STATISTICS

Paper 4040/12

Paper 12

Key Messages

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

Premature rounding or truncation of decimals in the middle of working should be avoided so that accuracy is not lost.

Candidates should develop the skill of holding the intermediate values of a calculation in the calculator to obtain maximum accuracy in the final answer.

Candidates should try to relate their knowledge to the specific requirements of a question rather than simply repeat memorised knowledge.

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.

General Comments

The overall standard of work was comparable to that of last year. Some very good marks were obtained, and there were few exceptionally low marks. As is noted regularly in these reports, there were again instances of marks being needlessly lost due to final answers not being given to the accuracy specifically stated in the question. In those parts of questions requiring comment related to results calculated there is still a tendency for some answers given to be mathematical rather than contextual (see **Question 10** below).

Any candidate 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. For example, if it is found that the mid-day temperature in a city is set to increase by 20°C by mid-century (see **Question 9** below) it should be obvious that a mistake has been made; this is far in excess of even the direst predictions of climate change scientists.

It may seem superfluous to remark that a question should be read carefully before an answer is attempted. Yet there was one question in particular on the paper (see **Question 2** below) where this was apparently not done.

Comments on Specific Questions

Section A

Question 1

Parts **(i)** and **(ii)** were generally answered best. It was clear from answers to the other parts that many candidates do not understand the terms “central tendency” and “dispersion”, for many gave a measure of dispersion when a measure of central tendency was requested, and vice versa. Few answered all parts correctly.

Answers: **(i)** mode **(ii)** range **(iii)** median **(iv)** variance or standard deviation **(v)** interquartile range **(vi)** mode

Question 2

The best answers to part **(ii)** were those which demonstrated that the candidate had read the question carefully, and in particular had understood that the key piece of information given was that there was a proposal to change the time of the class. Thus, when taking her sample, it was important that the instructor did not select, for example, all women who were in full-time employment, and who, presumably, would all have been against the change. The answers given below are not exhaustive; but whatever was suggested, to earn credit it had to be explained to be something that would affect the woman's ability, one way or the other, to attend at the new time.

Weak answers did not address the situation described, but reproduced what was apparently memorised material on avoiding bias *in general*. Thus in spite of the question stating clearly that this was a class for women, and that the instructor already knew their ages, it was quite common to see gender and age suggested for items of data needed.

Answers: **(i)(a)** quota **(i)(b)** systematic **(ii)** employment status, because working women may need to be at work in the afternoon; maternal status, because a woman with children may prefer afternoon attendance when her children are at school

Question 3

This was very well done, with only part **(iv)** causing problems. Success was most readily achieved by those who tried inserting different sets of three consecutive integers into their ordered list in part **(iii)**.

Answers: **(i)** 6 **(ii)** 3.9 **(iii)** 4 **(iv)** 3

Question 4

Whilst parts **(i)** and **(ii)** were almost always answered correctly, there were few fully correct answers to the next three parts. As is observed regularly in these reports, many candidates do not understand clearly what the regions of the different parts of a Venn diagram represent. In parts **(iii)** and **(iv)** common numerators seen were 27 and 6 respectively, and in part **(v)** little appreciation was shown that a denominator of 9 had to be used.

Answers: **(i)** 25 **(ii)** 6 actors have worked in Los Angeles and Rome but not Mumbai **(iii)** 40/48
(iv) 10/48 **(v)** 4/9

Question 5

Parts **(i)** and **(ii)** were almost universally well done. There were also many correct answers to part **(iii)**, but because past questions have usually asked about the radii of the charts, some candidates felt that squaring or taking square roots had to be done somewhere.

Answers: **(i)** $12 million **(ii)** 126° **(iii)** 4 : 3

Question 6

This was another question which was almost universally well done. Candidates understood very clearly this particular form of tabulation for the representation of the distances between different towns. Errors occurred occasionally in part **(ii)** when it was not realised that three distances only had to be added for the journey described in the question.

Answers: **(i)(a)** 35 in cell BC **(i)(b)** 24 in cell AC **(i)(c)** 21 in cell CE **(i)(d)** 37 in cell CD **(ii)** 81 km

Section B

Question 7

As was the case in the examination last year, most candidates were able to apply their knowledge of crude and standardised rates to fertility rates, and there were many good answers to parts **(i)**, **(ii)** and **(iii)**. However, as mentioned in the general comments above, this was yet again one of the questions where marks were sometimes lost through failure to follow the given accuracy instructions.

Good answers to part **(iv)** showed clear understanding that the task was to find the number of deaths in the city, as the number of births was already known from part **(iii)**. They further showed understanding that the calculation had to be based on the total population of the city, and not just the females. It was quite common in weaker answers to see this last point overlooked, with 18 450 being used in the working for deaths instead of 36 900. The least creditworthy attempts simply subtracted one of the death rates from one of the fertility rates and stopped at that point, again failing to appreciate that, whilst fertility rates applied only to the females, death rates applied to the whole population.

Very good general understanding of what was required was shown in part **(v)**.

