STATISTICS

Paper 4040/12

Paper 12

Key Messages

After performing any calculation it is worth pausing to consider if the answer obtained is a reasonable one for the practical situation of the question.

It is very important to carefully read the words of a question to understand precisely what is required.

Candidates should always try to relate their knowledge to the specific requirements of a question, including the specific context involved, rather than simply writing out memorised general theory.

If a question specifies a certain degree of accuracy for numerical answers, full marks will not be obtained if the instruction is not followed.

General Comments

The overall standard of work was higher this year. A substantial number of candidates obtained very good marks, and there were few exceptionally low marks. It has been noted regularly in these reports that marks are often lost due to final answers not being given to the accuracy specifically stated in the question. A definite improvement in this respect was observed this year.

It has also been noted previously that a student of Statistics ought to be able to observe whether or not the result of a calculation is reasonable in a given practical situation. If it is clearly unreasonable, the work can be checked to find the error. But some candidates still seem to give no thought to the answer they obtain, treating the exercise as one in Pure Mathematics, having no practical relevance. For example, in a cross country running race (see **Question 8** below), it should have been obvious that the mean time taken by the senior competitors following the most difficult route could not have been less than ten minutes.

It will seem superfluous to remark that a question should be read carefully before an answer is attempted. Yet there were several instances on the paper (see **Questions 2**, **4(ii)**, **8(vii)** below) where this most basic advice for answering examination questions was not followed, and where candidates seemed to assume what they thought was to be done.

When questions are asked which require written answers, there is a tendency for some candidates to respond in a very general way, repeating apparently memorised points, without relating their knowledge to the particular context of the question (see **Question 3** below). Also, for example, in a situation involving accidents in the construction industry (see **Question 7(iv)** below) there should have been no explanations in terms of death rates, when there was no mention whatsoever of deaths anywhere in the question.

Comments on Specific Questions

Section A

Question 1

In part **(i)** the mean and standard deviation were usually evaluated correctly. But in part **(ii)** many answers said that the standard deviation would be unchanged after the error made in the gauge readings had been corrected. It appears as though these candidates were confusing this question with the theory concerning the effect on the mean and standard deviation of a variable by adding (or subtracting) a constant to each observation in a set of data.

Answers: **(i)** 46.5, 4.46; **(ii)** mean smaller, standard deviation larger

Question 2

This question was not to test how statistical measures are calculated, but to test whether or not a candidate could deduce which measure was which for a set of measures already calculated, from their relative numerical values. Many fully correct answers were seen. Many answers were also seen where the candidate did not read the question properly, but assumed, completely erroneously, that this was a set of raw data from which the stated measures were to be found.

Answers: **(i)** 48, 43, 53, 6, 36; **(ii)** 53

Question 3

Strong answers in part **(i)** showed a good appreciation of the advantages and disadvantages of the different survey methods in this particular situation. Weaker answers tended to be vague, speculative, or of a very general nature. It is not enough in this type of question to say, for example, only that something is "easy": one might validly ask in what way is it "easy"; what is it that makes it "easy"?

The difference between closed and open questions was generally well understood in part **(ii)**. The main weakness in answers tended to be seen in part **(ii)(a)** where there was sometimes too much focus on the example given, rather than on closed questions in general.

The answers given below are far from exhaustive, but give examples of what would be considered good answers.

Answers: **(i)(a)** citizens not in the telephone directory are excluded, **(b)** better response rate, **(c)** a very wide range of people can be reached very quickly, people without internet access are excluded; **(ii)(a)** only a limited number of answers is possible, **(b)** any relevant open question

Question 4

Many fully correct answers to part **(i)** were seen. In contrast, part **(ii)** was rarely answered well. This was another instance of candidates not reading the question correctly. The variable is clearly stated to be "the number of these vaccines received...", and this only takes the values 1, 2 or 3.

Answers: **(i)** the following numbers inserted into the correct spaces: **(a)** 19, **(b)** 20, **(c)** 17, **(d)** 34; **(ii)** 2

Question 5

Many candidates obtained good marks on this probability question. Errors tended to occur most frequently in parts **(ii)** and **(iii)**, with incorrect denominators being used. Some candidates made the solution more complex than it needed to be in part **(iv)** by considering male and female authors separately, rather than all of the authors together as a group.

Answers: **(i)** $\frac{3}{8}$, **(ii)** $\frac{1}{8}$, **(iii)** $\frac{2}{5}$, **(iv)** $\frac{34}{195}$

Question 6

Many candidates now recognise that, for a histogram, the frequency of a class is not always represented simply by the height of the relevant column. In taking account of the column areas, however, occasional errors were made which indicated that the labelling of the vertical axis had either not been read carefully, or not properly understood. So for the "under 50km/h" class, for example, the height was multiplied by 50 rather than 5. Part **(iv)** was least well done, with the total class frequency being offered, rather than the fraction of it indicated in the question.

