STATISTICS

Paper 4040/12
Paper 1

Key messages

It is very important to read a question carefully to take note of any key statistical terms that it might contain.

It is as just as important to be able to explain and interpret the results of calculations as it is to be able to perform them accurately. Statistical analysis only has any real value when it is applied to a real-life practical situation.

When commenting on a practical situation following statistical analysis, use should be made of the results of the analysis, and of any specific information given in the question.

General comments

The overall standard of work involving calculations of a routine nature was good. This was particularly true of the topics of statistical measures, crude rates, line of best fit and pie charts. Answers were generally much more limited on the explanation and interpretation parts of questions. Such answers were sometimes speculative, without specific reference to the data of the question (see **Questions 2(c)**, **10(g)** below).

There was evidence on this paper of candidates apparently not appreciating the significance of statistical terms in the question, or even ignoring them altogether (see **Questions 6(c)**, **11(e)** below).

The need for students of this subject to appreciate the practical nature of Statistics in its application to a wide variety of real-life situations must again be emphasised. It is certainly important for a candidate to be able to perform routine calculations efficiently and accurately; but it is just as important to understand what the results of such calculations demonstrate, and their implications when applied to the practical context of the question.

Comments on specific questions

Question 1

Almost all candidates obtained correct answers.

Question 2

Almost all candidates obtained correct answers to **parts (a)** and **(b)**. There was scarcely any confusion between the different measures of central tendency.

In **part (c)** neither decision, whether to create more parking spaces or not, was regarded as correct in itself; but to gain credit the decision had to be based on the given data, or the measures calculated. Candidates who speculated in very general terms, making no reference to the data or the calculated measures, were not credited.

Question 3

Most candidates understood the nature of negative correlation in **part (a)**, though some tried to explain it in terms of the appearance of the points on a scatter diagram, rather than the relationship between the variables. Very good understanding of the scatter diagrams in **part (b)** was widely shown. The most common limitation was the omission of the terms ‘weak’ and ‘strong’ for the first two diagrams respectively.

Question 4

Many candidates completed the table successfully and went on to obtain full marks. The most common error, which was made frequently, was to misinterpret, for **part (c)**, the words ‘two more home matches than away matches’ as a multiple, rather than a difference, of the two values.

Question 5

Some misunderstanding of the diagram was seen in **parts (a)** and **(b)**. For example, in **part (a)** the number of applicants in the triple intersection was sometimes excluded. But the diagrams presented in **part (c)** were mostly clear, well labelled and fully correct. The most common error in **part (d)** was to include also the applicants who had studied Architecture only.

Question 6

Good answers to **part (a)** gave a clear advantage the first of the methods had over the second, or a clear disadvantage of the second which the first did not possess. Candidates generally performed well on **part (a)(i)**, where many pointed out that not all parents could be assumed to have internet access. Answers were most limited when a claimed disadvantage of a particular method, for example no response/reply, could be a feature of all the methods.

Very good understanding of closed questions was shown in **part (b)** though not all candidates provided the limited number of responses from which the respondent was expected to choose. Answers to **part (c)** tended to focus more on which of the groups it would be most fruitful for the headteacher to ask, rather than on which it would be easiest to obtain a census.

Question 7

Almost all candidates were able to find the crude rate of **part (a)** successfully. General success in **part (b)** was not as widespread, where a common error was to apply the given crude death rate to only the females, rather than the entire population.

Question 8

Answers to **part (b)** varied in quality. Many well-drawn correct graphs were seen, but there were many also with one, or even both, of two errors: the plotting of the points at the mid-points of the class intervals rather than the upper boundaries, and the joining of the points with a curve instead of straight lines. In **part (c)** sound knowledge was displayed of how to find these measures from the graph.

Answers to **part (d)** were more limited. The best showed clearly, by means of lines drawn on the graphs, the approach being taken, since there were several ways of solving the problem. Some outstanding answers concluded with a mathematical inequality supporting the reason for the decision taken. The most limited answers, earning no credit, simply read off the download speeds corresponding to 60 per cent and 10 per cent of the population.