Answers: **(i)** 88.7 **(ii)** 145, 828, 714, 87 **(iii)** 96.2 **(iv)** 1486 **(v)** migration of people into or out of the city

Question 8

There were many correct answers to part **(i)**, though not all candidates appreciated that this was a 'without replacement' situation. Most did not see the simple link between this part and the next, and attempted part **(ii)** as though it was completely unrelated to what had gone before. Unfortunately, in the analysis of the different cases this involved, one of the three possibilities was frequently omitted.

The quality of answers to the histogram was mixed, with many fully correct answers, but also many where no allowance was made for the different widths of the rectangles.

Whilst the number of fully correct answers to part **(vii)** was limited, a good number of candidates were able to obtain some marks on the question. The best answers showed clear understanding of the conditional element, ending with a division of probabilities, even though these might not be individually correct, it being sometimes thought that there were just three 3, 4, 5 cases. More limited answers finished at the point where the probability of the apartments having 12 rooms had been found, the conditional element not being recognised. A significant number of answers was seen in which it was thought that the only requirement was to find the probability of choosing three apartments each with 4 rooms. It should have been apparent that a question worth 6 marks must have involved more than one line of working for its solution.

Answers: **(i)** 35/204 **(ii)** 169/204 **(iii)** 54 **(iv)** 6 **(v)** rectangle of height 5 **(vi)** modal class **(vii)** 13/157

Question 9

Some candidates produced graphs of very high quality, the majority plotting points correctly. But the error of using mid-class values instead of upper class boundaries continues to be seen too often.

As has been pointed out before in these reports, good answers to this type of question give some indication on the graph (for example with lines drawn and labelled) of how the required information is being found. Credit can then be given for method, even if the answer is incorrect. Some progress appears to have been made in this respect, with, on this occasion, fewer graphs devoid of annotations than has been the case in the past.

Part **(iii)** was reasonably well done, although a significant number of answers was seen where the serious error of using a total frequency of 400 was made. Common errors in part **(iv)** were to add 2.5°C or even 20°C to the median previously found, and also to add a temperature increase to the interquartile range previously found. In the case where 20°C was being added, it should have been realised that this was a highly unrealistic increase.

In part **(v)**, thought processes were not always evident from answers presented. The best solutions were those where vertical lines were drawn on the graph at temperatures of 36°C and 34°C, with horizontal lines linking these to the respective cumulative frequencies.

Answers: **(i)** 8, 33, 85, 166, 245, 313, 350, 365 **(ii)** plot of cumulative frequencies at upper class boundaries joined by a smooth curve **(iii)(a)** 20.7°C to 21.3°C **(iii)(b)** 11°C to 12°C, dependent on correct method for, and accuracy of, quartiles **(iv)(a)** answer to part (iii)(a) + 2°C
(iv)(b) same answer as part (iii)(b) **(v)** 9, 10 or 11 days

Question 10

Following an observation made in this report last year on the clarity of plotted points, this year, almost always, Examiners were able to see points very clearly.

Very good marks were generally earned on the first three parts, with good understanding shown of the need to order data to find the semi-averages. By far the best way to proceed in part **(iv)** was to use the two given averages to find the equation of the line. Candidates who used the average they had calculated in part **(iii)** risked error by using values they could not be certain were correct, unlike the values for the other averages given in the question. Unfortunately many did exactly this, and as a consequence of working with their own (incorrect) average obtained an incorrect equation. Incorrect equations also resulted from working with a gradient accurate to only one significant figure.

In part **(v)** quite a lot of answers were written in purely mathematical language, when what was required was an appreciation of what was implied for the schools and teachers.

Reasonable skill was shown in part **(vi)** in drawing a line of best fit by eye, and in part **(vii)** in finding its equation. For the latter it was essential that points from the line drawn had to be used. When values were seen which were originally given in the table, Examiners only gave credit if the line drawn passed through the plot of these particular points.

In part **(viii)**, most candidates knew that this had something to do with educational provision as it related to the number of teachers employed. But a good number focused on the intercepts of the two equations rather than the gradients. Statements to the effect that Belport was better because it employed more teachers could not be accepted, as actual numbers for Belport were unknown.

Answers: **(ii)** (927+1085+1219+1361)/4 **(iii)** (559.75, 25.75) **(iv)** m = 0.0280 or 0.028, c = 10.00 to 10.11 **(v)** it indicates there are 10 teachers when there are no pupils **(vii)** m = 0.033 to 0.039, c = intercept of line drawn in part (vi) **(viii)** Belport, as gradient for Belport is higher, showing that the number of teachers per pupil there is higher than at Astra

Question 11

The answers below for part **(i)** are not exhaustive, but to gain credit specific advantages and disadvantages in the statistical analysis of data had to be provided. Thus references to a process being tedious or taking a lot of time were not considered acceptable. Also, what appear to be common assumptions about it being 'easier' to analyse a frequency distribution rather than a large set of data must be questioned; if a large set of data is held in a spreadsheet a wide range of statistical measures can be found almost instantaneously.

