Examiners' Report
Principal Examiner Feedback

Summer 2025

Pearson Edexcel GCSE Statistics Higher
Paper 1 (1ST0_1H)

Edexcel and BTEC Qualifications

Edexcel and BTEC qualifications are awarded by Pearson, the UK's largest awarding body. We provide a wide range of qualifications including academic, vocational, occupational and specific programmes for employers. For further information visit our qualifications websites at www.edexcel.com or www.btec.co.uk. Alternatively, you can get in touch with us using the details on our contact us page at www.edexcel.com/contactus.

Pearson: helping people progress, everywhere

Pearson aspires to be the world's leading learning company. Our aim is to help everyone progress in their lives through education. We believe in every kind of learning, for all kinds of people, wherever they are in the world. We've been involved in education for over 150 years, and by working across 70 countries, in 100 languages, we have built an international reputation for our commitment to high standards and raising achievement through innovation in education. Find out more about how we can help you and your students at: www.pearson.com/uk

Summer 2025
Publications Code 1ST0_1H_2506_ER

GCSE (9-1) Statistics – 1ST0
Principal Examiner Feedback – Higher Paper 1

Introduction

General comments

Most candidates responded to the challenges within this paper well and demonstrated understanding of a range of areas of the specification. They were generally confident at completing calculations and diagrams and demonstrated good statistical understanding when asked to interpret these. As seen in previous series, candidates found questions requiring interpretation in context and evaluation of approaches or techniques slightly more challenging.

Candidates should be encouraged to show full working and set this out clearly so that partial credit can be awarded if a fully correct solution is not obtained. They should also read the question carefully to identify the demand, for example whether an interpretation in context or conclusion is required. Where interpretation in context was required in a question some responses gave general interpretations without consideration of the situation that had been described.

Question 1

This question required candidates to interpret data on litter sizes presented in frequency polygons in order to draw conclusions relating to wild boar litter size. It also required consideration of how the method of data collection could reduce the reliability of the conclusions. This question was attempted by the vast majority of candidates.

Part (a) required a comment on whether the frequency polygons supported the conclusion that the average wild boar litter size in France is bigger than in Italy. There were some responses that identified that the mode for France and the mode for Italy could be compared to identify that the conclusion was supported with reference to the modal values or the values with the greatest frequency. It was also common to see comments that compared the medians or means which was condoned as the relative size of these could be discerned from the frequency polygon. However, there were a significant proportion of incorrect responses seen to this question with common incorrect responses being that all of the frequencies for France were greater than for Italy or comparing the maximum litter sizes for the two countries.

In part (b) candidates were asked to give another comparison of the distributions of the wild boar litter sizes in France and in Italy. This part of the question was better answered than part (a). Some responses were able to identify that there was a greater range for the wild boar litter sizes in France than the wild boar litter sizes in Italy and others gave the contextual interpretation of this – that the wild boar litter sizes were more variable in France than in Italy. Common incorrect responses made comparison of averages (which had already been compared in part a), commented on the highest

values for the two countries or compared the overall frequencies for France and for Italy.

In part (c) candidates were asked to discuss how the data collection could reduce the reliability of the conclusions. This was answered well with the majority getting this correct. The most common issues identified were the use of secondary data, the reliability of the source or the potential for the data to be out of date. Common wrong answers included references to unequal sample sizes and to not including the whole boar population in each country. Some responses referenced bias without any further explanation which was not acceptable as an answer.

Question 2

This question was based on interpretation of a choropleth map. The question was attempted by almost all candidates and it was clear from responses that they were familiar with this type of statistical diagram.

In part 2(a) candidates were asked to show that London has been shaded correctly on the choropleth map. Overall, this was answered well by most candidates with many responses demonstrating that the shading was correct. The most common approach was to show that the percentage increase for London was 11.4...% and that to state that this was in the 10-12% category. Where responses did not show this fully it was common to see a correct calculation of the percentage increase but without reference to the 10-12% category. A common mistake was multiplying by 10 or misunderstanding the components of the percentage change formula, for example, confusing the original and final values or using the wrong denominator in the calculation.