Answers: **(i)** 116, **(ii)** 62, **(iii)** 40, **(iv)** 6

Section B

Question 7

The question on crude and standardised rates continues to be answered exceptionally well, and there was good application this year of basic knowledge to the problem of industrial accident rates. There were a few cases in part **(iv)** of the reason referring to death rates, and as there was no mention of death rates in the question, this could be given no credit. Only limited understanding was shown in part **(vi)** as to why the rates found in the previous part were crude rates.

Answers: **(i)** 106.5; **(ii)** 40, 47.9, 75, 162.3; **(iii)** 102.0; **(iv)** Fastbuild, because its standardised accident rate is lower; **(v)** Kwikbuild 30.7, Fastbuild 32.5, Kwikbuild; **(vi)** crude rates, a standardised rate is to eliminate differences in population structures, so is meaningless for one category

Question 8

This question tested the abilities of candidates to interpret statistical information, presented in a mixture of pictorial and tabular forms, relating to a particular situation. Responses were generally very good, with many obtaining a high proportion of the marks available. Because of incorrect work a few candidates produced answers to part **(v)** which, with pause for thought, should have been recognised as being utterly unrealistic. When most of the competitors referred to took more than two hours to complete the route, it should have been immediately apparent that the mean time taken could not have been, as was seen more than once, less than ten minutes.

To answer part **(vii)** it was necessary to use all the sources of information: pictogram, pie chart and table. Once more, this was an instance of a question not being read carefully, because many candidates did not base their calculation on the senior competitors who had *chosen* the moderate route, as the question states, but on the senior competitors who had *completed* the moderate route. Such candidates thereby used only one of the sources of information, not all three.

Answers: **(i)** 280, **(ii)** (35/100) × 120, **(iii)** 72, **(iv)** 12, **(v)** 141 minutes, **(vi)** 17.3%, **(vii)** 37.5%

Question 9

Candidates responded to this question well, and a good number obtained correct answers to all the numerical parts. One of the errors sometimes made in part **(iii)** was to divide the 60% by 2 for the overweight people, but then to read BMI for a cumulative frequency of 30% rather than 70%. A general limitation in answers to parts **(iii)** and **(iv)** was the absence of explanatory method. Whilst this did not matter when answers were correct, marks for method could not be awarded when they were incorrect.

There is no single good answer to part **(v)**. But to earn both marks it was necessary to say not only how the health of the people of the country had changed, but to give this specific support by citing more than one of the changes which had occurred, or citing actual statistics as calculated earlier in the question.

Answers: **(i)(a)** 23.5–23.8, **(b)** 26.2–26.5, **(c)** 21.2–21.5, **(d)** 29.5–29.8; **(ii)(a)** 57%–59%, **(b)** 36%; **(iii)** 29; **(iv)** 22%; **(v)** the population became more unhealthy, because the percentage healthy decreased from 58% to 36%, and the percentage obese increased from 7% to 22%

Question 10

Almost all candidates had a clear idea of the required steps in the plotting of data, and the calculation of averages, to find the equation of the line of best fit. However, a common error seen this year in the calculation of the semi-averages derived from the ordering of x values and y values as though they were unconnected with each other, rather than linked pairs of values. This serious error resulted in the loss of marks for many candidates.

There were mixed answers to part **(vi)**, and only a minority seemed to appreciate the point in part **(vii)**.

Answers: **(ii)** their x coordinates are not in the set of the four lowest x coordinates; **(iii)** (2, 41), (4.5, 69), (3.25, 55); **(iv)** m = 11.0–11.4, c = 18–19; **(v)** 52; **(vi)** Science, because the gradient of the line of best fit is the greatest; **(vii)** it would have been very difficult to know if the pupils performed well in tests because they liked the subject, or they liked the subject because they performed well in tests in it.

Question 11

There were some excellent fully correct answers to this question, but also some very weak ones. Where good solutions were seen, marks were most frequently dropped in part **(iii)(b)** where the case of 0 eggs was omitted. In weaker answers, little solid progress was made beyond part **(i)**.

Answers: **(i)(a)** 0.81, **(b)** 0.19; **(ii)** 0.117; **(iii)(a)** 0.00972, **(b)** 0.982

STATISTICS

Paper 4040/13

Paper 13

Key Messages

Candidates should always try to relate their knowledge to the specific requirements of a question, including the specific context involved, rather than simply writing out memorised general theory.

It is sound examination practice to show method clearly, so that marks for method can be awarded even if the answer obtained is incorrect.

If a question specifies a certain degree of accuracy for numerical answers, the instruction must be followed for full marks to be credited.

General Comments

The overall standard of work was comparable to that of last year. A wide range of marks was seen, but there were few very high marks. The best performances were on **Questions 1**, **4** and **6** in ***Section A***, and on **Question 10** in ***Section B***.

This year candidates paid much better attention than has often been the case in the past to following accuracy instructions, where given, as in **Question 10**.

In questions which require written answers, candidates should try to relate their knowledge to the specific context of the question rather than simply repeat memorised knowledge of a general nature. The latter tended to happen especially in **Question 5**, resulting in little creditworthy work.

