STATISTICS

Paper 4040/12
Paper 1

Key Messages

A candidate should always be aware of whether or not the answer to a question which has been obtained is a sensible one. If the answer is clearly not sensible, the work should be checked to find the error.

A question should always be read carefully so that the information that is given, and the task to be carried out, are both clearly understood.

When a question asks for a specific method to be employed, credit can only be obtained for using that method.

General comments

The overall standard of work was good. Candidates showed reasonable awareness of the need to maintain accuracy in carrying out calculations, and to follow any accuracy instructions in presenting their answers.

It has been stressed regularly in these reports that a student of Statistics ought to know whether or not the result of a calculation is reasonable, especially in a particular practical situation. If it is obviously unreasonable, or even impossible, the work leading to it should be reviewed to find the mistake. Unfortunately some candidates seem to regard Statistics as a form of pure mathematics without practical relevance. For example, in costing a family's water bill over 80 days (see **Question 9(v)** below) it should have been obvious that it was highly unlikely to have been just one or two dollars, or many millions of dollars. In analysing the performance of students on university courses (see **Question 7** below) it should have been obvious that a pass rate of more than 100% is impossible.

There was evidence on this paper of candidates either reading not carefully enough, or ignoring, information given in the question (see **Question 9(v)**, **Question 10(vi)** and **(vii)** below). Marks were lost frequently as a consequence.

Comments on specific questions

Section A

Question 1

This was very well done. Almost all charts were drawn accurately and labelled clearly.

Answers: **(i)** 198, 54, 108

Question 2

This was also very well done. There were occasional errors in the use of the information given before part **(iii)** on scheduled departures.

Answers: **(i)** 0, 3 in correct cells **(ii)** 8, 12, 35 in correct cells **(iii)** 40, 10, 29, 6 in correct cells

Question 3

This question caused difficulties for many candidates. Few obtained full marks in part **(i)**. Whilst many knew the assumed mean method for finding the mean, they then often abandoned it for finding the standard deviation. The method not only has computational advantages for some types of data, but perhaps, even more importantly, aids understanding of the concept of dispersion when applying it to standard deviation. The instruction on the method to be used was emphasised in the question, and no credit was given if it was ignored.

In part **(ii)**, a few candidates gave for the range the value of the standard deviation from part **(i)**.

Answers: **(i)** 1002, 6.52 (using assumed mean method) **(ii)** 922, 19

Question 4

Performance here was very varied, with a good number of full mark answers, and a good number of answers earning no marks, being seen. Many candidates still demonstrate limited understanding of what the different regions of a Venn diagram represent.

Answers: **(i)** 17 **(ii)** 19 **(iii)** 35 **(iv)** 9 **(v)** 11

Question 5

This question was generally answered well. Most errors occurred either in part **(iii)**, where a denominator of 87 instead of 84 was used, or in part **(iv)**, where only one possible order of selection was considered, or probabilities for 'with replacement' rather than 'without replacement' were used.

Answers: **(i)** 1/29 **(ii)** 64/87 **(iii)** 23/28 **(iv)** 84/1247

Question 6

Most candidates obtained reasonable marks on this question. The best answers to parts **(i)** and **(ii)** showed, through clear annotation of the diagram, the method of solution. Many full mark answers to part **(iii)** were seen, demonstrating good understanding of the properties of a distribution which the mean and standard deviation measure. A common exception to this occurred amongst candidates whose answers to parts **(iii)(b)** and **(iii)(c)** were reversed. Presumably such candidates incorrectly interpreted 'least in agreement' as being associated with a small standard deviation rather than a large one.

Answers: **(i)** 54 **(ii)** 25, 39 **(iii)(a)** comfort **(b)** cost **(c)** punctuality

Section B

Question 7

The question on crude and standardised rates was again answered exceptionally well, and many fully correct answers were seen. Almost all candidates worked with the percentages properly, and in part **(vi)** arrived at a conclusion which was both correct and appropriate to the context. This was however one of the questions where impossible answers (referred to in general comments above) were occasionally seen. Candidates ought to have been aware that if they obtained a pass rate of more than 100% something had to be wrong somewhere.

Answers: **(ii)** 100%, 85%, 62.5%, 55% **(iii)** 76.8% **(iv)** 79.5% **(v)** 78.0% **(vi)** Hale, because it has the higher standardised pass rate

Question 8

In part **(i)** greater awareness was shown than has sometimes been the case in the past of the need to maintain accuracy in the calculations through appropriate use of the calculator. Even so, those candidates who used 21.3, or even 21.28, instead of 1277/60, when finding the standard deviation were unable to obtain full marks. General improvement was seen in the reduced use for finding standard deviation of the method based on $\Sigma f(x - \text{mean})^2$.

Good understanding of the reading of a histogram was demonstrated in part **(ii)**, with more candidates obtaining the number of days correctly than those simply reading off the heights of the columns.

There were few fully correct answers to part **(iii)**. The overwhelming reason for this was that many candidates seemed to think they had finished when they had calculated the mean time taken by Tariq walking home from work. This was also a question where completely unreasonable answers were presented. For the times given in the two tables, it should have been obvious that the answer could not have been as small as two or three minutes, or as long as two or three days.