It was clear from answers to **part (e)** that there is very limited general understanding of the principle on which the method of linear interpolation rests.

Question 9

It now seems to be better understood than has often been the case in the past that with histograms class frequencies cannot generally be found by simply reading column heights. But the principle of frequency being proportional to column area, though more widely now appreciated, is not always properly applied from the scale definition on the vertical axis. Thus in **part (a)** three incorrect answers were sometimes seen, but together they were in the correct ratio. This was allowed some credit.

The probability questions in **part (b)** were generally well answered, though there was occasional confusion between the percentages for stationary and moving times. There were some needlessly long and involved attempts in **part (c),** where both parts required only brief solutions. Candidates who appreciated that the 60 per cent had been broken down for them into a 24 per cent and four 9 per cent for the stationary periods tended to be more successful.

Question 10

Parts (d), **(e)** and **(f)** were very well answered. Almost all candidates knew how to find the required semi-average in **part (d)**, and use it to find a line of best fit in **part (e)**. It needs pointing out again that, when finding the equation of the line, the most accurate way is to use the averages, through which the line must pass, and which are known precisely. Finding the gradient of the line by reading the coordinates of other points on the drawn line is perfectly valid, but less accurate. In **part (f)** most candidates were able to use the equation of the line appropriately for the required prediction, but many rounded the resulting decimal value up rather than down.

Answers to the interpretative parts of the question, **parts (b)**, **(c)** and **(g)** were much more limited. Very few candidates realised in **part (b)** that there could be different routes with the same distance, or that in **part (c)** the question had referred to each 'occasion' on which Salman went jogging, not each 'day'.

The best answers to **part (g)** focused on the inherent unreliability of any prediction made for values beyond the range of values on which a statistical analysis has been carried out. There is the problem of extrapolation itself; but also, in this case, it is clear from the data collected and the plot of the points on the graph, that individual times on any jog vary considerably around the line of best fit. Much more limited answers focused speculatively on the songs and the music player, for example, how favourite songs might be repeated, the player might be dropped, the battery might fail etc.

Question 11

Good skills were shown by many candidates in interpreting the pie charts in **parts (a)** and **(b)**. The most common error was to assume that the totals for the charts were in the ratio of the radii, rather than the squares of the radii.

The percentage bar chart of **part (c)** was also well interpreted. But using this chart together with one of the pie charts in **part (d)** produced wider variation in the quality of answers. Typically, amongst those candidates not solving successfully, only one of the parts of the percentage calculation might be found, for example, the total quantity of roses exported, but not the quantity of roses exported to the Gulf States.

Good answers to **part (e)** were extremely rare. Very few candidates seem to have noted that the question states that the data are for individual flowers, and that their lives have symmetrical distributions. So very few used the given information in a proper analytical way, by finding from the given measures the expected lives of individual flowers in the bunch. A very common response was to choose Variety B, with the brief justification that this variety had the highest mean and range. These answers showed lack of appreciation of the fact that, whilst the large range indicated that some of the flowers could be expected to live for a long time, it also indicated that some could be expected to live for a very much shorter time. There were many possible correct answers to this question.

STATISTICS

Paper 4040/22
Paper 2

Key message

The full cycle of statistical enquiry is explored within this examination, from planning and data collection, to presentation and analysis, and finally to the interpretation of that analysis.

- Within planning and data collection, candidates should know the names of the different types of data, be able to discuss appropriate sampling methods and select samples using a variety of techniques.
- The presentation and analysis includes the production of pictorial representations, graphical techniques and statistical and probability calculations. Candidates should have an appreciation of the advantages and disadvantages of particular statistical diagrams and calculations, and thus be able to make judgments about which would be the most appropriate to use in a specific situation.
- The best statistical diagrams should be accurately drawn, and include essential information such as units, so that they are meaningful.
- Candidates scoring the highest marks in the numerical calculations will provide clear indications of the methods they have used in logical and clearly presented solutions.
- The interpretation of statistical analysis should include detailed explanations, in the context of the situation presented.