Comments on Specific Questions

Section A

Question 1

Good knowledge of these measures was shown, and how they are found. There were many full mark answers, but a mark was sometimes lost in the explanation for the median, the need to order the data initially being omitted.

Answers: **(i)** 8 is the mode, and definition **(ii)** 9 is the median, and definition **(iii)** 11 is the mean, and definition

Question 2

The variable was usually identified as discrete in part **(i)**, but it was rarely explained what feature of the variable made it so. Some used as an incorrect reason the fact that the cumulative frequency only has integer values. A common error in part **(iii)** was to enter cumulative frequencies into the table.

Answers: **(i)** X is discrete, as it only takes integer values **(ii)** 0, 4 **(iii)** 0, 5, 15, 10, 0, 7, 6, 7

Question 3

In part **(a)**, the way in which methods were different was identified more easily than the way in which they were similar. Good answers referred to whether or not there was a need for a sampling frame, or for random numbers, and whether or not the method was biased. Few expressed clearly the way in which they were similar.

In part **(b)**, whilst the correct choice between biased and unbiased was often made, the reasons offered were usually not creditworthy. In particular, the fact that one of these is a form of random sampling, whilst the other is not, was rarely recognised.

Answers: **(a)** similar in that both sample proportionately from the different age groups; different in that stratified random sampling requires a sampling frame, whilst quota sampling does not
(b)(i) because there are likely to be fewer words on the last page of the chapter than on other pages, the sample is likely to be biased **(ii)** because a systematic sample is a form of random sampling, the sample is likely to be unbiased

Question 4

This question was a good source of marks for many. Any errors that were made were usually in plotting cumulative frequencies at class mid points, and occasionally in part **(iii)(b)**, finding the percentage smaller than, instead of larger than, 37.2 mm.

Answers: **(i)** 0, 8, 18, 35, 46, 50 **(iii)(a)** correct reading from the graph presented **(b)** 14%–16%

Question 5

Candidates demonstrated that they had studied the advantages and disadvantages of the different forms of chart, but often the answers seen made no reference to the practical situation of the question: that is, the company producing wood. Answers such as “values can be compared” were not given credit. It also seems not to have been appreciated that part **(i)** was not about the advantages and disadvantages of a dual bar chart considered on its own, but the advantages and disadvantages of a dual bar chart “as opposed to a percentage bar chart”. So any disadvantage offered which applies to both could not be credited.

Answers: **(i)** it shows actual amounts of wood; it only shows amounts for individual sizes **(ii)** total amount of wood of all sizes produced **(iii)** pie chart, sectional bar chart **(iv)** change chart

Question 6

A substantial number of full-mark answers to this question was seen, with clear reasons well expressed in part **(iii)**.

Answers: **(i)** 5 **(ii)** none of these citizens speaks all three languages **(iii)(a)** no, because the person would only speak two of the languages **(b)** yes, because the person would speak all three of the languages **(c)** no, because the person would only speak one of these languages

Section B

Question 7

Correct answers were most frequently seen to part **(a)** and part **(c)(i)**. It was puzzling in part **(b)** to see products of two fractions frequently offered, when clearly three children were being chosen. In part **(c)(ii)** only a few candidates recognised that two sums of two products would have to be worked out, corresponding to the first choices being both blue, or both white.

Answers: **(a)** 88/105 **(b)(i)** 1/35 **(ii)** 1/7 **(c)** 86/189

Question 8

Good answers were those in which the measures for *y* were calculated correctly, then used with minimal work to find the measures for *x*. Marks were frequently lost in all of parts **(iv)** – **(viii)**.

For part **(iv)**, the values in column 6 were often found by squaring those in column 5. For part **(vi)** and part **(vii)** it was often assumed that there had been just 7 visitors to the gallery, not 105. And for part **(viii)** many started again with the original data, instead of following the instructions of the question to use what had just been calculated in the previous parts.

Most answers to part **(ix)** did not make a judgment by looking at the nature of this particular distribution, but simply repeated apparently memorised theory about the use of the interquartile range.

Answers: **(i)** 15, 32.5, 37.5, 45, 55, 65, 85 **(ii)** –12, –5, –3, 0, 4, 8, 16 **(iii)** –72, –55, –12, 0, 104, 112, 64 **(iv)** 864, 275, 36, 0, 416, 896, 1024 **(v)** 141, 3511 **(vi)** 1.34 **(vii)** 5.62 **(viii)(a)** 48.4 **(b)** 14.1 **(ix)** because the distribution is reasonably symmetrical, standard deviation is preferable

Question 9

Many candidates recognised in part **(i)** that they could not simply read off the heights of these columns to produce the frequency table, but some also did not. For part **(ii)** most formed the correct grouping, but a good number made the mistake of grouping the classes in part **(i)** in pairs. In the latter case the classes did not have equal width, so an appropriate histogram could not be produced in part **(iii)**. A frequent problem with the histogram was that the vertical axis was not properly labelled, so a possible follow through mark could not be awarded on the heights drawn. Candidates should note that a vertical axis labelled "fd" or even "frequency density" is not good enough; the labelling should be of the form given for the histogram at the start of the question.