Part **(ii)** was generally well answered, although a mark was commonly lost on the standard deviation through failure to maintain sufficient accuracy in decimals in the body of the working. For such a problem candidates should have the ability to retain intermediate values of maximum accuracy within the calculator, by making use of the memory. Too often premature rounding or truncation of decimals is seen. Most used the method for standard deviation based on Σfx and Σfx^2, which is far better for computational purposes than that which uses $\Sigma f(x - \text{mean})^2$.

Part **(iii)** aimed to test if candidates were able to focus on the particular numbers relevant to a question, when given a table containing a range of information. There were very mixed answers, with some giving more than one programme for one or both answers.

Good understanding was shown in part **(iv)**, and many clearly presented answers were seen.

Answers: **(i)** provides a concise summary of the data; original data are lost **(ii)** 3.66, 0.343 **(iii)(a)** Q **(iii)(b)** T **(iv)** 197/900

STATISTICS

Paper 4040/13

Paper 13

Key Messages

A valuable skill in statistical work is to be able to recognise when the results of a calculation or analytical process are reasonable.

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

If words in a question are emphasised they should be noted carefully by the candidate so that unnecessary errors are avoided.

General Comments

The overall standard of work was comparable to that of last year, with a wide range of marks being obtained. As is noted regularly in these reports, there were again instances of marks being needlessly lost when answers were not given to the required accuracy, where this was stated in the question (see **Questions 2**, **10** below).

A candidate of statistics ought to know whether or not the result of a calculation or analytical process is reasonable in a given practical situation. If it is clearly unreasonable, the work can be checked to find the error and the error corrected. If a plot of the values on a scatter diagram show clearly that as x increases y decreases, it ought to be obvious that, if found, a line of best fit with positive gradient must be wrong (see **Question 10** below).

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 (see **Question 6** below).

Comments on Specific Questions

Section A

Question 1

Answers to this question were mixed. It is clear that some candidates do not understand the terms “central tendency” and “dispersion”, for a measure of dispersion was sometimes given when a measure of central tendency was requested, and vice versa.

Answers: **(i)** median, mode **(ii)** interquartile range **(iii)** mean **(iv)** two from range, standard deviation, variance

Question 2

This was very well answered, with many candidates obtaining full marks. Good understanding was shown of the use of the square of the radius in part **(iv)**, though occasionally a mark was needlessly lost as a consequence of the accuracy instruction being ignored.

Answers: **(i)** Europe 164°, Asia 74°, North America 90°, Rest of the World 32° **(ii)** \$162 million
(iii) 4.9 cm to 5.1 cm **(iv)** 4.1 cm

Question 3

This was another very well done question, with many full mark answers being presented.

Answers: **(i)** and **(ii)** column totals: 220, 440, 660; 540, 125, 665; 1390, 785, 2175; 2150, 1350, 3500

Question 4

Where errors occurred they were mainly in part **(iii)**, where the value for the total number of handball players was occasionally used instead of the value for those who play only handball.

Answers: **(i)** 1; one girl did not play any of the three sports **(ii)(a)** 7 **(ii)(b)** 2 **(iii)** 19

Question 5

This question and the next were by far the least well answered in ***Section A***. Whilst almost all recognised the need for rectangle heights to correspond to frequency densities, many errors were made in using the one given height to deduce correctly the standard class width.

Answers: **(i)** 40 **(ii)** 7, 32, 4, 2 **(iii)** 2.71

Question 6

It was clear that most candidates knew about systematic sampling, and there was scarcely any confusion with other types of sampling. But in part **(a)(i)** there was a tendency to give examples of biased outcomes rather than the causes of such outcomes. Answers to part **(a)(ii)** were reasonable, though rarely complete, either the first or second steps (or even both) in the process being omitted. In part **(b)**, stratification was clearly understood, but only the strongest answers gave stratification directly relevant to the surveys being carried out. Weaker answers offered criteria which might be employed in general, such as gender, age or occupation.

Answers: **(a)(i)** occurs when there is a regular pattern in the population listing **(a)(ii)** three basic steps to be given: listing the population; starting the selection at a random point; selecting every 19th candidate from the list after the starting point **(b)(i)** into smokers and non-smokers **(b)(ii)** into those who live near an airport and those who do not

Section B

Question 7

In this question, part **(b)** was answered far better than the other two parts. The diagram was well understood and there were many correct answers.

In part **(a)** not everyone appreciated that the case of the person not having the disease had to be considered as well as the case of the person having the disease, and furthermore that the test result had to be negative in the former case to give the correct result. Nevertheless some correct solutions were seen.

But there were very few correct solutions to part **(c)(i)**. Almost all failed to consider in their working that if Laura went into exactly one shop it meant that she did not go into the other. Consequently 0.8 and 0.3 were usually absent from the working. In part **(c)(ii)** some candidates did not seem to recognise the numerical comparison which had to be made in order to give a decision.