General comments

In terms of planning and data collection, most candidates demonstrated that they knew the correct statistical language to describe data (**Question 7a**). They were usually able to select a simple random sample (**Question 2a**), but when considering whether or not that sample was representative (**Question 2b**), details of the necessary calculations were often missing. Most candidates could suggest a suitable factor for dividing the candidate population into strata (**Question 8a**), but some did not express why their chosen factor was appropriate in this situation.

Presentation of data was examined with a stem-and-leaf diagram (**Question 4)**. As is often the case with pictorial or graphical presentations of data, the units were sometimes missing from the key or the key was missing entirely. Without the inclusion of a key with units, the diagram simply displays some numbers, but with no meaning attached to them. Most candidates were able to provide a correct advantage of a stem-and-leaf diagram, although some gave the advantage in terms of ease of drawing rather than in terms of the information that it provides. Candidates were provided with a scale in order to draw a box-and-whisker diagram (**Question 7**). Most candidates were able to use the scale correctly to produce an accurate representation of the data, although a few misread the scale and plotted the upper quartile of 6.5 either at 6.6 or at 7.

Questions that involved probability and statistical calculations, such as **Questions 1b**, **3**, **6b**, and **8** were usually well presented, so that those who made an error in the calculation were able to obtain some marks for partially correct methods. In **Questions 5, 9** and **10**, however, some solutions were incomplete or unclear in terms of the methods being used.

Finally, in terms of interpretation of data, most candidates started the paper well with a correct interpretation of the time series graph (**Question 1a**). In **Question 4c** some candidates described the shape of the distribution rather than suggesting an explanation for it. In **Question 7e** it was necessary to provide general comparisons in the context of the question. Some tried to do this, but made errors such as reducing the phrase ‘number of letters per word’ to ‘number of words’. When considering a comparison of the medians, for example, the best answers were general comments in the context of the question, such as ‘the words in the Sports News article are generally longer’ rather than simply ‘the Sport News article has a bigger median’.

Comments on specific questions

Question 1

In **part (a)**, most candidates correctly concluded that the factory owner had been successful, and provided an appropriate reason by considering the downward trend of the graph.

In **part (b)**, candidates needed to find the differences between the trend line values and the actual number of absences for each of the 5 late shifts. These five differences, divided by 5, provide the estimate for the seasonal component which, as the late shift values are below the trend line, should be a negative value. Most candidates correctly found the number of absences for each late shift, but a few simply divided this value by 5, without finding any differences. Those that found appropriate differences sometimes made an accuracy error with the readings from the trend line, and some made the subtractions the wrong way around, resulting in positive values.

A reading of 11 from the trend line should then be added to their seasonal component from **part (b)**, to obtain the answer to **part (c)**. As this is an estimate for a number of people, it should be given to the nearest whole number. Many candidates correctly added their seasonal component to 11, but this was often not rounded to the nearest whole number.

Question 2

Nearly all candidates were able to use the list of random numbers correctly to identify the eight buses that should be selected in **part (a)**. A few candidates incorrectly repeated the value of 31 and a small number found a systematic, rather than a simple random sample.

In **part (b)**, many candidates observed that there were four buses of each type in their sample and correctly calculated the expected sample size in each strata, 3.6 and 4.4, concluding that the sample was representative. Those that did not gain full marks often made the general statement that random sampling should produce a representative sample or concluded that as the numbers of each bus type were equal the sample was representative, but did not provide numerical justification to support this. A minority established that the expected stratum sizes were 3.6 and 4.4, but concluded incorrectly that local buses were ‘over represented’.

Question 3

Many fully correct responses were seen in **part (a)**, with just a small number of candidates providing a correct expression in the unknown, and then struggling with the required algebraic manipulation. Some candidates appeared not to have read the information at the start of the question and assumed that they were to find the mean and standard deviation of the two results presented, rather than understanding that the means and standard deviations in the table were for the whole school. Calculations of the scaled result, z, were almost always correct.