For part **(v)** marks were available for correct method following earlier errors, but the method had to be clearly shown to earn these.

Answers: **(i)** 24, 36, 32, 21, 18, 22, 19, 28 **(ii)** frequencies 24, 68, 80, 28 **(iv)** 19.3 cm–19.4 cm **(v)** 84%

Question 10

The question was very well done and a good source of marks for many candidates. The most common error was in part **(v)**, where the percentages in the second table were sometimes used instead of the percentages for the standard population.

Candidates understand very well that it is the standardised rate that has to be used to make fair comparisons in this type of situation; but they understand less well why it is that one town can have a higher crude rate, but a lower standardised rate, than the other.

Answers: **(ii)** 7.18 **(iii)** 5.56, 7.83, 11.86 **(iv)** 6.49 **(v)** 7.90 **(vi)** populations of the towns are differently structured in terms of age groups **(vii)** because its standardised death rate is lower, Eastbury

Question 11

It was a pity that, at the outset in part **(i)**, some candidates did not label the plotted points as instructed, as this labelling was needed later in the question to make judgments on the performance of the trainees. It was in these judgment parts, **(iv)** and **(vi)**, that most marks were lost.

Good answers were able to point out in part **(iv)** that results for A and B seemed to follow (different) straight lines, whilst those for C were quite erratic. Once the experienced technician's result was plotted in part **(v)** they then further added that B's results were clearly accurate. The best revised lines drawn in part **(vii)** passed very closely through B's results and that of the experienced technician.

Answers: **(ii)** (37.5, 104.5) **(iv)** results of A and B fall approximately on straight lines, whilst results of C are erratic **(vi)** experienced technician's result fits results of B very well, so it seems that results of B are accurate **(viii)** correct reading from revised line drawn

STATISTICS

Paper 4040/22
Paper 22

Key Message

The most successful candidates in this examination were able both to calculate the required statistics and to interpret their findings. In the numerical problems, candidates scoring the highest marks provided clear evidence of the methods they had used in logical, clearly presented solutions. In questions requiring written definitions, justification of given techniques and interpretation, the most successful candidates provided detail in their explanations with clear thought given to the context of the problem, where appropriate.

General Comments

In general, candidates did better on the questions requiring numerical calculations than on those requiring written explanations; in particular, candidates did well on the numerical parts of **Questions 8** and **10**. It was particularly pleasing to see, in **Questions 8(iii)** and **10(c)(iii)**, clearly laid out logical solutions. There were however three numerical questions that caused difficulty this year, namely parts **(ii)** and **(iii)** of **Question 2** and **Question 4(ii)**. Answers to questions requiring written explanations, such as **Questions 8(v)**, **10(c)(i)** and **10(c)(ii)**, were sometimes too vague or insufficiently detailed. However, in **Question 1**, for example, there were some very good descriptions of different types of variable and in **Question 7(a)(i)** clear purposes stated for finding moving average values. Graphs and charts were often accurately produced where necessary, but a common error in **Question 5** was for the vertical axis label to be missing.

Question 9, on probability and expectation, proved to be the least popular of the optional ***Section B*** questions, with each of the remaining ***Section B*** questions proving equally popular.

Comments on Specific Questions

Section A

Question 1

Most candidates found it easier to find examples for parts **(ii)** and **(iv)** than to produce the required definitions for parts **(i)** and **(iii)**. There were, however, some good responses seen in all parts and, in general, candidates did better on this question than on similar questions in the past. In part **(i)** the most common correct definitions seen for a discrete variable were ‘a variable whose outcomes can only take specific or exact values’ or ‘a variable which can be counted’. The most common incorrect answer seen was where candidates thought that discrete variables must take whole number values. Many correct examples were seen in part **(ii)** including height, weight and length. In part **(iii)** many candidates correctly described a qualitative variable as one which does not involve numbers or one which can only be described in words. The mark was not awarded to explanations which simply said that a qualitative variable is one which has quality, as further explanation was required. In part **(iv)** many correct examples of a discrete quantitative variable were seen including shoe size, the number of people on a bus and the number of leaves on a tree.

Question 2

Most candidates correctly identified the mode in part **(i)**. Many candidates struggled, however, with the remainder of this question. In part **(ii)**, for example, many candidates, rather than adding 1 to the total frequency before dividing by 2 to find the correct position for the median, simply divided 29 by 2. In part **(iii)** many candidates, correctly, made an attempt to work with a cumulative frequency of 18, but a common incorrect answer seen was $h = 7$. This occurred as a result of using an incorrect method for finding the position of the median for ungrouped data, as was also seen in part **(ii)**.

Answers: **(i)** 6; **(ii)** 5; **(iii)** 6.

Question 3

The majority of candidates scored full marks in part **(i)(a)** of this question. In part **(i)(b)** there were many good answers, with the most common errors being either not multiplying by 2 or not realising that if Maria eats the chocolates then this implies that the situation is 'without replacement'. Part **(ii)** proved to be more difficult for many candidates. Some candidates incorrectly included 3 attempts in their working or found only the probability of exactly 3 attempts. Some candidates produced tree diagrams with probabilities written on the branches, but no indication of further working which might have gained them some method marks. Some candidates missed the fact that in this part unwanted chocolates were returned to the tin, and thus incorrect denominators were sometimes seen.

