available six marks by calculating the cumulative frequencies correctly, using the stated scales, and starting a cumulative frequency graph of any sort from the correct point, in this case (0,9).

Answers: **(i)** 9, 20, 30, 36, 56, 61, 64

Question 5

For many candidates this question was clearly a source of six very easily-obtained marks, as they were able, in just a few lines of working, to obtain the correct equation in the required form, and then use it to obtain the requested estimate. Others, however, appeared to have very little idea of what to do. This is a topic which in previous years has appeared almost exclusively as part of a ***Section B*** question, and therefore optional. Centres and candidates should note that it is possible for any part of the syllabus to be examined in either of the two sections of either paper.

Answers: **(i)** $y = 0.68x + 14.4$ **(ii)** 36.9 (36.8 and 36.88 also allowed)

Question 6

A majority of candidates managed to score the first five marks available for this question, giving an adequate explanation in answer to **(i)** and then correct numerical answers to **(ii)**, **(iii)** and **(iv)**. One matter relating to **(iv)** is of concern, though. Some candidates lost the second mark available for this part because, having obtained a correct value for the mean, they not only then rounded it to an integer, but in some cases explicitly stated that this was necessary because the mean of data of this sort could only be an integer, which is totally incorrect.

A very small number of candidates correctly selected the median in **(v)**. However, apart from that, hardly any marks at all were scored by any candidates for **(v)** and **(vi)**. The reason for this was, quite simply, that candidates did not read the question sufficiently carefully. It specifically asked about the most appropriate measure of average 'to represent these data'. In other words, comments were required which referred specifically to the figures given in the question, but on almost all submitted scripts there were simply general 'textbook comments' such as 'the mode cannot be used in further statistical calculations', and 'the median is only one value from the data'. The reason for the mean not being the appropriate measure to use was that it was distorted by the one very extreme value, 16, in the data. The reason for the mode being inappropriate was that it happened to be the lowest value of those in the data.

Answers: **(ii)** 3 **(iii)** 1 **(iv)** 4.57

Section B

Question 7

It was pleasing to see the higher proportion of candidates than in previous years who knew that in a histogram, area is proportional to frequency; however, by no means all were able either to use this fact correctly, or to apply the scale on the vertical axis required by the question. Very few were able to obtain the correct result to **(i)(b)** despite the topic having been asked on a number of occasions in the past. Whether this was due to the presence of negative values of x is uncertain. However, many candidates coped successfully with the negative values when calculating the mean and standard deviation, although it was disappointing to see how few were able to complete the latter calculation totally correctly. As was the case last year, quite a number of candidates again included frequencies in their calculation of the mean, but not in that of the standard deviation.

Answers: **(i)(b)** 5 cm **(ii)(a)** 12.2 **(b)** 11.4

Question 8

This was not only by far the least popular question in ***Section B*** as regards the number of candidates attempting it, but was also well answered by only a very small number of those who did attempt it. Indeed, a lot of those who did attempt it only managed to score the three marks available in **(b)(i)**. Few candidates scored any marks in **(a)(i)**, comments generally just being in very general terms about outcomes being exclusive or independent. The comments being looked for were that the probabilities in the second column were those of a set of exclusive and exhaustive outcomes of one event (i.e. a visitor called to see one and only one of the inhabitants) which therefore of necessity summed to 1, whereas the probabilities in the third column were those of possible outcomes of different events (i.e. whether each of the inhabitants was at home) which therefore bore no relationship to each other. Some candidates did manage to answer **(a)(ii)** and **(a)(iii)** correctly, but correct answers to **(a)(iv)** were very rare indeed. Hardly any candidates at all realised that the probabilities of four different events needed to be multiplied together. Of those candidates who used valid methods to answer **(b)(ii)** and **(b)(iii)** many did not score as many marks as they might have done because they approximated the values they obtained mid-way through their solutions to integers, presumably believing this had to be done as they were dealing with 'people'.