Answers: **(a)(i)** 0.05 **(a)(ii)** 0.1, 0.9 in second column **(a)(iii)** 0.9075 **(b)(i)** 13/33 **(b)(ii)** 4/5 **(b)(iii)** 4/13 **(c)(i)** 0.0558 **(c)(ii)** unlikely as 0.0558 > 0.05

Question 8

There were very few completely correct answers to part **(a)** because of the graphs presented in part **(a)(ii)**. Candidates do not seem to have observed the emphasis given to the word “appropriate”, because almost all produced a totally inappropriate graph. As the variable is discrete, full credit could only be given where a step polygon was drawn.

The first five parts of part **(b)** were generally well answered, though with occasional errors through the misreading of scales. Good appreciation was shown in part **(b)(vii)** that there would be no change, but a mark was frequently dropped in part **(b)(vi)** because the "5 minutes" given in the question was absent from the answer.

Answers: **(a)(ii)** step polygon required **(b)(i)** 42 **(b)(ii)** 35 **(b)(iii)** 55 to 56 **(b)(iv)** 180 **(b)(v)** 6th or 7th **(b)(vi)** increased by 5 minutes **(b)(vii)** unchanged

Question 9

The calculation of crude and standardised death rates is well known by most candidates, and there were many good answers to the first three parts.

The explanatory parts were less well done. In part **(iv)**, few focused on the population age structures, and in part **(v)** it was usual to see only the first of the reasons given below, though credit was also given for the observation that town B must have the healthier environment. In part **(vi)** there was widespread recognition that the rate would not change, but incomplete explanation as to why this was so.

Answers: **(i)** $p = 9$, $q = 40$ **(ii)** 4.2 per thousand **(iii)** 7.3 per thousand **(iv)** the proportions of the population of town A in the different age groups match exactly the proportions of the standard population in the different age groups **(v)** town B has a larger population than town A; town B has a much smaller group death rate amongst the elderly than town A **(vi)** value unchanged; CDR is calculated using only total population and total deaths, and both would be unchanged

Question 10

A good proportion of candidates answered the first four parts well, with accurately plotted points and accurately calculated averages, leading to a good line of best fit. But for others the fact that y decreased as x increased resulted in a common error, it being assumed that the smallest values of x always had to be paired with the smallest values of y, when calculating the semi-averages. This error meant that the location of the plotted averages on the grid, and the line subsequently drawn through them, bore no relationship whatsoever to the pattern of the plotted data. The line had a positive gradient when clearly the trend of the data indicated the gradient should be negative. When this happened the candidate ought to have realised something was wrong and paused for reflection, instead of continuing regardless.

In part **(v)** the accuracy instruction was sometimes ignored.

The best answers in part **(vii)** were those which illustrated the dangers of extrapolation with contextual examples, commenting on the likely performance in this situation of very young children or elderly people.

Answers: **(ii)** overall (10.7, 18.7); lower (8, 23.7); upper (13.3, 13.7) **(iv)** gradient: value rounding to –1.9; intercept: value rounding to 39 **(v)** 12 minutes **(vi)(a)** reasonably well **(vi)(b)** A **(vii)** would not be valid for substantial extrapolation; for example, the line of best fit indicates an impossible time of zero for someone who is about 20 years old

Question 11

The quality of answers to this question was variable. Even though basic computation of mean and standard deviation was required, marks were routinely lost. Sometimes this was the result of calculation errors, sometimes the result of using incorrect formulae.

In part **(iv)**, as emphasised in the question, the results from part **(iii)** had to be used. Few candidates were able to do this successfully. The few good answers seen used the 250 and 750 appropriately and obtained the required values quickly and easily. Unsatisfactory answers went back to the original x values and started again.

Answers: **(i)** –1250, 750, 2000, 3750, 7500 **(ii)** –8, 0, 5, 12, 27 **(iii)** 5.02, 85.9896 **(iv)(a)** 2005 **(iv)(b)** 5 374 350 **(v)** dollars squared

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 and graphical work than on those requiring written explanations; in particular, candidates did well on the numerical and graphical parts of **Questions 1**, **5**, **7** and **10**. Answers provided to questions requiring written explanations, such as **Questions 7(a)(i)**, **10(ii)(d)** and **11(iv)(d)**, were sometimes too vague. Where candidates needed to provide some interpretation of their calculated statistics, such as in comparing the interquartile ranges in **Question 9(b)(iv)**, some otherwise strong candidates seemed to struggle.

Question 8, on probability, 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

The majority of candidates were able to apply correctly the laws of probability relating to independent and mutually exclusive events. The most common errors were for candidates simply to add the probabilities of A and B in part **(i)(b)**, without subtracting the intersection, and to multiply the probabilities of C and D in part **(ii)(a)**.

Answers: **(i)(a)** 0.03 **(i)(b)** 0.32 **(ii)(a)** 0 **(ii)(b)** 0.64

Question 2

In part **(i)** of this question, a new value was being added to a set of data and candidates were asked to explain the effect on the mean and the standard deviation. Many candidates stated, incorrectly, that the mean would increase and that the standard deviation would stay the same. Such candidates had confused the idea of adding a constant to each data item, rather than adding a single value to the set of data items. In part **(ii)** the concept being tested was the effect on the mean and standard deviation of adding to each item a constant and of multiplying each item by a constant. Some candidates, incorrectly, assumed that the addition of the bonus would affect the standard deviation.