Answers: **(i)(a)** 1/19, **(b)** 15/38; **(ii)** 16/25

Question 4

Correct answers, together with a correct reason, were seen in the work of the more able candidates in part **(i)**. Many candidates did not realise that the fact that the data contains extreme values is both the reason for the choice of the median as the measure of central tendency and the reason for the choice of the interquartile range as the measure of dispersion. Some candidates did not notice the presence of extreme values in the data (namely some large masses). Candidates tended to be more successful with part **(ii)** of this question. The most common error was for candidates to find one third, rather than two thirds, of 19 before adding it on to 12.

Answers: **(ii)** 25

Question 5

In part **(i)** most candidates were able to name the chart as a percentage sectional or a percentage component bar chart. The interpretation of this chart required in part **(ii)** was very well done by the vast majority of the candidates, with almost all candidates getting the correct numbers of males and most of those also getting the correct numbers of females. Accurately drawn dual bar charts were seen in part **(iii)**, although some candidates omitted the label on the vertical axis. In part **(iv)** it was encouraging to see that many candidates recognised that the dual bar chart provided them with actual numbers rather than simply percentages. Some candidates, however, simply stated that the dual bar chart was easier to read, without explaining that this was because it shows actual numbers.

Answers: **(ii)** 42, 36, 22; 36, 24, 60

Question 6

Many candidates found this to be the most difficult of the ***Section A*** questions. It was quite common in part **(i)** for candidates to give just one pair, rather than stating all of the pairs of the independent and mutually exclusive events. Candidates often gave A and D as their only answer to part **(i)(a)** and B and C as their only answer to part **(i)(b)**. In part **(ii)** many candidates successfully wrote down that the probability of A is 1/6 and the probability of B is 1/2. These two probabilities were then often simply added together by candidates, rather than, using the fact that these are independent events, finding the intersection by multiplying the two probabilities together and then subtracting this from the total. Some good attempts were made in part **(iii)**, but it was very rare to see both the smallest and the largest possible values both correct.

Answers: **(i)(a)** A and B, A and C, A and D, **(b)** B and C, C and D; **(ii)** 7/12; **(iii)** 0 and 5/6

Section B

Question 7

Many candidates were able to give two correct purposes of finding moving average values in part **(a)(i)**. Any two from 'to smooth out/eliminate the variation', 'to look for the trend', 'to find the seasonal components' or 'to make predictions' were required. In part **(a)(ii)** many candidates gave the correct answer of 3, although a considerable minority gave the incorrect answer of 4. In part **(a)(iii)** answers were usually consistent with any value suggested in part **(a)(ii)**, though many stopped at 'because *n* is odd/even' without the further explanation as to whether or not the moving average values correspond to original data points, which was required for full marks. In parts **(b)(i)** and **(b)(iii)** the values of *a*, *b* and *c* were usually correct, the centred moving average values were usually correctly plotted and a reasonable trend line was obtained. Candidates often had difficulty, however, in part **(b)(ii)**, with finding the seasonal component and, in part **(b)(iv)**, with using that seasonal component to estimate sales. Those candidates who correctly found the seasonal component in part **(b)(ii)**, by finding the differences between sales and moving average values for quarter II and then averaging these differences, often went on to correctly use the seasonal component in part **(b)(iv)**. A common error in part **(b)(iv)** was for a reading to be taken from the trend line, but for the seasonal component not then to be added to this reading.

Answers: **(a)(ii)** 3; **(b)(i)** $a = 75$, $b = 70.75$, $c = 82$; **(ii)** –9.6; **(iv)** 54 400

Question 8

Most candidates obtained the correct ratio in part **(i)**. In part **(ii)** most candidates produced correct or almost correct price relatives with some candidates omitting the 100s in the first column. Some weaker candidates appeared not to understand the term 'price relatives' and entered what looked like total expenditure values in the various categories. These candidates sometimes did the part **(ii)** calculation in part **(iii)** and recovered. Others continued with their very wrong values. Many candidates, however, did this part perfectly, producing well set out solutions and giving their answer to the required degree of accuracy. Part **(iv)** was also done perfectly by many candidates, though some ignored the instruction to use the index from part **(iii)** and did the calculation by finding the new cost in each category. Many candidates correctly gave reasons specific to the context of the problem presented, for example, that the amount of electricity used may have changed, that the number of hours worked may have changed or the amount of ingredients used may have changed for their answer to part **(v)**. A few candidates referred, incorrectly, to the inaccuracy introduced by rounding errors and some referred, incorrectly, to changes in prices. Vague answers, such as the weights or the quantities may have changed, were sufficient for some of the marks, but in order to score full marks the reasons provided needed to be in the context of the problem.