Answers: **(i)** 21.3, 2.47 **(ii)** 4, 8, 17, 21, 10 **(iii)** 54

Question 9

Many curves were correct in part **(ii)**, but many were also incorrect, the cumulative frequencies being plotted at class mid-points. In part **(iii)** candidates were very good at finding the median and interquartile range from the curve, with the best answers clearly demonstrating, with lines drawn on the graph, how this was done. They were less successful on the percentile, some either incorporating 375 into the calculation, or assuming that they only had to find the cumulative frequency corresponding to 375 litres.

In attempting to calculate the water bill in part **(v)**, many errors were seen. The information given about the mean and median being approximately the same seems to have been largely ignored (referred to in general comments above), seriously impeding any possible progress in obtaining a correct solution. In attempting to find the total amount of water consumed by the family over this period of time, it was common to see neither the value of the median, nor the 80 days, used. This resulted frequently in highly unrealistic answers (also referred to in general comments above).

To allow for small differences in the graphs drawn by candidates, in parts **(iii)**, **(iv)** and **(v)** examiners accepted values within specified ranges. The answers below give examples of values falling within these ranges.

Answers: **(i)** 4, 15, 35, 60, 74, 80 **(iii)(a)** 360 litres **(b)** 85 litres **(c)** 60 **(iv)** 0.4 **(v)** $92

Question 10

The early parts of the question were well answered with candidates displaying good ability in interpreting the given pictorial information. In part **(v)**, examiners experienced some difficulty, when answers were incorrect, in understanding the candidate's method: with no words in answers it was difficult to decide whether the candidate was working on totals, the Commercial Hotel, or all the others except the Palm Beach. In order for marks to be awarded for method on incorrect answers, the method must be shown clearly.

Very many answers to parts **(vi)** and **(vii)** showed again (referred to in general comments above) that the information given in the questions had either not been read properly, or ignored. These are clearly business visitors and so are more likely to stay in a particular hotel. This needed to be considered in forming the appropriate number of cases for the probability calculations.

Answers: **(i)** 15 000 **(ii)** 25 000 **(iii)** 15.4% **(iv)(a)** 1/2 **(b)** 1/2 **(c)** 3/5 **(v)** 350 **(vi)** 0.86 **(vii)** 0.0325

Question 11

Answers to the first three parts were generally very good, with clearly plotted points, a correctly calculated upper semi-average, and a well explained method in finding the line of best fit. Most candidates used the averages in finding the latter, but a few used instead points chosen from their line. Candidates need to be aware that, whilst it is perfectly correct to choose points from the line, it commonly results in loss of accuracy: values read from the line are almost always less precise than the calculated averages, through which the line must necessarily pass.

In part **(iv)** a mark was frequently lost as a result of the *additional* number of weeks not being found. And in part **(v)** the idea of extrapolation was not mentioned as often as it might have been.

Some excellent explanatory answers were seen to part **(vi)**, but there was also some confusion between parts **(a)** and **(b)** as to in which part the focus should be on the gradient, and in which the intercept.

Answers: **(ii)** (6.5, 62) **(iii)** $y = –8.5x + 117.25$ **(iv)** 6 **(v)** involves substantial extrapolation beyond the range of the given data, and the same trend might not be maintained **(vi)(a)** George, because the constant, *c*, in his equation is the largest **(b)** Joseph, because the gradient, *m*, in his equation indicates his line has the steepest negative gradient.

STATISTICS

Paper 4040/13
Paper 1

Key Messages

A candidate should always be aware of whether or not the answer to a question which has been obtained is a sensible one. If the answer is clearly not sensible, the work should be checked to find the error.

A question should always be read carefully so that the information that is given, and the task to be carried out, are both clearly understood.

When a question asks for a specific method to be employed, credit can only be obtained for using that method.

General comments

The overall standard of work was good. Candidates showed reasonable awareness of the need to maintain accuracy in carrying out calculations, and to follow any accuracy instructions in presenting their answers.

It has been stressed regularly in these reports that a student of Statistics ought to know whether or not the result of a calculation is reasonable, especially in a particular practical situation. If it is obviously unreasonable, or even impossible, the work leading to it should be reviewed to find the mistake. Unfortunately some candidates seem to regard Statistics as a form of pure mathematics without practical relevance. For example, in costing a family's water bill over 80 days (see **Question 9(v)** below) it should have been obvious that it was highly unlikely to have been just one or two dollars, or many millions of dollars. In analysing the performance of students on university courses (see **Question 7** below) it should have been obvious that a pass rate of more than 100% is impossible.

There was evidence on this paper of candidates either reading not carefully enough, or ignoring, information given in the question (see **Question 9(v)**, **Question 10(vi)** and **(vii)** below). Marks were lost frequently as a consequence.

Comments on specific questions

Section A

Question 1

This was very well done. Almost all charts were drawn accurately and labelled clearly.

Answers: **(i)** 198, 54, 108

Question 2

This was also very well done. There were occasional errors in the use of the information given before part **(iii)** on scheduled departures.

Answers: **(i)** 0, 3 in correct cells **(ii)** 8, 12, 35 in correct cells **(iii)** 40, 10, 29, 6 in correct cells

Question 3

This question caused difficulties for many candidates. Few obtained full marks in part **(i)**. Whilst many knew the assumed mean method for finding the mean, they then often abandoned it for finding the standard deviation. The method not only has computational advantages for some types of data, but perhaps, even more importantly, aids understanding of the concept of dispersion when applying it to standard deviation. The instruction on the method to be used was emphasised in the question, and no credit was given if it was ignored.