Answers: **(i)** Stay the same, decrease **(ii)** 12800, 1050

Question 3

There were some good attempts at this question, with many candidates producing well organised solutions. Some candidates got incorrect probabilities, but were nonetheless able to use expected values to decide whether or not the game was fair. A few candidates, incorrectly, attempted to compare probabilities, rather than expected values.

Answers: **(i)** ¼, ¾, fair game **(ii)** $3

Question 4

Many candidates struggled to deal with the times in this question. It was necessary to find the mean number of minutes early/late for the two groups of candidates before trying to combine them. In part **(ii)** many candidates were able to quote the correct formula for standard deviation, but again they frequently used times rather than the number of minutes late in this formula.

Answers: **(i)** –36, 8.59 **(ii)** 11.9

Question 5

Most candidates were able to use the change chart, together with the figures provided, to calculate the quantities of the commodities produced in 2012. They then, usually successfully, displayed this information in the form of a dual bar chart. A mark was lost by some candidates for insufficient labelling of the vertical axis, where it was necessary to state that the units were 'millions of tonnes'. In part **(iii)** some candidates did not explain sufficiently clearly that the advantage of a dual bar chart over a change chart is that the original data is not lost.

Answers: **(i)** 80.7, 96.8, 22.1, 17.7

Question 6

Most candidates correctly identified the heights of the players as continuous, quantitative data and the towns of birth of the players as discrete, qualitative data. In part **(b)**, the majority of candidates were able to identify the chart correctly as a sectional, component or composite bar chart, but many did not recognise that this chart was more appropriate than a histogram, as the data presented here is discrete. Many candidates simply stated that the sectional bar chart was easier to understand than a histogram. In part **(b)(iii)**, it was common to see the answer given as simply the number of matches played in the cup in which the team scored 2 or more goals, rather than this expressed as a fraction of the total number of matches played in the cup. The denominator of 11 was frequently incorrect or missing entirely.

Answers: **(b)(iii)** 3/11

Section B

Question 7

In part **(a)(i)**, it was necessary for candidates to consider the merits of obtaining moving average values in this particular situation. Therefore they needed to consider whether the number of visitors at a tourist attraction is likely to be subject to seasonal variation, and to conclude that this is likely. Many candidates simply stated, in general terms, the purpose of calculating moving average values, without relating their comments to the particular situation identified. Parts **(a)(ii)** and **(iii)** were completed correctly by many candidates with a few, incorrectly, giving an answer of 3 for part **(ii)**.

The calculations in parts **(b)(i)** and **(ii)** were completed correctly by most candidates and the graph plots in part **(iii)** were mostly accurate, with a suitable trend line drawn. Most candidates correctly interpreted the trend line in the context of the problem presented. In part **(v)**, some candidates did not take the reading from the trend line at the correct place and others did not subtract 11.25 from their reading. The most common error, however, was not to give the final estimate of the number of patients admitted to the hospital as a whole number.

Answers: **(a)(ii)** 4 **(b)(i)** 168, 1308, 218 **(b)(ii)** 213.5, 215, 215.5, 217, 219.25, 221.25

Question 8

Most candidates were successful with part **(i)** of this question, although it was surprisingly common to see incorrect responses of 60/100 and 40/130 for parts **(a)** and **(d)**, respectively. There were many fully correct solutions seen to part **(ii)**, with some errors caused by some candidates unnecessarily trying to consider the males and females separately and omitting some of the possible combinations. Part **(iii)** was more challenging, but some good attempts were seen with the most common error being multiplication by 3 instead of 6. Most candidates found part **(v)** the most challenging, although some fully correct solutions were seen. Some candidates were unable to attempt the final part of this question and a common incorrect answer seen in part **(v)(a)** was 4/13.

Answers: **(i)(a)** 1/5 **(i)(b)** 8/15 **(i)(c)** 23/30 **(i)(d)** 2/7 **(ii)** 105/299 **(iii)** 700/3427 **(iv)** 4, 3, 6
(v)(a) 40/121 **(v)(b)** 34/121

Question 9

Candidates found part **(a)**, and in particular part **(a)(i)**, of this question difficult. A common error seen in part **(a)(i)** was for the true class limits to be given as 50 and 60. Candidates who were successful with parts **(a)(ii)** and **(iii)** usually went on to complete the numerical parts of **(b)** correctly.

In part **(b)(i)**, many candidates correctly chose the median and explained that it is not affected by extreme values. Almost all candidates found the correct cumulative frequencies in **(b)(ii)** and most then tried to find the 25^{th} and 75^{th} values, as required for the interquartile range in **(b)(iii)**. Most candidates then applied the correct formula, but many used wrong values for class boundaries and class widths. In part **(b)(iv)**, it was necessary for candidates to interpret this value. Many thought that a smaller interquartile range indicated smaller masses in general, rather than a smaller dispersion of the masses. Again in part **(b)(v)**, incorrect identification of class boundaries led to wrong answers for some who used a correct approach. Common wrong methods involved trying to divide the whole population proportionately, rather than just the 450–750 group.