Another, less common, correct approach was to start from the 10-12% category as shaded on the choropleth and apply a 10% increase and a 12% increase to 16.6 billion which gained the first marks. The final mark could then be awarded for concluding the 18.5 billion lay within the calculated range. Some responses using this approach did not apply both the 10% increase and the 12% increase and so were not sufficient to show that London had been shaded correctly.

Part (b) asked which region had the lowest percentage change in mileage driven between 2020 and 2021. This was very well answered with almost all candidates able to correctly identify that this was Yorkshire and the Humber.

In part (c) candidates were asked to comment on the conclusion that the percentage change in the mileage driven between 2020 and 2021 is the same for Wales and South West England. This part of the question saw a high proportion of fully correct responses with many candidates giving examples of how the two regions could have different percentages and lie within the same range. It was also common to see responses that commented on the intervals or groups which was sufficient to award the mark. The most common incorrect answer was to reference the wrong interval or to make reference to differing initial mileages for the two areas.

In part (d) candidates were asked to explain whether or not the choropleth supported the conclusion that the number of motor vehicles using the roads has increased the most in Scotland between 2020 and 2021. Some fully correct responses were seen, but it was more common to award 1 mark for a partially correct comment. Fully correct responses generally referred to only knowing the percentage change of mileage rather than of motor vehicles so the conclusion was not supported, or indicating that the choropleth did not show information about the number of motor vehicles using the road and not supported. Some candidates commented that it was possible to comment on the increase in mileage but that people were driving further rather than there being an increase in the number of vehicles.

It was common to award 1 mark for responses to part (d). The most common reason for this was for saying yes it is correct as Scotland is shaded darkest or has the highest/greatest percentage on the choropleth map. A much smaller proportion of responses referenced not knowing if the percentage change was positive or negative. The most common incorrect response was to indicate that the figures for 2020 were not known.

Question 3

This question was based on drawing and interpreting stem and leaf diagrams together with use of median and interquartile range to make comparisons. Overall this question was answered well by the majority of candidates.

In part (a) candidates were asked to explain what is meant by secondary data. This was answered very well by the majority of candidates. The most common correct answer seen was ‘Data collected by someone else’. Other correct responses were data collected by somebody else, data you did not correct yourself and data that has already been gathered or collected before. Responses which did not gain the mark were in the minority and these were generally not specific enough for example 'not from its primary source'.

Completing the stem and leaf diagram (part b) was also very well answered. The majority of candidates managed to complete the stem and leaf diagram and included an acceptable key. Where full marks were not awarded this included stem and leaf diagrams where the data had not been ordered or where responses missed off 1 or 2 pieces of data (often one of the repeated values). Some responses did not include a key or gave a key that was incomplete, for example 2|3 alone. Only a minority of responses included units with the key (acceptable provided the units were correct) and it was even more uncommon to see a key that had incorrect units included (which was not awarded the mark).

Part (c) of the question asked candidates to work out the median age for the winners of the Best Actress category from 2000 to 2022. This was well answered by the majority of candidates who were able to find the correct median of 35. Working was often seen on the stem and leaf diagram showing crossing off from each end in order to find the middle value. Where the correct median was not found some responses gave 34 which came from 23/2 = 11.5, leading to midpoint of 33 and 35. Other incorrect responses

came from finding the median of the Best Actor category rather than the Best Actress category. Others attempted to find the median from the correct stem and leaf diagram but crossed off the values in the stem and leaf diagram in the wrong order.

In part (d) candidates were asked to give a reason why it would be more appropriate to use the interquartile range than the range as a measure of spread for the data presented in the stem and leaf diagrams. This was usually answered well by the majority of candidates who made reference to one of extreme values, outliers or anomalies. Common incorrect answers included stating that the interquartile range is more accurate or gave a description of what the IQR is.