In part **(ii)**, a few candidates gave for the range the value of the standard deviation from part **(i)**.

Answers: **(i)** 1002, 6.52 (using assumed mean method) **(ii)** 922, 19

Question 4

Performance here was very varied, with a good number of full mark answers, and a good number of answers earning no marks, being seen. Many candidates still demonstrate limited understanding of what the different regions of a Venn diagram represent.

Answers: **(i)** 17 **(ii)** 19 **(iii)** 35 **(iv)** 9 **(v)** 11

Question 5

This question was generally answered well. Most errors occurred either in part **(iii)**, where a denominator of 87 instead of 84 was used, or in part **(iv)**, where only one possible order of selection was considered, or probabilities for 'with replacement' rather than 'without replacement' were used.

Answers: **(i)** 1/29 **(ii)** 64/87 **(iii)** 23/28 **(iv)** 84/1247

Question 6

Most candidates obtained reasonable marks on this question. The best answers to parts **(i)** and **(ii)** showed, through clear annotation of the diagram, the method of solution. Many full mark answers to part **(iii)** were seen, demonstrating good understanding of the properties of a distribution which the mean and standard deviation measure. A common exception to this occurred amongst candidates whose answers to parts **(iii)(b)** and **(iii)(c)** were reversed. Presumably such candidates incorrectly interpreted 'least in agreement' as being associated with a small standard deviation rather than a large one.

Answers: **(i)** 54 **(ii)** 25, 39 **(iii)(a)** comfort **(b)** cost **(c)** punctuality

Section B

Question 7

The question on crude and standardised rates was again answered exceptionally well, and many fully correct answers were seen. Almost all candidates worked with the percentages properly, and in part **(vi)** arrived at a conclusion which was both correct and appropriate to the context. This was however one of the questions where impossible answers (referred to in general comments above) were occasionally seen. Candidates ought to have been aware that if they obtained a pass rate of more than 100% something had to be wrong somewhere.

Answers: **(ii)** 100%, 85%, 62.5%, 55% **(iii)** 76.8% **(iv)** 79.5% **(v)** 78.0% **(vi)** Hale, because it has the higher standardised pass rate

Question 8

In part **(i)** greater awareness was shown than has sometimes been the case in the past of the need to maintain accuracy in the calculations through appropriate use of the calculator. Even so, those candidates who used 21.3, or even 21.28, instead of 1277/60, when finding the standard deviation were unable to obtain full marks. General improvement was seen in the reduced use for finding standard deviation of the method based on $\Sigma f(x - \text{mean})^2$.

Good understanding of the reading of a histogram was demonstrated in part **(ii)**, with more candidates obtaining the number of days correctly than those simply reading off the heights of the columns.

There were few fully correct answers to part **(iii)**. The overwhelming reason for this was that many candidates seemed to think they had finished when they had calculated the mean time taken by Tariq walking home from work. This was also a question where completely unreasonable answers were presented. For the times given in the two tables, it should have been obvious that the answer could not have been as small as two or three minutes, or as long as two or three days.

Answers: **(i)** 21.3, 2.47 **(ii)** 4, 8, 17, 21, 10 **(iii)** 54

Question 9

Many curves were correct in part **(ii)**, but many were also incorrect, the cumulative frequencies being plotted at class mid-points. In part **(iii)** candidates were very good at finding the median and interquartile range from the curve, with the best answers clearly demonstrating, with lines drawn on the graph, how this was done. They were less successful on the percentile, some either incorporating 375 into the calculation, or assuming that they only had to find the cumulative frequency corresponding to 375 litres.

In attempting to calculate the water bill in part **(v)**, many errors were seen. The information given about the mean and median being approximately the same seems to have been largely ignored (referred to in general comments above), seriously impeding any possible progress in obtaining a correct solution. In attempting to find the total amount of water consumed by the family over this period of time, it was common to see neither the value of the median, nor the 80 days, used. This resulted frequently in highly unrealistic answers (also referred to in general comments above).

To allow for small differences in the graphs drawn by candidates, in parts **(iii)**, **(iv)** and **(v)** examiners accepted values within specified ranges. The answers below give examples of values falling within these ranges.

Answers: **(i)** 4, 15, 35, 60, 74, 80 **(iii)(a)** 360 litres **(b)** 85 litres **(c)** 60 **(iv)** 0.4 **(v)** $92

Question 10

The early parts of the question were well answered with candidates displaying good ability in interpreting the given pictorial information. In part **(v)**, examiners experienced some difficulty, when answers were incorrect, in understanding the candidate's method: with no words in answers it was difficult to decide whether the candidate was working on totals, the Commercial Hotel, or all the others except the Palm Beach. In order for marks to be awarded for method on incorrect answers, the method must be shown clearly.

Very many answers to parts **(vi)** and **(vii)** showed again (referred to in general comments above) that the information given in the questions had either not been read properly, or ignored. These are clearly business visitors and so are more likely to stay in a particular hotel. This needed to be considered in forming the appropriate number of cases for the probability calculations.