Answers: **(a)(ii)** 0.09 **(iii)** 0.31 **(iv)** 0.0896 **(b)(ii)** 0.373 **(iii)** 0.818

Question 9

It is impossible to avoid the conclusion that many candidates learn the methods associated with the calculation of death rates by rote, without having much understanding of what they are doing or why they are doing it. This was borne out again by the marks obtained for the different parts of this question by many candidates. On many scripts, most or all of the ten marks available for **(i)** and **(ii)** (totally calculation) were obtained, but then none or hardly any of the remaining six which required comment or explanation. **(iii)** is an excellent example of the need to look at what information is given in the question which might help in answering it. The only information given about Lowshire was its population, yet reference to it was very rarely seen. The question asked why the crude and standardised death rates were identical. The crude death rate is based on an area's population, the standard death rate on the standard population, but hardly any candidates at all used this knowledge to spot that the area's population had the same age structure as the standard population. In **(iv)** only a very small number appeared to know that the standardised death rate should be used because, being based on the standard population, it enables 'like to be compared with like'. The decision asked for in **(v)** was one that has been requested very frequently in the past and tended to be better answered than the other later parts of this question. **(vi)(b)** provided another example of most candidates not using information given in the question which would enable them to evaluate their answer. The reference to 'a representative cross-section' meant that while the crude death rate would decrease as there would be fewer old people in Highshire, the standard population, and hence the standardised death rate, would not be altered.

Answers: **(i)** X = 7, Y = 5000, Z = 385 **(ii)(a)** 21.44 per thousand **(b)** 23.56 per thousand

Question 10

Most candidates knew how to obtain values of the medians and quartiles from the graph, and hence to obtain the inter-quartile ranges and then the difference between them. Appropriate tolerances were allowed for the accuracy of graph reading. For the later parts of the question, however, far fewer were able to interpret correctly what the question was asking, and then to use the graph to obtain their answers. In **(iii)** many candidates realised that what was required was the time for which the vertical distance between the two curves was at a maximum, but then gave a value around 30 minutes, rather than the correct one of approximately 10 minutes. A far smaller number realised that **(iv)** required the numbers of patients above each doctor's graph at 40 minutes to be summed. In questions such as **(v)** where both a decision and a reason for it are required, the decision, even if correct, cannot score a mark if either no reason, or an incorrect reason, is given. A majority of candidates failed to interpret the graph correctly in answering **(v)**. Also the relevance of the information in the question "assuming the week is a typical week" appeared to be totally ignored by many. The cumulative frequency of 33 for a time of 12 minutes for Dr Jones meant that 33 of Dr Jones' patients had to wait for less than 12 minutes before being seen, whereas only 20 of Dr Smith's patients had waited for that period of time before being seen. A patient seeing Dr Jones was therefore more likely to have to wait for less than 12 minutes than one seeing Dr Smith. In **(v)(b)**, by a similar argument, a patient wishing to minimise the chance of having to wait for 33 minutes or more should choose to see Dr Smith.

Answers: **(i)** Smith 15-16 minutes (exclusive), Jones 12.5-13 minutes **(ii)** Smith 23, 11-12(exclusive); Jones 34, 7-8(exclusive); IQR difference = 16 minutes **(iii)** Any value in the range 9-12 minutes **(iv)** 13

Question 11

While many candidates gave the correct reason required to answer **(i)**, that the *x*-values of the points were not in ascending order, the number giving one particular incorrect reason was sufficient to cause concern. This latter reason was that the mean had to be calculated first so that those values below the mean could be used to calculate one semi-average and those above, the other. <u>This is not correct</u>. In **(ii)** a small number of candidates gave the clear indication that they believed the question to be incorrect, and that the first four points given in the table had to be used to obtain one semi-average. The quality of the graphs presented was generally good; probably the most common error in well- drawn graphs was for the axes not to have been located correctly, leading to at least one plotted point being off the graph grid. Only a surprisingly low number of candidates managed to identify the 'rogue point' correctly, despite it being the only point well away from the general 'straight line shape' of the rest of the points. Hardly any candidates were able to give both a correct decision, and a correct reason for reaching it, in **(viii)**. The inclusion of the 'rogue point' in the calculations for the line drawn in **(v)** meant that the lower end of that line was higher than that of the best straight line drawn through all the other points in **(vii)**, and so the estimate obtained from the first line would be higher than the true value.