Working out the interquartile range for the winners of the Best Actress category (part e) was answered well by a good proportion of candidates, but less so than finding the median (part c). Successful responses took one of two approaches to finding the interquartile range – using systematic crossing off to find the middle of the two halves of the data, or use of the formula to find the position of the lower quartile and upper quartile in the list, and then subtracting. Where responses were not fully correct there were a significant proportion which were awarded 1 mark for a subtraction with one of the correct quartiles, this generally followed from an attempt to use the formula to find the position of the lower quartile and of the upper quartile in the list. Common incorrect responses included 16 (from $45 - 29$) or use of 49.5 (from $45 + 54 = 99, 99/2$).

The extended response (part f) asked for a comparison of the distribution of the ages for the winners of the Best Actor category with the winners for the Best Actress category. The full range of possible marks were awarded. The strongest responses were clearly structured and logical often considering comparing medians and interpreting followed by comparing interquartile ranges and interpreting.

The best responses for the comparison of medians gave a comparison of the median age of Best Actor before 2000 with the median age of Best Actress before 2000 together with a comparison of the median age of Best Actor between 2000 and 2022 with the median age of Best Actress between 2000 and 2022, and were able to interpret this in context. Although some successful responses made these comparisons as overall comments, for example ‘the median age for Best Actor is always greater than for Best Actress, so Best Actor was older on average’, it was more common to see a pairwise comparison before 2000 and a pairwise comparison for the 2000 to 2022 time period often with both being correctly interpreted in context (although a single interpretation of one of these in context was sufficient for the contextual interpretation mark to be awarded). Where all three marks for the comparison of medians and interpretation were not awarded, it was common to see the comparison of medians with no contextual interpretation or with an incorrect contextual interpretation which did not reference the age or which incorrectly linked median to the consistency of the age of the ages. In other instances responses did not compare the Best Actors and Best Actresses, but instead looked at how the median for Best Actor had changed over time and how the median for Best Actresses had changed over time this was awarded 1

mark if both Actors and Actresses were compared in this way and a further 1 mark if both were then correctly interpreted in context.

Some responses made reference to comparing medians but did not indicate which time period was being considered or that this was true for both time periods, for example "The median for best actor was higher than best actress". In this case 1 mark was awarded if a correct comparison was made with a further mark being awarded if this was linked to the correct contextual interpretation (Best Actors being older on average).

The comparison and contextual interpretation of interquartile range was not answered quite as well as the comparison of medians as the contextual interpretation proved to be more challenging. Responses generally saw similar structure to the comparison of interquartile ranges as was seen in the comparison of medians with it being common to see a pairwise comparison before 2000 and a pairwise comparison for the 2000 to 2022 time period. The contextual interpretations often showed misunderstanding of what interquartile range showed or omitted the idea of age.

Common issues with contextual interpretations were to link the statistic being compared to the incorrect meaning in context, for example incorrectly interpreting median as being related to consistency of ages, or not to include the contextual aspect and make general comments e.g. so Best Actor more consistent. Other candidates attempted context but did not link this to age, for example misinterpreting the comparison of median as meaning that more Actors won than Actresses.

Question 4

Part (a) of this question was attempted by most candidates, however the question proved challenging with only a minority of responses gaining full marks. The most common response for full marks was “appropriate because it is bivariate”. Other correct responses indicated that the approach was appropriate as it will show the relationship between the two variables. Many candidates also gave a correct justification but forgot to mention ‘appropriate’ with it.

Some responses made reference to the idea of there being two variables, but did not identify that the data was paired and so did not gain credit. Others that did not gain marks referenced seeing correlation which repeated the information given in the question. Another response that did not score was to indicate that this was appropriate without any reason being given.

Completion of the table from the histogram (part b) was answered well by the majority of candidates. If a fully correct response was not seen then some were able to gain 1 mark for a correctly labelled frequency density axis, the most common instances of this just used the heights of either or both of the first two bars to label the axis. A common error was to read the scale incorrectly and give an answer of 33.2. A common error when attempting to label the axis was to mark the major gridlines as increasing in 2’s rather than 1’s.