Answers: **(i)** 15 000 **(ii)** 25 000 **(iii)** 15.4% **(iv)(a)** 1/2 **(b)** 1/2 **(c)** 3/5 **(v)** 350 **(vi)** 0.86 **(vii)** 0.0325

Question 11

Answers to the first three parts were generally very good, with clearly plotted points, a correctly calculated upper semi-average, and a well explained method in finding the line of best fit. Most candidates used the averages in finding the latter, but a few used instead points chosen from their line. Candidates need to be aware that, whilst it is perfectly correct to choose points from the line, it commonly results in loss of accuracy: values read from the line are almost always less precise than the calculated averages, through which the line must necessarily pass.

In part **(iv)** a mark was frequently lost as a result of the *additional* number of weeks not being found. And in part **(v)** the idea of extrapolation was not mentioned as often as it might have been.

Some excellent explanatory answers were seen to part **(vi)**, but there was also some confusion between parts **(a)** and **(b)** as to in which part the focus should be on the gradient, and in which the intercept.

Answers: **(ii)** (6.5, 62) **(iii)** $y = –8.5x + 117.25$ **(iv)** 6 **(v)** involves substantial extrapolation beyond the range of the given data, and the same trend might not be maintained **(vi)(a)** George, because the constant, *c*, in his equation is the largest **(b)** Joseph, because the gradient, *m*, in his equation indicates his line has the steepest negative gradient.

STATISTICS

Paper 4040/22
Paper 2

Key message

This examination requires candidates to be able to produce statistical diagrams, to calculate statistics and to interpret findings. The best statistical diagrams should be accurately drawn, taking care to use scales correctly, and have clearly labelled axes and, where appropriate, a key. Candidates scoring the highest marks in the numerical problems will provide clear indications of the methods they have used in logical and clearly presented solutions. In questions that require written definitions, justification of given techniques or interpretation, the most successful responses will include detailed explanations with, where appropriate, clear consideration of the context of the problem.

General comments

As usual, candidates did better on the questions requiring numerical calculations than on those requiring written explanations; in particular, candidates did well in the numerical parts of **Question 3** on scaling, **Question 7(ii)** on moving averages and the early parts of **Question 8** on weighted aggregate cost indexes. They also did well in **Question 4** on sampling and **Question 11(ii)**, which involved interpreting pictorial representations of data. A numerical question that caused difficulty this year was the final part of **Question 6**, on finding the standard deviation when combining sets of data. The final parts of **Question 9**, on expectation, also caused some difficulty. Answers to questions requiring written explanations, such as **Questions 4(iii)**, **7(i)**, **10(iv)** and **11(vi)** were sometimes insufficiently detailed; however in **Question 1(iii)**, for example, there were some good descriptions of the nature of variables and in **Question 7(iv)** clear descriptions of the trend. Other questions that caused difficulty this year were the parts of **Question 2** that required interpretation in words of a probability statement and the final parts of **Question 5** where candidates needed to consider their assumptions when making probability calculations.

Question 9, on probability and expectation, proved to be the least popular of the optional ***Section B*** questions, with **Question 8**, on price relatives and index numbers, **Question 10**, on linear interpolation, and **Question 11**, on pictorial representations of data, proving to be the most popular.

Comments on specific questions

Section A

Question 1

In part **(i)**, many candidates correctly identified that *B*, the number of cars in the car park at 9 am, was not a variable, but it was less common to see *E*, the height of the headteacher's car, also correctly identified. A common error in part **(ii)** was to see *B* rather than *C* given as the answer. In part **(iii)** it was good to see correct use being made of the terms 'continuous data', to describe the height of each car, and 'qualitative data', to describe the colour of each car.

Answers: **(i)** *B* and *E* **(ii)** *C*

Question 2

In part **(i)**, most candidates used the correct formula to find the required probability, although it was not uncommon to see independence being assumed and the probabilities of *A* and *B* simply being multiplied. Many candidates found it difficult to interpret in words what their answer represented; indeed some, often those who had assumed independence, made incorrect reference to *A* and *B* being independent events rather than simply putting into words what their answer represented. In order to score this mark, it was necessary to express that this is 'the probability of *A* and *B*'. It was not enough simply to describe 0.6 as 'the events in *A* and *B*' or 'the common outcomes of *A* and *B*', as some reference to this being 'the probability of…' was also required. In part **(ii)** correct language in the context of probability was important. Examples of correct answers are 'the probability of *A* **or** *B* but not both' or 'the probability of *A* only **or** *B* only' with the word 'or' being important in these descriptions to express the union. Candidates tended to be most successful with part **(iii)** of this question, correctly stating that *C* and *D* are mutually exclusive events.

Answers: **(i)** 0.6

Question 3

There were many fully correct solutions seen to part **(i)** of this question. Occasionally candidates rounded or truncated too early within their working and did not obtain the correct final answers. It was apparent in part **(ii)** that many candidates did not realise the purpose of obtaining scaled marks. Thus incorrect answers suggesting that the practical mark was better were often seen, with reasons such as 'because her raw mark was higher', 'because the mean was higher' or 'because the standard deviation was lower' being given. Some candidates also tried to compare the raw and scaled marks rather than simply stating that she did better in the written test, because it was in that test that her scaled mark was higher. Candidates found part **(iii)** a little more difficult than part **(i)**, although those that began with the correct expression, with two occurrences of the unknown, usually had the algebraic skills required to solve the equation.