In **part (b)** it was important to give a response in the context of the question. Many candidates talked in vague terms about scaled results being useful for comparison, rather than explaining that in this case the scaled results could be used to compare a candidate’s high jump result with their long jump result. It was important that the answer made it clear that the scaling allowed for comparison between the jumps, which could not otherwise be easily compared due to them having different means and standard deviations.

Question 4

In **part (a)** most candidates had a basic idea of how to draw a stem-and-leaf diagram. The most common error was for the key to be missing or expressed without units. As previously stated, the key is essential to make the diagram meaningful. Another common error was for the number 4 to be missing from the stem. This created a false impression of the data and led to difficulties for candidates when they came to **part (c)**. It was pleasing to see many candidates making an effort to line up the data in the leaves and present that data in ascending order. The completed stem-and-leaf diagram should resemble a bar chart, in the sense that the lengths to which the data extends from the stem gives a visual impression of the frequency for each interval, hence the need to line up the leaves.

Most candidates treated the data as if it were discrete in **part (b)**, and thus a common incorrect answer was 63 – 26 = 37 km/h. It was important here to read the information at the start of the question, which explains that the data represents speeds which have been recorded to the nearest km/h (emboldened in the question). Thus the maximum possible range is 63.5 – 25.5 = 38 km/h.

Those candidates who had drawn a correct stem in their stem-and-leaf diagram, including the number 4, were often able to state, in **part (c)**, that the data appeared bimodal. It was necessary, however, in this part to provide a possible explanation for this bimodal shape. Examples of correct answers seen were 'perhaps there were two different speed limits on the road, one for cars and another for lorries' or 'some cars saw the speed camera and slowed down whereas others continued to travel very fast'. A very small number of candidates looked back at the order of the original data and correctly suggested an example that would cause a temporary slowing down of the traffic, such as a slow-moving vehicle.

In **part (d)**, many candidates correctly stated an advantage of the stem-and-leaf diagram over a histogram as being that the original data is not lost. A significant minority stated that the stem-and-leaf diagram was easier to draw, and whilst this may be true, the choice of one diagram over another should be made because of the information that it provides and not because of the ease of its construction.

Question 5

In **part (a)** only a small number of candidates both found $P(A) + P(B) = 1.35$ and correctly concluded that, since this sum is greater than 1, these events cannot be mutually exclusive. Some candidates correctly found $P(A) + P(B) = 1.35$, but either did not complete the explanation or stated, incorrectly, that this result not equalling 1 provided the justification required. Another common error was to incorrectly assume independence, and give the intersection of both events, $P(A \cap B)$, as $P(A) \times P(B) = 0.45$, and then to state that the non-zero result supported exclusivity.

In **part (b)** the addition rule, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, was correctly used by a majority of candidates to show that the intersection of events occurred with probability 0.5. The most common error was to assume independence and state $P(A \cap B) = 0.45$.

It was common, in **part (c)**, to see $P(A) \times P(B) = 0.45$. Unfortunately it was not common to see this result compared with the result for $P(A \cap B)$ obtained in the previous part. Candidates needed to show that $P(A \cap B) = 0.5 \neq P(A) \times P(B) = 0.45$, and then to draw the conclusion that the events are therefore not independent. Occasionally the alternative correct method of comparing $P(A \cup B) = 0.85$ with the addition rule result when independence is assumed, namely $P(A) + P(B) - P(A) \times P(B) = 0.9$, was seen.

Only a small number of candidates were successful in **part (d)**. A Venn diagram might have helped candidates to see that the probability of either event A or event B, but not both, occurring can be found by doing either $P(A \cup B) - P(A \cap B)$ or $P(A) + P(B) - 2P(A \cap B)$.

A Venn diagram might have proved useful in **part (e)** too. To find the probability that neither event occurs, it was necessary to do the calculation, $1 - P(A \cup B)$. The most common incorrect answer seen came from doing $(1 - P(A)) \times (1 - P(B))$, which assumes independence.