Part (c) required candidates to use the information in the table to complete the histogram. This was less well answered than completion of the table, but there were still a good proportion of correct responses seen.

Correct responses gave a correctly labelled frequency density axis and had the correct height for the bar for $26 < t \leq 28$. Responses that did not gain this mark sometimes had the height of the bar correct, but lost the mark due to incorrect labelling on the axis often using frequency values rather than frequency density values.

Part (d) of this question presented candidates with a histogram and the conclusion that the median was in the $50 < t \leq 52$ minute time interval. They were asked to decide whether the histogram supported the conclusion and show calculations to support their answer. The best responses generally took the approach of showing a calculation for the median position within the data and demonstrated that there were too many runners in the first three time intervals for the median to lie within the $50 < t \leq 52$ interval. When calculating the median position the use of $\frac{n}{2}$ or $\frac{n+1}{2}$ were accepted, and it was condoned if candidates used the total frequency from part (b) of the question (which differed from the total frequency in this histogram as some skiers who complete the classic stage of the race did not complete the full race). Although this was the most common correct approach there were other correct responses seen, for example showing that there was an equal total frequency in the $44 < t \leq 46$ and $46 < t \leq 48$ groups as in the $50 < t \leq 52$, $52 < t \leq 54$, $54 < t \leq 56$ and $56 < t \leq 58$ groups which demonstrated that the median should lie in the $48 < t \leq 50$ interval. Using the frequency density only of the bars to calculate the median class was rare, but when it was seen, the candidate tended to gain two marks.

Many responses scored just one mark as they showed calculations for the median position and concluded that this would fall within the $48 < t \leq 50$ time interval, however did not demonstrate why this position would be within this interval. There were also instances of responses gaining just 1 mark due to errors in evaluating the calculations.

Common incorrect answers included not showing calculations and using the modal group as a justification rather than working with median.

In part (e) of this question candidates were asked to comment on the appropriateness of use of a matched pairs approach to see whether use of a different wax on skis improved performance. This was generally answered poorly with few fully correct responses seen. The strongest responses identified that this was appropriate and made reference to the approach controlling for the difference in skier performance or differing levels of equipment, with skier performance being the more commonly seen of the two. It was more common to award 1 mark for part (e) for identifying that this approach was appropriate as you can see the effect of changing the was or for the non-contextual response of appropriate as this controls for extraneous variables.

Many incorrect responses involved suggesting use of a repeated measures approach with a single skier rather than commenting on the approach suggested as was required.

In other instances responses focussed on things that had already been accounted for, for example saying that skier performance might be different for the two skiers, and then stated that the approach would not be appropriate.

Question 5

This question was based around use of statistical techniques to collect data on the effectiveness of an advertisement campaign and vouchers to promote use of leisure centres. Part (e) then followed on from this, presenting candidates with a plan, process and present data to investigate whether the number of people using the leisure centre has increased following the advertisements and vouchers.

In part (a) candidates were asked to comment on the use of cluster sampling for correcting data on the advertisements with the clusters being the leisure centres. Those candidates who attempted a contextual reason here did so well; most frequently commenting on that the sample only considered people at the leisure centres rather than those who may have seen the adverts but decided not to visit. Non-contextual answers often referred to the sampling being biased or not representative, where this was accompanied by an indication that this was not suitable this was awarded 1 mark. Responses which suggested an alternative sampling method or talked about the time or cost involved were not awarded any marks.

In part (b) candidates were asked to give a reason why a pilot study would be an appropriate thing to do. This was generally very well answered with common correct responses including to get an idea of the response rate, to check for any issues with the questions or to see if changes needed to be made. A minority of responses described what a pilot study was rather than indicating why this would be appropriate in this case and were not awarded marks. Other incorrect responses indicate that the pilot survey would give more data or did not show understanding of what a pilot study was.

Part (c) required candidates to explain why it is appropriate to use the random response technique in the given context. This question was well answered with candidates recognising that people would potentially be embarrassed to answer the question and not wish to answer honestly. Common answers which were also accepted were to make answering the question anonymous. However, those who did not achieve the mark were often focusing on the word ‘random’ in the question, saying that it would eliminate bias or give everyone an equal chance.