Answers: **(i)** 40, 25 **(iii)** 85

Question 4

Many candidates found the correct number of employees required from each age group, and most then went on to find the correct sample in part **(i)**. A few candidates listed the person allocated '38' twice, and some gave a simple random rather than a stratified sample. In part **(ii)** most candidates were able to work backwards from the composition of the representative sample to find the number of men in the original population. There were some excellent, well explained answers given to part **(iii)**. Many candidates suggested that marital status was an important factor to consider, giving as their reason that married people might have children and therefore need to have working hours that fit around them. Others mentioned distance of work from home or mode of transport taken to work as an appropriate factor to consider when stratifying, giving clear reasons as to how this might affect views on working hours. Some candidates gave status within the workplace or the length of time that employees had been working for the company, without providing a justification in terms of a possible impact on their views on the proposal to change working hours.

Answers: **(i)** 15, 38, 64, 29, 04, 70, 47, 55 **(ii)** 50

Question 5

In part **(i)** most candidates used a correct method, 1 – 1/5 – 1/3, to obtain the answer. A few candidates incorrectly did 1 – 1/5 · 1/3. Many candidates approached part **(ii)** by finding the probability that Bashir does not choose chicken on any one occasion and then multiplying it by itself to find the probability of him not choosing chicken on each of two consecutive days. Some tried the alternative method of considering the 'non-chicken' possibilities, but often did not consider them all. Very few candidates correctly stated in part **(iii)** that they had assumed independence when making their calculation, and fewer still were therefore able in part **(iv)** to comment on whether or not this was justified, by considering if choice on one day is likely to be influenced by choice on the previous day. Many candidates struggled to answer parts **(iii)** and **(iv),** simply describing the calculation they had carried out in part **(ii)**.

Answers: **(i)** 7/15 **(ii)** 16/25

Question 6

Many candidates answered part **(i)** correctly although some simply found the mean of the two means. Most candidates struggled with part **(ii)**, often squaring the mean and multiplying by the number of packets, but omitting to use the given standard deviation for each employee in their calculation at all. The most successful candidates usually started by writing out the formula for standard deviation and inserting the known values, before attempting to rearrange the formula for the sum of the squares. In part **(iii)** many candidates attempted to use their answers from parts **(i)** and **(ii)** in a correct formula for the standard deviation, but some did not realise the need to use the previous parts and others used 2 rather than 41 for the total frequency.

Answers: **(i)** 29.0 **(ii)** 16 393, 18 416 **(iii)** 2.8

Section B

Question 7

It was pleasing to see some well-presented solutions to this question. In part **(ii)** there was no guidance provided in the question to suggest what the table should contain, as there has often been in the past. Many candidates, however, did not provide quite enough detail in their answer to part **(i)**. Here candidates needed to explain that a 3-point moving average would be appropriate, and this detail was often missing. Most candidates did, however, explain that the period was odd (without stating the value of this period) and/or that the moving average values would coincide with the timing of original data plots. As previously stated, solutions to part **(ii)** were usually clearly set out, as requested, with each of the seven moving average values corresponding to the correct time slot in a table. A few candidates only provided three of the seven moving average values. In part **(iii)** plots tended to be accurate, with just a few candidates misreading the scale, and sensible trend lines were usually drawn. In part **(iv)** many correct descriptions of the declining trend in visitor numbers were seen. Many candidates demonstrated some understanding of the concept of the seasonal component in part **(v)** by finding an appropriate difference between a number of visitors and a moving average value, either by using values directly from the table or by using readings from their graph. Some candidates, however, only found one such difference rather than finding all three and dividing by three. Those candidates who had shown some understanding of a seasonal component were usually able to use it correctly to make an estimate in part **(vi)**, although, as in the past, some candidates simply took a reading from the trend line without attempting to apply the seasonal component.

Answers: **(ii)** 573, 566, 560.7, 534.3, 512.7, 489.7, 480.7 **(v)** 325 to 345 **(vi)** 745 to 785

Question 8

Correct responses were usually seen to parts **(i)** and **(ii)(a)**. Many candidates in part **(ii)(b)**, however, were only able to interpret partially the figure they had obtained for the weighted aggregate cost index. The most common error was to describe the increase of 12.9% as an increase in the expenditure rather than an increase in the overall cost or prices. Most candidates correctly identified that this increase took place between 2012 and 2014. Most candidates were able to use the weights in the table to find the overall expenditure, given a value for the rent in part **(iii)**. Some candidates did not appear to realise in part **(iv)** that they could use their answers to parts **(ii)** and **(iii)** to obtain the estimate for the overall expenditure in 2014. Many candidates incorrectly referred to prices when trying to come up with a reason why the weights may have changed in part **(v)**. They did not appreciate that any price changes have already been accounted for in the calculation and that it is a change that might have occurred in the quantity (in this case of raw materials) that could account for an inaccurate estimate.