Answers: **(ii)** 100s in first column; 108, 122, 97 in second column; **(iii)** 101.3 or 101.4; **(iv)** $42 600

Question 9

This was the question omitted by most candidates or started and abandoned after part **(i)(a)**. Those candidates who continued with this question were usually able to score both marks in part **(i)(a)**. Although some perfect answers were produced, many of those who did attempt the whole question did not seem to have understood the basic process and, for example, thought that 'exactly 2 points' required 2 heads and 1 tail. In part **(i)(b)** some candidates gave partially correct responses by calculating $\frac{1}{4} \times \frac{1}{4} \times \frac{3}{4}$, but many missed the fact that 2 points could be achieved in 3 ways. In part **(ii)** many candidates correctly realised that X could take the values 0, 1, 2 or 3, but the probabilities of each of these outcomes were usually incorrect and very often these probabilities did not sum to 1. In part **(iii)** some correct attempts to use expectation were seen, although sometimes the working was rather disorganised. In part **(iv)** many candidates realised that the probability of a head is 1/5. Again some correct attempts to use expectation were seen, but many candidates had abandoned this question by this stage.

Answers: **(i)(b)** 9/64; **(ii)** 27/64, 27/64, 9/64, 1/64; **(iii)** $22; **(iv)** $20 so should not risk throwing extra coin

Question 10

Candidates did well on the numerical parts of this question. Correct answers with clear working were often seen in parts **(a)(i)** and **(a)(ii)**. In part **(b)** correct values for the mean and standard deviation were usually seen in the first three rows of table. A common error was to see the values 22 and 6 for the mean and standard deviation, respectively, in the final row of the table. In part **(c)(i)** many candidates achieved one of the two available marks. They were often able to state that the marks for Algebra were better, but a correct comparison using the class standard deviations was much less common. The more able candidates were able to state that the marks in Algebra were generally more varied and there were more correct responses to this question than on a similar question requiring the comparison of interquartile ranges last year. Part **(c)(ii)** was a difficult question requiring candidates to compare Priyanka's mark with the class mean in terms of the class standard deviation for each of Algebra and Geometry. It was necessary to show that she did better in Geometry because her mark was two standard deviations above the mean in this subject, whereas in Algebra her mark was only one standard deviation above mean. Only the most able candidates were able to score both of the marks in this part. Many candidates scored full marks in part **(iii)** with clearly set out solutions.

Answers: **(a)(i)** 64; **(ii)** 1.5; **(b)** 10, 3; 5.5, 1.5; 58, 15; 11, 3; **(c)(iii)** 12.5.

Question 11

Responses to questions of this sort have improved over the years. In part **(i)** many candidates obtained the correct simple random sample and it was pleasing to see that the number 47 was not repeated too often. In part **(ii)** many candidates correctly found a systematic sample, although a few used an interval of 9 instead of 10, and of the three types of sample being tested in this question, the systematic sample was the least well done. Most candidates were able to find the correct sample sizes in part **(iii)**, even though they had not been provided with the necessary totals. Many correct answers were seen in part **(iv)**, but some candidates did not give enough detail when explaining why Asad's sample was not accurate. It was necessary to state that Asad's sample over-represents machine A (or under-represents machine B or C) and that Omar's sample accurately represents jars filled by each machine. In part **(v)** many correct stratified samples were seen and in part **(vi)** correct sample sizes were calculated, which included the need to round answers to the nearest whole number. Part **(vii)**, which required candidates to explain why it was more appropriate to stratify by machine, was a difficult final part to this question. Only the most able candidates looked back to the start of the question to find the purpose of sampling in this case. As the purpose was to check the mass of jam in each jar, it was appropriate to stratify by machine, rather than packer, as it was the machine that was responsible for the mass of jam. While only the most able candidates were successful in this part of the question, there were more correct responses than in a similar part on last year's paper.

Answers: **(i)** 47, 00, 51, 32, 85, 11, 67, 05, 10; **(ii)** 01, 11, 21, 31, 41, 51, 61, 71, 81; **(iii)** 2, 3, 4; **(v)** 44, 03, 59, 14, 27, 20, 78, 60, 81; **(vi)** 4, 5

STATISTICS

Paper 4040/23
Paper 23

Key Message

The most successful candidates in this examination were able both to calculate the required statistics and to interpret their findings. In the numerical problems, candidates scoring the highest marks provided clear evidence of the methods they had used in logical, clearly presented solutions. In questions requiring written definitions, justification of given techniques and interpretation, the most successful candidates provided detail in their explanations with clear thought given to the context of the problem, where appropriate.

General Comments

In general, candidates did better on the questions requiring numerical calculations, than on those requiring written explanations; in particular, candidates did well on **Questions 3 and 4** and the numerical parts of **Questions 7** and **8**. It was particularly pleasing to see in **Questions 3(iv)** and **8(v)**, on finding the standard deviation and the interquartile range, respectively, clearly laid out logical solutions. There were however two numerical questions that caused difficulty this year, namely parts **(ii)** to **(iv)** of **Question 1** and **Question 2**. Answers to questions requiring written explanations, such as **Questions 6(a)** and **8(vii)**, were sometimes too vague. However, in **Question 6(b)**, for example, there were some very good explanations provided for the difference between discrete and continuous variables. Graphs and charts were often accurately produced where necessary, but a common error in **Question 5** was for labelling to be missing.