Answers: **(a)(i)** 50, 61 **(a)(ii)** 49.5, 60.5 **(a)(iii)** 45, 65 **(b)(ii)** 12, 43, 72, 86, 94, 98, 100 **(b)(iii)** 480
(b)(v) 62.3

Question 10

Most candidates demonstrated a good understanding of price relatives in their answers to this question. The numerical parts of **(i)**, namely parts **(a)** and **(c)**, were usually correct. In part **(i)(b)**, candidates needed to explain that the price relative of 110 for equipment indicates that the price or cost has increased by 10% between 2010 and 2012. A few stated incorrectly that it indicated that the expenditure had increased by 10%. In part **(i)(d)**, most candidates correctly drew a two-way table, with values of 100 for each category in 2010 and the price relatives that they had calculated for 2012.

Again it was the numerical parts of **(ii)**, **(b)** and **(c)**, which candidates found the most straight-forward, and many fully correct solutions were seen. In part **(a)**, it was necessary to explain that the ratio of the expenditure on the three different categories could be used to calculate the weights. In part **(d)**, candidates needed to consider the reliability of the result they had achieved and also what might contribute to an unreliable result. In the particular context provided, these reasons might have been that the number of employees, or number of hours worked, had changed or that the amount of equipment used had changed. These features had not been considered within the calculation of the weighted aggregate cost index. Some candidates gave incorrect answers, such as that there might have been inflation; this is a feature included within the figures for the price relatives, and thus not a potential source of inaccuracy. Other answers which did not gain credit were those which were too vague and did not relate specifically to the problem presented, such as simply that the weights may be incorrect. It was necessary to provide a reason as to why this might be the case.

Answers: **(i)(a)** 107 **(i)(c)** 98 **(ii)(b)** 104 **(ii)(c)** 5990

Question 11

In part **(i)**, candidates needed to consider the reliability of the sampling technique described. They needed to think about the specific situation and consider who might be waiting for a bus at 7 am on a Monday morning. In the best answers, the candidates described the method as not representative of the whole population, because it was likely to contain a group of people such as workers or college candidates who are likely to have similar requirements in terms of the buses they want to catch.

In part **(ii)**, candidates needed to explain the need to number the population list from either 00 to 59 or 01 to 60 before selecting a simple random sample. Some candidates, who described writing names on pieces of paper and drawing them from a hat, did not appear to have read the question carefully enough. There was a high proportion of correct answers to **(ii)(b)**, with the most common errors being the inclusion of both 00 and 60 or the inclusion of 15 twice.

Many candidates were able to find the correct systematic sample in part **(iii)** although, as with the previous part, some candidates did not read the wording of part **(a)** sufficiently clearly and described the process for selecting the whole sample rather than giving a sufficiently detailed description of the selection of the first term. It was necessary to state that the first term came from a number between 00 and 09 (or between 01 and 10) randomly selected from the table.

In the final sampling method, stratified sampling, the numbering of the people within the age groups needed to be considered. Some candidates appeared to be using the ages themselves for the numbering of the groups, and thus it was common to see 07 excluded and 82 and 16 included in the stratified sample. The correct numbering that should have been to give each of the age groups in the question was 00 to 19, 20 to 49 and 50 to 59 (or 01 to 20, 21 to 50 and 51 to 60) respectively. In part **(iv)(d)**, it was necessary to explain the benefit of a sample stratified by age group when considering the views of the population to the proposed bus timetable change. Thus it was necessary to consider the relevance of age to this particular problem. In the best answers, candidates explained that the different age groups are likely to want buses at different times. Most candidates made a general comment about stratified samples not being biased, which did not relate specifically to the problem that had been presented. Only the very strongest candidates scored this final mark on the paper.

Answers: **(ii)(b)** 15, 08, 00, 31, 52, 47 or 15, 08, 60, 31, 52, 47 **(iii)(b)** 04, 14, 24, 34, 44, 54
(iv)(a) 2, 3, 1 **(iv)(c)** 17, 55, 25, 07, 35, 42

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 the numerical parts of **Questions 4**, **6**, **7**, **8** and, for those who attempted it, **Question 9**. It was particularly pleasing to see, in **Question 9(v)**, clearly laid out logical solutions, as these were essential in this 8-mark question. Answers to questions requiring written explanations, such as **Questions 5(i)** and **10(a)(ii)**, were sometimes too vague.

In **Questions 6(ii)**, **7(iv)(b)**, **11(b)(iii)** and **11(b)(iv)** it was necessary for candidates to provide some interpretation of their calculated statistics and graphs. In **Questions 6(ii)** and **11(b)(iv)**, some otherwise strong candidates seemed to struggle to interpret and explain their results.