Answers: **(i)** 100, 100, 120, 103 **(ii)(a)** 112.9 **(iii)** $19 950 **(iv)** $22 500

Question 9

This was the least popular question in ***Section B***. Of those who attempted the question, most were able to find all the possible totals in part **(i)**, but few could find all the correct probabilities. The probabilities of winning a total of \$6 and a total of \$2 were most commonly found correctly, but in the case of the other amounts, that could be won in more than one way, the probabilities were often incorrect. It was quite common for the probabilities in the table produced not to add to one. Almost all candidates with some form of distribution used the correct method in part **(ii)**. In parts **(iii)** and **(iv)**, most candidates were able to work out the correct probabilities for choosing two green beads, with and without replacement. However, they were often unsure how to proceed from here. Common wrong attempts in part **(iii)** involved comparing $25/36 \cdot 4$ with 6 and in part **(iv)** a common wrong answer was \$7.50, which came from $2/3 \cdot x = 5$. Some candidates did produce excellent clear solutions.

Answers: **(i)** amounts: \$2, \$3, \$4, \$5, \$6; with probabilities: 4/25, 8/25, 8/25, 4/25, 1/25 **(ii)** \$3.60 **(iii)** should play gold bonus game **(iv)** \$2.50

Question 10

A surprising number of candidates made an error either with the class boundaries or with the class widths in part **(i)**. Most candidates were able to state the class containing the median in part **(ii)**, but some candidates were unable to estimate the median accurately in part **(iii)** or the number of letters with a mass less than 75 g in parts **(v)** and **(vi)**, due to errors with class boundaries and widths. In part **(iv)** candidates were advised in the question to refer to the data in the table when giving their explanation. There were, however, a lot of rather vague references to extreme data. A good response provided a detailed reference to the particular extremes seen in this table, such as the existence of a few very large masses and the impact this will have on the mean compared to the median. In part **(vii)** candidates needed to spot that there was a large difference between the estimated cost of posting the 100 letters, namely \$72, and the actual cost, \$67.50. Most candidates simply made references to rounding errors, which do not explain the size of this difference. Only the most able candidates were able explain that linear interpolation assumes an even distribution of data within an interval, which these figures suggest was not the case in this instance.

Answers: **(i)** 59.5 and 69.5, 10 **(ii)** 70–79 **(iii)** 72.8 **(v)** 60 **(vi)** \$72

Question 11

Most candidates were able to name the charts in part **(i)** as a change chart and a percentage sectional/component/composite bar chart. A good number of candidates were able to extract information from these charts to be able to produce the two-way table and to complete the percentage sectional bar chart in parts **(ii)** and **(iii)**. A few candidates made errors when interpreting the scale of the change chart. In part **(iv)** candidates needed to use both charts to make comments on both the number and the proportion of standard cars. Some candidates made comments about the other categories of car, but most were able to make at least one of the correct statements: that between 2004 and 2014, the number of standard cars had increased and the proportion of standard cars had remained the same. Most candidates were able to produce a correct dual bar chart in part **(v)** with correct labelling and a key given in most cases. The most common error was a missing label on the vertical axis. Few candidates, however, in part **(vi)** were able to give the advantage that a composite bar chart has over a dual bar chart. Commonly seen errors were to say that the composite bar chart is easier to draw or the vague answer that it is easier to interpret. The specific advantage that the totals are clearly displayed was required.

Answers: **(ii)** 65, 45, 15; 60, 54, 36 **(iii)** 40%, 36%, 24% **(v)** 10 and 50, 18 and 36, 24 and 12

STATISTICS

Paper 4040/23
Paper 2

Key message

This examination requires candidates to be able to produce statistical diagrams, to calculate statistics and to interpret findings. The best statistical diagrams should be accurately drawn, taking care to use scales correctly, and have clearly labelled axes and, where appropriate, a key. Candidates scoring the highest marks in the numerical problems will provide clear indications of the methods they have used in logical and clearly presented solutions. In questions that require written definitions, justification of given techniques or interpretation, the most successful responses will include detailed explanations with, where appropriate, clear consideration of the context of the problem.

General comments

As usual, candidates did better on the questions requiring numerical calculations than on those requiring written explanations; in particular, candidates did well in the numerical parts of **Question 3** on scaling, **Question 7(ii)** on moving averages and the early parts of **Question 8** on weighted aggregate cost indexes. They also did well in **Question 4** on sampling and **Question 11(ii)**, which involved interpreting pictorial representations of data. A numerical question that caused difficulty this year was the final part of **Question 6**, on finding the standard deviation when combining sets of data. The final parts of **Question 9**, on expectation, also caused some difficulty. Answers to questions requiring written explanations, such as **Questions 4(iii)**, **7(i)**, **10(iv)** and **11(vi)** were sometimes insufficiently detailed; however in **Question 1(iii)**, for example, there were some good descriptions of the nature of variables and in **Question 7(iv)** clear descriptions of the trend. Other questions that caused difficulty this year were the parts of **Question 2** that required interpretation in words of a probability statement and the final parts of **Question 5** where candidates needed to consider their assumptions when making probability calculations.

Question 9, on probability and expectation, proved to be the least popular of the optional ***Section B*** questions, with **Question 8**, on price relatives and index numbers, **Question 10**, on linear interpolation, and **Question 11**, on pictorial representations of data, proving to be the most popular.