Answers: **(ii)** Either ADFH or BCEG **(iii)** (19,100), (12,72), (26,128) **(vi)** F (11.0,96)

STATISTICS

Paper 4040/02

Paper 2

General comments

The overall standard of work submitted was quite similar to last year, although there were noticeably fewer instances of very low marks being achieved than has been the case in previous recent years. Extremely high marks were also relatively rare, parts of some *Section B* questions being found quite testing by all but the very best candidates. The *Section B* question on probability was, however, answered more successfully than the corresponding question last year, although that on expectation proved just as unpopular as ever.

Two of the major sources of loss of marks are still occurring although, it is pleasing to note, with a reduced frequency. When a question instructs that an answer should be given to a particular level of accuracy, the final mark is lost if this is not done. And when what is specifically requested is a comment in the context of a question, general 'textbook' comments will not score any marks, (see the comments on **Question 9** for an example of this).

There is, however, a third general cause of candidates losing marks needlessly. This is through failure to read exactly what a question states, and to answer accordingly. Examples of this are given in the specific comments on **Questions 3** and **8**.

Finally, candidates must once again be commended for the excellent standard of most of the diagrammatic work presented.

Comments on specific questions

Section A

Question 1

As is usually the case with questions on this topic, many candidates answered clearly and correctly. For others, answers appeared to be pure guesswork, often having been changed a number of times. As has been mentioned in previous reports, the correct classification of a variable is essential if further analysis of it is to be valid.

Question 2

A disappointingly low proportion of candidates gave B as the correct statement, many giving as their reason for it being incorrect the mistaken belief that cyclical variation occurs over the same period of time in any series. It is seasonal variation which always occurs over the same period of time, cyclical variation does not. Candidates should note that when a question asks for a reason as to why a statement is incorrect, it is not sufficient just to state that it is not correct. For example, many simply gave as the 'reason' for statement *C* being incorrect, "The method does not remove the long term trend". To score the mark, it was necessary to give a further clarification/justification, e.g. "The method removes variation, not the trend", or "The method does not remove the trend, it reveals it", or other equivalent wording.

Answer: **(i)** *B*

Question 3

Many answers presented to the previous question and this one, both of which relate to the method of moving averages, gave the impression that the topic had merely been learned 'by rote', with little clear indication of any understanding of what was being done or why it was being done. This was highlighted on many scripts by the difference in quality between the answers presented for the first part and the other parts of this question. The vast majority of graphs presented in **(i)** were of very high quality indeed, but many such graphs were then followed by work which gave clear evidence either that the candidate did not know how the graph was to be 'used', or that no attempt had been made at all to 'use' it. An example of the general comment about noting the exact wording of a question is that if a question states 'use your graph', a candidate should immediately ask himself/herself "In what way is the graph to be used?" The purpose of plotting a time series graph is to see whether there appears to be any pattern in the figures. What the graph here showed was that the pattern repeated itself every five years, (a comment that there was a peak or a trough every five years was sufficient to show that the point had been realised and to earn the mark). A five-point moving average was therefore required. Answers given to **(iii)** frequently provided further evidence of a candidate's failure to understand what the method of moving averages is designed to achieve. A frequently-seen incorrect comment was that the required moving average was even-pointed because figures had been given for an even number of years.

Question 4

Questions on this topic are rarely well-answered by the majority of candidates, and this one was more poorly-answered than most. This was because not only was there the customary confusion between independent events and mutually exclusive events, but also because the method used in **(i)** by many candidates was, at that stage of the answer, not valid. From the information given in the question, the answer to **(i)** requires use of the general addition formula of probability, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. What many candidates used was the multiplication property of independent events, which is invalid here because, although it is eventually shown that A and B are independent, this has not yet been proved. When trying to show whether events are mutually exclusive or independent many candidates still give descriptive answers containing expressions such as 'they are independent because they have nothing to do with each other'. Descriptive answers never score any marks in this context. The only way to show that two events are mutually exclusive is to show that $P(A \cap B) = 0$, (or that $P(A \cup B) = P(A) + P(B)$); and, within the syllabus of this subject, the only way to show that two events are independent is to show that $P(A \cap B) = P(A) \times P(B)$.