Question 9, on probability and expectation, proved to be the least popular of the optional ***Section B*** questions, although those that attempted it generally scored high marks. **Question 7** on linear interpolation proved to be the most popular of the optional questions.

Comments on Specific Questions

Section A

Question 1

The majority of candidates were able to correctly identify 'the number of items of mail' as a discrete variable and 'the distance run by a number of athletes during 1 hour' as a continuous variable, although there were slightly more errors in part **(ii)** than part **(i)**. In parts **(iii)** and **(iv)**, it was necessary to give true lower and upper class limits for each of these variables, with candidates often being more successful in identifying the lower than the upper class limit.

Answers: **(i)** Discrete **(ii)** Continuous **(iii)** 5 and 9 **(iv)** 4.5 and 9.5

Question 2

There were some good attempts by many candidates in part **(i)** to explain the term 'base year'. Many scored one of the two available marks, by describing the base year as a reference point or the point in time to which other index numbers are referred. The second mark was for stating that it is when an index number takes the value 100, and it was less common to see this part of the answer given.

In part **(ii)**, many candidates found it difficult to express the meaning of the term 'weight' in this context. A few candidates were able to say that weights were used to calculate a weighted cost index, but not that it is a means of taking into account the different levels of importance of items in an index number. For the second mark, it was necessary to state that the weight is usually based on the expenditure on each item and this was rarely seen.

As with part **(i)**, it was common for candidates to score one mark in part **(iii)**. It was necessary to describe the term 'price relative' as the ratio of two prices, or as showing the proportional or percentage change in the price of an item, for the first mark and then, for the second mark, to state that this is relative to the base year.

Question 3

It was rare to see the correct answer to part **(i)**, that the samples of male and female mice are of different sizes. Some candidates, incorrectly, stated that the division should be by 45 rather than 2, because there are 45 mice.

In part **(ii)**, errors in the calculation of the mean included candidates multiplying the sums of lengths by the respective frequencies and use of the incorrect formula from part **(i)**. In the calculation of the standard deviation, some candidates used their mean correct to only one decimal place, when a greater degree of accuracy was required in order to find the standard deviation correct to one decimal place.

Answers: **(ii)** 164.1, 10.2

Question 4

The majority of candidates were successful with this question, particularly parts **(i)** and **(ii)**. In part **(iii)**, it was quite common to see an incorrect initial equation, often with only one occurrence of the unknown, rather than two.

Answers: **(i)** 136 **(ii)** 30.75 **(iii)** –26

Question 5

Correct items of detail missing from the bar chart for part **(i)** are the key/legend, the vertical scale, the label on the vertical axis and the title. Some candidates gave answers which were too vague, such as 'labels' or 'axes'. The vast majority of candidates scored at least one of the two marks available for this part of the question.

In part **(ii)**, candidates needed to describe the bar chart as either Sectional, Component or Composite. The most common error was to see it described as a 'stacked bar chart'.

Most candidates correctly noted, in part **(iii)**, that had the bar chart illustrated percentages then the bars would have been of equal heights.

In part **(iv)**, many candidates correctly stated the disadvantage of using pictograms to illustrate frequencies as the difficulty to determine the exact frequency when partial pictures are used. Some however incorrectly referred to the difficulty of construction of a pictogram, rather than considering the impact of the diagram once constructed.

In part **(v)**, most candidates correctly identified that nothing had changed and therefore a change chart could not be used.

Question 6

The most common error in part **(i)** was to see the weighted cost index given as the final answer, rather than the percentage increase. There were however many fully correct solutions and the majority of candidates scored at least three of the four available marks.

There were more difficulties, however, with part **(ii)**, with surprisingly few candidates appreciating the need to compare 5.56% with 7% in order to establish that the farm was more profitable than it had been, because the income had increased by more than the costs.

Answers: **(i)** 5.56%

Section B

Question 7

Parts **(i)**, **(ii)**, **(iii)** and **(iv)(a)** of this question were completed successfully by the majority of candidates.

Some candidates did not appear to understand what they were being asked to do in part **(iv)(b)**. Many did not find the distances of the quartiles from the median. Those that did were often able to state correctly that these distances were approximately equal, and that this is what would be expected in a distribution of heights of adult males.

In part **(v)**, some candidates were able to identify correctly that the curve would be steepest in the middle of the graph, where the frequency is greatest or where the greatest change in the cumulative frequency occurs. The most common error was to identify the part of the graph where the greatest change in frequency occurs.

Answers: **(i)** 144, 1376, 3589, 6148, 7857, 8562, 8585 **(ii)(a)** 170 – under 175 class **(ii)(b)** 171.4
(iii)(a) 165 – under 170 class **(iii)(b)** 166.7 **(iv)(a)** 9.2

Question 8

Most candidates were successful with part **(i)** of this question, realising the need to include the factor of ½ (the probability of selecting each bag) within their calculations.

In part **(ii)**, some candidates did not realise that the selection of two balls implies that they will not be replaced and hence the most common error was to see denominators of 11 and 11 in the two two-factor products, rather than 11 and 10.