Comments on specific questions

Section A

Question 1

In part **(i)**, many candidates correctly identified that *B*, the number of cars in the car park at 9 am, was not a variable, but it was less common to see *E*, the height of the headteacher's car, also correctly identified. A common error in part **(ii)** was to see *B* rather than *C* given as the answer. In part **(iii)** it was good to see correct use being made of the terms 'continuous data', to describe the height of each car, and 'qualitative data', to describe the colour of each car.

Answers: **(i)** *B* and *E* **(ii)** *C*

Question 2

In part **(i)**, most candidates used the correct formula to find the required probability, although it was not uncommon to see independence being assumed and the probabilities of *A* and *B* simply being multiplied. Many candidates found it difficult to interpret in words what their answer represented; indeed some, often those who had assumed independence, made incorrect reference to *A* and *B* being independent events rather than simply putting into words what their answer represented. In order to score this mark, it was necessary to express that this is 'the probability of *A* and *B*'. It was not enough simply to describe 0.6 as 'the events in *A* and *B*' or 'the common outcomes of *A* and *B*', as some reference to this being 'the probability of…' was also required. In part **(ii)** correct language in the context of probability was important. Examples of correct answers are 'the probability of *A* **or** *B* but not both' or 'the probability of *A* only **or** *B* only' with the word 'or' being important in these descriptions to express the union. Candidates tended to be most successful with part **(iii)** of this question, correctly stating that *C* and *D* are mutually exclusive events.

Answers: **(i)** 0.6

Question 3

There were many fully correct solutions seen to part **(i)** of this question. Occasionally candidates rounded or truncated too early within their working and did not obtain the correct final answers. It was apparent in part **(ii)** that many candidates did not realise the purpose of obtaining scaled marks. Thus incorrect answers suggesting that the practical mark was better were often seen, with reasons such as 'because her raw mark was higher', 'because the mean was higher' or 'because the standard deviation was lower' being given. Some candidates also tried to compare the raw and scaled marks rather than simply stating that she did better in the written test, because it was in that test that her scaled mark was higher. Candidates found part **(iii)** a little more difficult than part **(i)**, although those that began with the correct expression, with two occurrences of the unknown, usually had the algebraic skills required to solve the equation.

Answers: **(i)** 40, 25 **(iii)** 85

Question 4

Many candidates found the correct number of employees required from each age group, and most then went on to find the correct sample in part **(i)**. A few candidates listed the person allocated '38' twice, and some gave a simple random rather than a stratified sample. In part **(ii)** most candidates were able to work backwards from the composition of the representative sample to find the number of men in the original population. There were some excellent, well explained answers given to part **(iii)**. Many candidates suggested that marital status was an important factor to consider, giving as their reason that married people might have children and therefore need to have working hours that fit around them. Others mentioned distance of work from home or mode of transport taken to work as an appropriate factor to consider when stratifying, giving clear reasons as to how this might affect views on working hours. Some candidates gave status within the workplace or the length of time that employees had been working for the company, without providing a justification in terms of a possible impact on their views on the proposal to change working hours.

Answers: **(i)** 15, 38, 64, 29, 04, 70, 47, 55 **(ii)** 50

Question 5

In part **(i)** most candidates used a correct method, 1 – 1/5 – 1/3, to obtain the answer. A few candidates incorrectly did 1 – 1/5 · 1/3. Many candidates approached part **(ii)** by finding the probability that Bashir does not choose chicken on any one occasion and then multiplying it by itself to find the probability of him not choosing chicken on each of two consecutive days. Some tried the alternative method of considering the 'non-chicken' possibilities, but often did not consider them all. Very few candidates correctly stated in part **(iii)** that they had assumed independence when making their calculation, and fewer still were therefore able in part **(iv)** to comment on whether or not this was justified, by considering if choice on one day is likely to be influenced by choice on the previous day. Many candidates struggled to answer parts **(iii)** and **(iv),** simply describing the calculation they had carried out in part **(ii)**.

Answers: **(i)** 7/15 **(ii)** 16/25

Question 6

Many candidates answered part **(i)** correctly although some simply found the mean of the two means. Most candidates struggled with part **(ii)**, often squaring the mean and multiplying by the number of packets, but omitting to use the given standard deviation for each employee in their calculation at all. The most successful candidates usually started by writing out the formula for standard deviation and inserting the known values, before attempting to rearrange the formula for the sum of the squares. In part **(iii)** many candidates attempted to use their answers from parts **(i)** and **(ii)** in a correct formula for the standard deviation, but some did not realise the need to use the previous parts and others used 2 rather than 41 for the total frequency.

Answers: **(i)** 29.0 **(ii)** 16 393, 18 416 **(iii)** 2.8

Section B

Question 7

It was pleasing to see some well-presented solutions to this question. In part **(ii)** there was no guidance provided in the question to suggest what the table should contain, as there has often been in the past. Many candidates, however, did not provide quite enough detail in their answer to part **(i)**. Here candidates needed to explain that a 3-point moving average would be appropriate, and this detail was often missing. Most candidates did, however, explain that the period was odd (without stating the value of this period) and/or that the moving average values would coincide with the timing of original data plots. As previously stated, solutions to part **(ii)** were usually clearly set out, as requested, with each of the seven moving average values corresponding to the correct time slot in a table. A few candidates only provided three of the seven moving average values. In part **(iii)** plots tended to be accurate, with just a few candidates misreading the scale, and sensible trend lines were usually drawn. In part **(iv)** many correct descriptions of the declining trend in visitor numbers were seen. Many candidates demonstrated some understanding of the concept of the seasonal component in part **(v)** by finding an appropriate difference between a number of visitors and a moving average value, either by using values directly from the table or by using readings from their graph. Some candidates, however, only found one such difference rather than finding all three and dividing by three. Those candidates who had shown some understanding of a seasonal component were usually able to use it correctly to make an estimate in part **(vi)**, although, as in the past, some candidates simply took a reading from the trend line without attempting to apply the seasonal component.