Answers: **(i)** 0.2 **(ii)(a)** not mutually exclusive because $P(A \cap B) \neq 0$ **(b)** independent because $P(A \cap B) = P(A) \times P(B)$

Question 5

The vast majority of candidates scored all the four marks available for the calculation part of this question, but many scored poorly on the other parts. A common reason for this is almost certainly that insufficient thought appears to be given to exactly which words are being used. Price, price relative, expenditure, amount, cost and value are all terms which appear to be used interchangeably by many, even though each has a precise separate meaning. The figures in the table were clearly stated to be price relatives, and the question asked what certain of these figures indicated. Price relatives are exactly that, they give information about the relative size of prices, and so the fully-correct answer to **(i)** was that the price of food has risen by 4% between the years 2003 and 2004, and to **(ii)** that the price of 'other items' did not change between 2004 and 2005. Any reference to price relatives or expenditure, rather than price, in these answers meant that the marks were lost. The answer being looked for in **(iv)** was a comment indicating that the spending pattern of the national population was unlikely to be similar (and therefore to have different weightings) to that of candidates. It was not relevant to refer to the sample size, (that was not even given in the question).

Answer: **(iii)** 104.45

Question 6

One of the very pleasing aspects about scripts presented in answering this paper was the large number of candidates who scored the first two marks available for this question, because in previous years class limits have not been handled at all well. Many candidates referred explicitly to the values 9.5 and 14.5, showing that they understood the continuous nature of the data, even though it had been recorded to the nearest integer. However, quite a number then failed to heed the hint which had been given them by what was being asked in **(i)** when it came to using class limits to find the values required in **(iii)**. Having correctly given 9.5 as the true lower class limit of the 10 - 14 class in **(i)**, they then proceeded to use 10 as the lower limit in their calculations in **(iii)(a)**. Although many did give a correct solution to **(iii)(a)**, they then lost the final mark in that part for failing to give their final answer to 1 decimal place. In this context, it must always be remembered that 0 is just as much a digit as any other. In both parts of **(iii)**, answers based on various ways which textbooks teach of determining the required items were allowed.

Answers: **(ii)** 13, 32, 55, 84, 112, 130, 137 **(iii)(a)** 10.0 (from either 34.5th or 34.25th item)
(b) 30.6 (from 114th item) or 30.9 (from 114.54th item) or 30.5 (from 113.71th item, the exact answer of 30.45 also allowed in this case)

Section B

Question 7

Among the weakest candidates there was confusion almost from the start between houses and people. For many candidates, though, half-marks and a little above were very common. Only a minority interpreted correctly the information given in bold print in the stem before **(vi)**. This eliminated seven people from consideration, leaving 103 to be chosen. However, it was exceptional for an answer to this question to be awarded full marks. All but a very few candidates failed to interpret correctly in numerical terms the wording of **(v)**, 'they live in the same house with no other people'. This meant that once the first of the two people had been chosen from among the 16 who lived in 'two-person households', the second person could <u>not</u> be <u>any</u> one of the other 15, but only the one other person with whom the already-chosen person lived.

Answers: **(i)** 40 **(ii)** 110 **(iii)(a)** 0.125 **(b)** 0.675 **(iv)** 0.273 **(v)** 0.00133
(vi) 0.138

Question 8

Despite the detailed column headings given in the table in **(a)**, many candidates did not interpret the summations correctly and tried to make the calculations of the mean and standard deviation far more complicated and lengthy than they really were. Indeed, it was disappointing to see how few candidates were able to make a reasonably accurate attempt at calculating the standard deviation. Some candidates who did manage to evaluate the mean and standard deviation correctly then lost the final mark in each case through failing to give their result to the required level of accuracy.

Attempts at **(b)** were far more frequently correct, but here again answers were made far lengthier than they need have been. There were two reasons for this; one was that in a few cases candidates failed to realise that the values of p and q were actually given in the question. For more generally, however, the over-lengthy answers were the result of the method used. Most commonly this involved restating, for each unknown value, an equation involving the raw and scaled marks, means and standard deviations. It is much quicker to present an answer based on multiples of the standard deviations. Candidates should ask themselves, "How many standard deviations…….?" For example, to find the value of r, consider Xavier's raw score of 69. This is 2 marks above the raw mean of 67. As the raw standard deviation is 6, 69 is one-third of a standard deviation above the mean. The value of r, the corresponding scaled mark, will therefore be one-third of a scaled standard deviation above the scaled mean. So $r = 100$ plus one-third of $12 = 100 + 4 = 104$, etc.