Parts **(iii)**, **(iv)** and **(v)** were usually correctly calculated by those who attempted them, although by part **(v)** some candidates were leaving this question blank.

In order to answer part **(vi)** of this question, it was necessary to notice that the probability that the first ball will be white is the same in each situation. Thus the comparison can focus on the fact that, with fewer balls in X than Z, the probability that the second ball being white is greater in part **(iv)** than part **(v)**.

Answers: **(i)** 5/11 **(ii)** 31/55 **(iii)** 21/121 **(iv)** 3/55 **(v)** 2/77

Question 9

Most candidates were able to correctly find the probability of getting three 1s in part **(i)**.

In part **(ii)**, the most common error was for candidates to calculate the probability of any three numbers except 1, rather than three of the same number except 1.

In part **(iii)**, some candidates did not consider the number of ways in which exactly two 1s and some other number could be achieved, however the majority correctly multiplied by a factor of three.

Most candidates correctly calculated the prize multiplied by the probability for each outcome and summed their results in part **(iv)**. Some candidates, however, left this as their final answer or subtracted the entry fee of $1 from this amount, rather than considering the profit of the organiser and performing the subtraction the other way around.

Clearly set out working was essential in part **(v)**. Most candidates achieved this, with the most common error being that some did not appreciate that the essential difference between this game and the previous one was that the selected cards are not replaced. Thus, for example, denominators of 10, 10 and 10 were sometimes seen instead of 10, 9 and 8. Such candidates did, however, usually appreciate the fact that the only possible ways of getting three cards the same are three 1s or three 2s. Another common error was for the factor of 3 to be missing from the probability of exactly two 1 s and some other number. The method for the calculation of expectation was usually correct. Those candidates that attempted this part of the question were usually able to achieve at least 3 of the available marks, with some scoring all 8 marks. Some candidates, however, did not attempt this part of the question.

Answers: **(i)** 1/216 **(ii)** 5/216 **(iii)** 5/72 **(iv)** 67 cents **(v)** 61 cents

Question 10

Most candidates correctly named the sampling method as quota sampling in part **(a)(i)**.

Answers to part **(a)(ii)** were often given in terms that were too general, such as to state simply that using a list of registered voters would be biased, without providing a reason. Candidates needed to state that not all inhabitants would appear in a list of voters.

Common incorrect answers of 17 and 1 were seen in parts **(a)(iii)(a)** and **(a)(iii)(b)** respectively, where candidates had ignored the second and third visits to the properties.

In part **(a)(iv)**, table headings were often incorrect with dwelling numbers 1 to 12, rather than number of children 0 to 7, appearing as one of the row/column headings. Dwelling types D, S and B usually correctly appeared as the other headings.

In part **(a)(v)**, candidates needed to use the fact that the bigger the sample, the more accurate it is likely to be. Some candidates incorrectly chose semi-detached houses over detached houses, because they said that there were too many children in the detached houses.

In part **(b)**, the simple random and stratified samples were usually correct. In part **(b)(ii)**, some candidates were picking values from the random number table at regular intervals, rather than selecting just the first value for the systematic sample from the random number table and then selecting every tenth person.

Answers: **(a)(iii)(a)** 23 **(a)(iii)(b)** 3 **(b)(i)** 21, 32, 07, 42, 57, 59 **(b)(ii)** 07, 17, 27, 37, 47, 57
(b)(iii) 21, 32, 07, 42, 57, 17

Question 11

Correct responses to part **(a)(i)(a)** described the trend as the long-term pattern after regular variations have been removed. The key to scoring the mark was to describe the increase/decrease as 'long-term', 'general' or 'over-time'.

In part **(a)(i)(b)**, some candidates described seasonal components rather than seasonal variation. The key here was to include the idea of variation that repeats itself, such as describing a regular variation over a fixed (relatively short) time period.

The majority of candidates were unable to explain the meaning of cyclic variation in part **(a)(i)(c)**. The key here was to describe long-term variation following a general pattern, but over variable lengths of time.

In part **(a)(ii)**, a common error was to state that the trend, rather than the seasonal variation, is removed from a time series when moving averages are calculated. Candidates were rarely able to explain that this is done by smoothing out the variations over one time period.

The calculations in part **(b)(i)** were usually all correct, with accurate plots in part **(b)(ii)**.

In part **(b)(iii)**, most candidates correctly stated that the sales rose initially, but then declined thereafter.

Many candidates did not see that a single trend line was not appropriate in part **(b)(iv)**. Those that did recognise this were not always able to express the reason as being because the trend in the early quarters was very different from that in the later part of the period.

Suitable trend lines, ignoring the plots before 2007, were drawn by many candidates for part **(b)(v)**. Some candidates incorrectly drew lines that included the early plots and others ignored the instruction to draw a straight line and drew curves.

In part **(b)(vi)**, most candidates correctly identified that the seasonal component for quarter one was negative and therefore sales are likely to be smaller than indicated by the trend line.

Answers: **(b)(i)** $x = 105$, $y = 205$, $z = 22.5$