Question 9, on probability, proved to be the least popular of the optional ***Section B*** questions and it was also the question that those who attempted it found the most difficult. **Question 8** on linear interpolation proved to be the most popular of the optional questions.

Comments on Specific Questions

Section A

Question 1

Most candidates were successfully able to calculate the mode in part **(i)**. Many candidates struggled, however, with the remainder of this question. In part **(ii)**, for example, many candidates, rather than adding 1 to the total frequency before dividing by 2, to find the correct position for the median, simply divided 46 by 2. In part **(iii)** many candidates, correctly, made an attempt to work with a cumulative frequency of 29, but a common incorrect answer seen was $k = 12$. This occurred as a result of using an incorrect method for finding the position of the median for ungrouped data, as was also seen in part **(ii)**. It was rare in part **(iv)** to see a correct answer of 9.

Answers: **(i)** 17 **(ii)** 1 **(iii)** 11 **(iv)** 9

Question 2

This proved to be a difficult question on expectation. Many candidates were unable to deal with the fact that 40% and 60% of letters are sent by 1st and 2nd class post, respectively, with many ignoring this information altogether in their solutions. Attempts were seen to multiply incorrect probabilities by the number of days in an attempt to find the expected number of days for a letter to be delivered. Those candidates who have successfully found the probabilities were often able to go on and correctly find the expectation.

Answer: 2.1

Question 3

Most candidates were able to find the required totals and the mean estimate. In part **(iv)** the correct answer was seen in many cases, with well set out solutions, but some candidates appeared not to know the correct formula for the variance or the standard deviation.

Answers: **(i)** 430 **(ii)** 17.2 **(iii)** 8131 **(iv)** 5.42

Question 4

As with question 3, many candidates produced fully correct solutions to this question. The most common error seen in part **(i)** was to have two, rather than one, unknown in the equation. A few candidates had a correct expression but were unable to rearrange correctly and then solve it. In part **(ii)** many candidates had a correct standardised term, with the unknown standard deviation the denominator, but this did not always appear in a fully correct equation.

Answers: **(i)** 33 **(ii)** 25

Question 5

Many accurate charts were seen in both parts of this question. Marks were, however, sometimes lost due to a lack of labelling of the vertical axes. It was very important in part **(i)** to label the vertical axis as 'expenditure (in dollars)' and in part **(ii)** as 'percentages'. Some weaker candidates did not start their scales on the vertical axes at 0 and thus the height of their bars were not proportional to the expenditure (in part **(i)**) or the percentage of the expenditure (in part **(ii)**).

Answers: **(ii)** 27%, 33%, 40%; 31%, 33%, 36%

Question 6

There were many partially correct responses to part **(a)**. Candidates were often able to explain what is meant by a qualitative variable, namely one with non-numerical outcomes, but they were not always able to explain why this means that it is not possible to illustrate such data in the form of a histogram. In addition candidates needed to explain that in a histogram area is proportional to frequency and, with no class widths, calculation of such an area is not possible. In part **(b)** there were some well explained comparisons made and this was good to see in a question requiring written explanation. Examples of correct comparisons seen were, "*a discrete variable can only take certain values within its range, whereas a continuous variable can take all values within its range*" and "*a discrete variable is counted whereas a continuous variable is measured*". A commonly seen incorrect answer was that discrete variables can only take whole number values. Answers to part **(c)(i)** were often incorrectly given as 14.5 or 14, whereas answers to part **(c)(ii)** were more often correct.

Answers: **(c)(i)** 15 **(c)(ii)** 14.5

Section B

Question 7

Some candidates who embarked upon this question abandoned it after part **(i)**. The most common error in part **(i)** was caused by candidates not noticing that each box contained 3 cricket balls. Most candidates who found the correct ratio in part **(i)** were able to go on and find correct price relatives in part **(ii)** and then use these values to find a correct weighted aggregate cost index in part **(iii)**. The majority of candidates were also able to find a correct estimate for the total cost of running the club in part **(iv)**. Fewer candidates, however, were able to provide reasons why their estimate may be very different from the true cost. The most commonly seen correct answers in part **(v)** were that number of balls used may have changed or that the number of hours worked by the groundsman may have changed. Some commonly seen incorrect answers were those which had been accounted for within the information included in the calculation, for example "*the wage rate may have changed*" or "*there may have been inflation*". Some answers did not give sufficient detail, for example, "*the groundsman may have become ill*", without giving further details as to how this would affect the calculation. In this case it would be necessary to add that he would then be able to work fewer hours.