Some candidates lost marks in **(ii)** and **(iii)** through failure to read the words in the question stem 'for each pupil', giving instead just the overall total marks for the papers.

Answers: **(a)(i)** 510 **(ii)** 14.6 **(iii)** 9266 **(iv)** 7.2 **(b)(i)** p = 100, q = 12, r = 104, s = 75, t =115, u = 47 **(ii)** 196, 193, 196 **(iii)** 313, 307, 316

Question 9

Answers to this question continued the improved trend of recent years shown by answers to questions on this topic. However, there are still a number of areas where there is room for general improvement. Most worryingly, some candidates are clearly still not learning the correct method of systematic sampling, that once the first item has been selected randomly, the others are taken at a fixed interval from the population listing. (It is frequently difficult to tell what candidates are doing, but in some instances it appears to be selection of numbers at fixed intervals from the table of random numbers. This is NOT the correct procedure.) In this case the first randomly-selected person was number 03, and because a 10% sample was required, the others selected were therefore 13 23 33 43 and 53. As is often the case, another common error was that in **(iv)** answers given were just general comments about the methods of sampling, whereas the question clearly referred to the three samples which had been chosen earlier. Very few candidates scored both marks available in **(v)** because knowledge of quota sampling seemed rather sketchy. The two points which needed to be made were that the quota of how many men and women were to be interviewed would be fixed in advance, but that it would be then up to the interviewer to choose exactly who these men and women were. Finally, in **(i)(b)** the question clearly indicated that the numbers which had not been used needed to be identified individually, and this was not always done.

Answers: **(i)(a)** 05, 53, 50, 12, 25, 37 **(b)** 61 and 82 out of range, 53 already selected **(ii)(a)** 01 **(b)** 10 **(c)** 03 **(d)** 13, 23, 33, 43, 53 **(iii)(a)** 2 men, 4 women **(b)** 03, 12, 47, 34, 51, 55

Question 10

Every year it seems, it can almost be guaranteed that, if there is a question in *Section B* on the topic of expectation, it will prove to be by far the least popular question, as regards the number of candidates attempting it, in that Section. This was once again true this year, even though half of the available marks were for the 'pure' probabilities requested in **(i)**, and the expectation was directly related only to the three marks in **(iii)**. A common error in **(i)** was to see the probability of all predictions being incorrect regarded as the complement of the probability of all being correct. Indeed, it was not uncommon among candidates attempting this question to see such an incorrect result for **(i)(b)** being given alongside perfectly correct results for **(a)**, **(c)** and **(d)**, even though this meant that the four probabilities did not sum to 1. What was pleasing, however, was the number of candidates who, even if their answers to earlier parts of the question did not score well, nevertheless gave perfectly correct explanations in their answers to **(v)**.

Answers: **(i)(a)** 1/48 **(b)** 15/48 **(c)** 23/48 **(d)** 9/48 **(iii)** $0.75 **(iv)** 400

Question 11

This question was attempted by almost all candidates, but a sizable proportion failed to realise exactly what the question was requiring them to do, i.e. extract some information from one diagram, manipulate it into a form which displayed a different aspect of the data, and then to present that form in another diagram. The most common error, as might have been expected, concerned those candidates who presented their entire answer in the form of the two-subject combinations, rather than extracting information about the separate subjects as the question required. However, among the large number who did do what the question asked, answers very often scored full or nearly full marks. The bar charts presented were of a high quality, the percentage calculation in **(iv)** was usually correct, and the vast majority were able to give a valid explanation in **(vi)**.

Answers: **(i)** B and C 36, B and P 40, C and P 24 **(ii)** B and C 20, B and P 32, C and P 48 **(iii)** Males B 114, C 90, P 96 Females B 65, C 85, P 100 **(iv)** 65.1%