Question 6

This question required the calculation of the mean and standard deviation from a grouped frequency table, with the complication that the data represents age. In **part (a)** the candidates were given a class mid-point which they had to explain. It was necessary for them to demonstrate, through their calculation of this mid-point, that they understood that the class boundaries for the 18–24 age group are 18 and 25, and therefore to do the calculation $\frac{(18+25)}{2} = 21.5$. Most candidates were successful, although a significant number either obtained the correct mid-point from an incorrect calculation or obtained an incorrect mid-point.

There were many fully correct solutions to **part (b)**. Working tended to be presented well, often making use of the blank parts of the table to form columns for the mid-points, x, the values of fx and the values of fx^2, and then making use of the space under the table for appropriate totals. Some candidates who had incorrect mid-points were able to score all but one of the marks, by clearly showing that they were using the correct method. A small number of errors were seen in the calculation of the standard deviation, often from the use of $(fx)^2$ rather than fx^2.

In **part (c)** it was necessary to increase the mean by two years and not to change the standard deviation. It was quite common for candidates to leave both the mean and the standard deviation unchanged.

Question 7

In **part (a)**, many candidates correctly described the data as 'discrete and quantitative'. A few did not respond to the mark prompt, that for two marks at least two terms would be expected, and only provided one of the correct terms.

There were many fully correct solutions seen in **part (b)**, but also a large number that gave an incorrect answer of 4.41, from simply calculating $\frac{4.5\times100}{102}$. Some candidates scored part marks by correctly multiplying 4.5 by 102, to obtain the number of letters in the original article, but then simply divided by 100, without using the information in the question to make the necessary adjustment for the number of letters in the new article, before dividing by 100.

Most candidates correctly calculated the largest value and the lower quartile using the information provided in the table in **part (c)**. They were then also usually able to extract the appropriate information from the table to draw the box-and-whisker diagram in **part (d)**, although a few candidates incorrectly used the mean rather than the median. Some candidates misread the scale provided, and plotted the upper quartile at 7 or 6.6 rather than at 6.5.

When making two comparisons of data presented in box-and-whisker diagrams, as in **part (e)**, an attempt should be made to compare both the central tendency and the spread, as this form of pictorial representation shows these two aspects clearly. The comparisons should also be made in the context of the question so, in this case, the fact that the median for the Sports News article is larger can be interpreted to suggest that 'in general, words in the Sports News article have more letters'. Similarly when comparing the spread, the fact that the interquartile range for the Sports News article is bigger can be interpreted as 'the words in the Sports News article are generally more varied in length'. A few candidates gave two answers which were effectively the same, such as 'the number of letters per word is generally less in the World News article' and 'the number of letters per word is generally more in the Sports News article'. A pair of comments like this will only score one mark.

Question 8

Most candidates were able to suggest a suitable stratifying factor (typically gender, age or living on/off campus) to divide students into strata when answering **part (a)**. The justification in terms of the impact of that factor on spending patterns was not always given.

In **part (b)** most candidates computed the price relative $\left(\frac{105}{103}\right)\times100=101.9$ correctly, although a few incorrectly subtracted the price relatives for 2017 and 2018.

In **part (c)** many candidates recognised that the value of 100 indicated that food prices had not changed since 2016. Those that lost the mark did so often because they referred to the 'price relative' not having changed rather giving an interpretation of what an unchanged price relative means in practice.

The majority of candidates got the price relatives for transport and clothing correct in **part (d)**, as these represented a simple increase and decrease, respectively, from the base year. The calculation for entertainment was more difficult as the increase had taken place between 2017 and 2018. Commonly seen incorrect answers were 107, 108 and 105.8, and it was quite rare to see the correct answer of 106 from correct working, $\frac{115}{108}\times100$.

The correct answer 105.9 from correct working was by far the most common answer in **part (e)(i)**, with arithmetic errors the main reason for a (rare) loss of a mark. A few candidates did not give the answer to the requested degree of accuracy.

Correct explanations in context were frequently seen in **part (e)(ii)**. The use of 'expenditure' for 'price' and failure to mention the time period were the main reasons for loss of marks.