Answers: **(ii)** 90, 102, 105, 103 **(iii)** 102 **(iv)** $21 675

Question 8

Most candidates found the correct modal class in part **(i)** and the correct cumulative frequencies in part **(iii)**. It was rare however see a correct answer for the maximum possible value of the range in part **(ii)**. Common incorrect answers seen included 197, the maximum frequency, and 186, the difference between the maximum and minimum frequency. In parts **(iv)** and **(v)** estimates of the median and the interquartile range were usually correct with clearly set out working. The position of the median, for example, was usually correctly identified for grouped data by taking the total frequency and dividing it by two. In part **(vi)** most candidates were unable to comment that the distribution was not symmetrical. In part **(vii)** some candidates correctly identified that the gradient would be steepest around the 2 – under 3 class, but the reason was sometimes incorrectly given as 'the difference in the frequency is greatest at this point', rather than that class frequency is greatest at this point or that the difference in the cumulative frequency is greatest at this point.

Answers: **(i)** 2 – under 3 **(ii)** 8 cm **(iii)** 12, 209, 242, 255, 379, 401, 412, 500 **(iv)** 4.62 **(v)** 2.57, 5.97, 3.4 **(vi)(a)** 2.04 or 2.05 and 1.35

Question 9

This was the least popular of the ***Section B*** questions. In part **(a)(i)** many candidates correctly stated that mutually exclusive events were those that could not occur at the same time. In part **(a)(ii)** candidates needed to give an example of a pair of mutually exclusive events. These were often incorrect as they were not outcomes of the same experiment, for example, a candidate may give one event as getting a head on a coin and the another event as getting a 6 on a die. A correct pair of mutually exclusive events would be, for example, getting a 6 on a die and getting a 4 on the die. In part **(a)(iii)(a)** candidates often incorrectly stated that *A* and *B* are not mutually exclusive because the sum of the probabilities is not 1, rather than stating that the sum of the probabilities is greater than 1. The answer to part **(a)(iii)(b)** was often correct. Most candidates were able to find the correct probability in part **(b)(i)**. Parts **(b)(ii)** and **(b)(iii)** were also usually correct with any errors occurring in the denominator of the fraction. Fewer candidates were successful with part **(b)(iv)**, with a common error being that the problem was considered to be with rather than without replacement. Occasionally addition rather than multiplication of the fractions was seen. Most solutions to part **(b)(v)** were incorrect, however some candidates correctly had denominators of 60 and 59 in their expressions. There are a number of ways to solve this problem, the most commonly correct method seen being $7/60 \times 12/59 + 28/60 \times 13/59$.

Answers: **(a)(iii)(b)** 0.3 **(b)(i)** 2/5 **(ii)** 23/35 **(iii)** 11/25 **(iv)** 1/177 **(v)** 112/885

Question 10

Most candidates produced a correct simple random sample in part **(i)**, with the most common error being the inclusion of the number 18. In part **(ii)(a)** the most common error was 03 being given for the largest possible two-digit number for the first person selected. In part **(ii)(b)** the most common error was 05 for the first person selected and this occurred sometimes in cases where the candidate had got the previous part correct. In part **(ii)(c)** values outside the range were sometimes seen. The stratified samples in parts **(iii)** and **(iv)** were usually correct. In part **(v)** reasons for each conclusion needed to be stated clearly. In the case of the sample stratified by friend/relative it was necessary to state that this sample is also representative in terms of age group. In the case of the sample stratified by age group it was necessary to state that this sample over-represents friends or under-represents relatives.

Answers: **(i)** 12, 00, 07, 09, 01 **(ii)(a)** 00, 02 **(ii)(b)** 00; **(ii)(c)** 03, 06, 09, 12 **(iii)(a)** 3 friends, 2 relatives **(iii)(b)** 06, 09, 08, 04, 02 **(iv)(a)** 2 from Group I, 2 from Group II, 1 from Group III **(iv)(b)** 11, 13, 10, 02, 09

Question 11

Some candidates misunderstood the question in part **(i)** and gave general reasons for calculating moving averages rather than reasons for calculating a 5-point moving average specifically. It was necessary to explain that each cycle is of length 5 days. A general question regarding the purpose of calculating moving averages appears in part **(vi)**. In part **(ii)** most candidates gave as a correct answer that each cycle contains an odd number of observations. Alternatively, the correct answer could be expressed as being because the moving average values are at the same point in time as the original values. In part **(iii)** the plots were usually correct. Most candidates spotted the clear cyclical pattern. An alternative answer would have been that there is no clear upward or downward long-term trend. The calculations in part **(iv)** and the plots in part **(v)** were usually correct. In part **(vi)** most candidates correctly stated the purpose of calculating moving averages, namely to eliminate seasonal variation or to find the trend, but the subsidiary question regarding how well this had been achieved in this case was less well answered. It was necessary for candidates to state that the purpose had been achieved well in this case. In part **(vii)** most candidates were able to draw a suitable trend line. However some candidates followed too closely the earlier moving average values and ignored the later ones. In part **(viii)** a common incorrect answer was $q = 3$. In part **(ix)** working was sometimes missing, which might have been worth a mark had it been shown.

Answers: **(iv)** $x = 127$, $y = 24.8$ **(viii)** $q = –3$ **(ix)** 17