Answers: **(ii)** 573, 566, 560.7, 534.3, 512.7, 489.7, 480.7 **(v)** 325 to 345 **(vi)** 745 to 785

Question 8

Correct responses were usually seen to parts **(i)** and **(ii)(a)**. Many candidates in part **(ii)(b)**, however, were only able to interpret partially the figure they had obtained for the weighted aggregate cost index. The most common error was to describe the increase of 12.9% as an increase in the expenditure rather than an increase in the overall cost or prices. Most candidates correctly identified that this increase took place between 2012 and 2014. Most candidates were able to use the weights in the table to find the overall expenditure, given a value for the rent in part **(iii)**. Some candidates did not appear to realise in part **(iv)** that they could use their answers to parts **(ii)** and **(iii)** to obtain the estimate for the overall expenditure in 2014. Many candidates incorrectly referred to prices when trying to come up with a reason why the weights may have changed in part **(v)**. They did not appreciate that any price changes have already been accounted for in the calculation and that it is a change that might have occurred in the quantity (in this case of raw materials) that could account for an inaccurate estimate.

Answers: **(i)** 100, 100, 120, 103 **(ii)(a)** 112.9 **(iii)** $19 950 **(iv)** $22 500

Question 9

This was the least popular question in ***Section B***. Of those who attempted the question, most were able to find all the possible totals in part **(i)**, but few could find all the correct probabilities. The probabilities of winning a total of \$6 and a total of \$2 were most commonly found correctly, but in the case of the other amounts, that could be won in more than one way, the probabilities were often incorrect. It was quite common for the probabilities in the table produced not to add to one. Almost all candidates with some form of distribution used the correct method in part **(ii)**. In parts **(iii)** and **(iv)**, most candidates were able to work out the correct probabilities for choosing two green beads, with and without replacement. However, they were often unsure how to proceed from here. Common wrong attempts in part **(iii)** involved comparing $25/36 \cdot 4$ with 6 and in part **(iv)** a common wrong answer was \$7.50, which came from $2/3 \cdot x = 5$. Some candidates did produce excellent clear solutions.

Answers: **(i)** amounts: \$2, \$3, \$4, \$5, \$6; with probabilities: 4/25, 8/25, 8/25, 4/25, 1/25 **(ii)** \$3.60 **(iii)** should play gold bonus game **(iv)** \$2.50

Question 10

A surprising number of candidates made an error either with the class boundaries or with the class widths in part **(i)**. Most candidates were able to state the class containing the median in part **(ii)**, but some candidates were unable to estimate the median accurately in part **(iii)** or the number of letters with a mass less than 75 g in parts **(v)** and **(vi)**, due to errors with class boundaries and widths. In part **(iv)** candidates were advised in the question to refer to the data in the table when giving their explanation. There were, however, a lot of rather vague references to extreme data. A good response provided a detailed reference to the particular extremes seen in this table, such as the existence of a few very large masses and the impact this will have on the mean compared to the median. In part **(vii)** candidates needed to spot that there was a large difference between the estimated cost of posting the 100 letters, namely \$72, and the actual cost, \$67.50. Most candidates simply made references to rounding errors, which do not explain the size of this difference. Only the most able candidates were able explain that linear interpolation assumes an even distribution of data within an interval, which these figures suggest was not the case in this instance.

Answers: **(i)** 59.5 and 69.5, 10 **(ii)** 70–79 **(iii)** 72.8 **(v)** 60 **(vi)** \$72

Question 11

Most candidates were able to name the charts in part **(i)** as a change chart and a percentage sectional/component/composite bar chart. A good number of candidates were able to extract information from these charts to be able to produce the two-way table and to complete the percentage sectional bar chart in parts **(ii)** and **(iii)**. A few candidates made errors when interpreting the scale of the change chart. In part **(iv)** candidates needed to use both charts to make comments on both the number and the proportion of standard cars. Some candidates made comments about the other categories of car, but most were able to make at least one of the correct statements: that between 2004 and 2014, the number of standard cars had increased and the proportion of standard cars had remained the same. Most candidates were able to produce a correct dual bar chart in part **(v)** with correct labelling and a key given in most cases. The most common error was a missing label on the vertical axis. Few candidates, however, in part **(vi)** were able to give the advantage that a composite bar chart has over a dual bar chart. Commonly seen errors were to say that the composite bar chart is easier to draw or the vague answer that it is easier to interpret. The specific advantage that the totals are clearly displayed was required.

Answers: **(ii)** 65, 45, 15; 60, 54, 36 **(iii)** 40%, 36%, 24% **(v)** 10 and 50, 18 and 36, 24 and 12