In **part (e)(iii)** some candidates correctly referred to the likely price differences across the country, which may result in different price relatives/weights, or the fact that spending patterns would be likely to be different amongst the non-student population, resulting in different weights, to explain why the calculated index would be inappropriate to use for the whole country. Some candidates gave insufficient detail in their answers, referring too vaguely to differences in weights.

In **part (f)** many candidates correctly used their weighted aggregate cost of living index to estimate Amare's expenditure in 2018, but did not quote the result to 3 significant figures. As the weighted aggregate cost index has been given correct to one decimal place, it is not sensible to give the resulting estimate to any greater degree of accuracy than the requested 3 significant figures.

Question 9

Many fully correct probabilities were seen in **part (a)**, with at least a correct numerator seen in almost all scripts. A small number of candidates incorrectly gave the denominator as 36.

Most candidates also correctly found the probability that the goose had a tag in **part (b)**.

In **part (c)** the denominator was sometimes incorrect, with 90 or 39 being the most common errors, but many gave a correct probability.

It was pleasing to see so many fully correct answers to **part (d)**, even from those candidates that had made earlier errors. The most common errors within this question were to have 90 as the denominator in both parts of the product or to omit to multiply the product of the two probabilities by 2.

Most candidates correctly multiplied the probability of surviving the jump, 0.86, by the probability of reaching the sea given that they have survived the jump, (1 – 0.1) or 0.9, to obtain the given result in **part (e)**.

It was pleasing to see so many correct answers to **part (f)**. Expectation can sometimes be a difficult topic, but perhaps the context helped candidates to work out what to do here.

Part (g) was more difficult, as it was necessary to start by replacing the number of goslings in nests in the east cliffs with $2x$ and the number of goslings in nests in the west cliffs with x. Some candidates who did this produced a fully correct equation in x and most of those went on to solve it correctly and give a correct total.

Question 10

Many candidates found this final question on the paper difficult, as it was a more challenging question on linear interpolation than in some previous years. In **part (a)** it was necessary to establish first that the time of the 39th train working down from the top, or the 11th train working up from the bottom, is needed. Once it is established that this lies in the 16–24 class interval, linear interpolation can then be used. Unfortunately the work of many candidates suggested that they did not get this far. Those that did often correctly found that the 39th train was 19.2 minutes late. It was quite common to see this given as the final answer, rather than going on to use the remaining information in the question to find the arrival time of the train.

Part (b) also proved to be tricky for many candidates. The mode was a common incorrect answer, with the justification being that it is the most common value. The median was also a common incorrect answer with many justifying that choice by saying that it is not affected by extreme values. Some candidates gave answers which were not measures of central tendency. Those that correctly gave the mean as the answer usually justified it correctly by saying that it will be affected by the extreme values, with the best answers stating that the extreme value in the 32 to 40 minutes late interval would make the mean larger than the median.

Part (c) also required linear interpolation, but this time working from a number of minutes late back to a frequency. More of the attempts at this part, than **part (a)**, received at least part marks. Many candidates made an attempt to find the number of trains that were 15 or more minutes late or the number of trains that were 30 or more minutes late, thereby achieving some of the marks. At this stage in the paper the working was often unclear, and candidates appeared to get a bit muddled in their calculations. Those that correctly estimated that 15 trains were delayed by between 15 and 30 minutes, and 2 trains were delayed by more than 30 minutes, were usually able to complete the question correctly.

The final two parts of this question required interpretation of a pair of frequency polygons. Perhaps because some candidates were rushing at the end of the examination, many gave the answer to **part (d)** incorrectly as 12.5. It was necessary to think about how a frequency polygon is constructed and to add together the frequencies of the first 2 classes, namely 11 and 13.

In **part (e)** it was necessary to notice that there were more bus journeys than train journeys that were a lot of minutes late and more train journeys than bus journeys that were fewer minutes late and thereby to come to the overall conclusion that bus journeys were generally more minutes late than train journeys.