Designing a random response question (part d) was answered well by a good proportion of the candidates. The strongest responses gave a clear description of a how to obtain a random outcome, most often using ‘flip a coin’, then indicated that the respondent should answer yes for one outcome and answer the question for the other together with a suitable question related to whether the person had used more than one of the vouchers. Responses which were not fully appropriate were also often able to gain partial credit. Most responses were awarded the first two B marks, writing an appropriate question and describing how to obtain a random outcome. The main error seen in such answers was to ‘tick no’ for certain outcomes rather than yes, thus failing to remove the possible embarrassment. Unfortunately, many responses were simply a

question with response boxes rather than including any random response technique or included a question that did not find out the information required.

The extended response (part e) provided candidates with a plan to investigate whether the number of people using the leisure centres had increased after the advertisements and vouchers had started and asked for a discussion of the appropriateness of the plan. This part of the question was attempted by the majority of candidates although the length of responses and also the number of marks awarded varied significantly.

The most successful responses were the ones that considered points in the plan in detail, linking the different elements of the plan with appropriate or not appropriate and a reason. However, some responses focussed on each bullet point in turn often did not appreciate the linkage between the elements. The weakest responses made statements about aspects of the plan being appropriate or not appropriate without any reason being given. Quite a few responses were seen where the answer provided was a description of how to carry out the study rather than justifying the appropriateness of the current study design.

In general, responses scoring partial marks were often better at commenting on the data collection and often scored two marks here for saying that the time frame was appropriate, with data collection before and after the publicity started, or that the data was representative of all leisure centres. Following on from a correct comment about one of the approaches described in the plan they had to state whether the feature they highlighted was appropriate or not to achieve two marks here.

Many students commented on the imbalance of 2 weeks before the four after as being inappropriate which did not gain marks. Other common remarks that did not gain credit were to classify this data collection as secondary data and then state that it was an unreliable source, suggesting that the leisure centres may not provide the data, that the time period was not long enough, that the process would be time consuming or there was too much data.

For the processing and presenting data, most marks were awarded for recognition of a time series graph showing a trend or change over time and for the unreliability of predicting results due to extrapolation. A minority of responses noted the error in using mean added to the trend line and corrected this by discussing mean variation added to trend line. Many responses linked seasonality to the seasons of the year, this meant that accessing the 4-point moving average mark was limited as it was common for these responses to state that the time frame did not cover the year. In these cases there was no appreciation shown that a 7-point moving average was better due to the 7 days in a week; with many incorrectly suggesting that a 6-point average would be better. Another common misunderstanding was that moving averages smooth fluctuations and do not produce a trend line, so it was not common to award marks for the fourth bullet point. Some also spent time saying what Nico should be doing instead, which often scored zero as the question had asked for the appropriateness of the plan given rather than improvements or alternative approaches.

Question 6

This question required candidates to calculate Spearman's rank correlation coefficient (part a), interpret given correlation coefficients (part b) and identify the correlation coefficient which would represent perfect agreement (part c). Overall this question was very well answered and responses demonstrated familiarity with the statistical technique and how to interpret the results.

Part (a) was answered very well by candidates. Around half of responses had fully correct calculations for Spearman's rank correlation coefficients and reached a correct answer. Where responses were not fully correct the marks awarded were fairly equally split between those scoring 0,1, 2 and 3 marks. Where responses were awarded 3 marks this was often due to slips in ranking, slips in evaluation of the difference in the ranks or correctly substituting into the formula but with errors in evaluation. Responses awarded 2 marks were those where the ranking and difference in ranks were successfully completed, but where the formula was not correctly used, for example a common error was including '1-' in the numerator rather than outside of the fraction. Where 1 mark was awarded this was for the ranking of the yellowness of cheese, a common error following this was in squaring the negative numbers and producing negative answers.

In general, it was uncommon to see reverse ranks, but responses using these showed similar levels of success to those using ranking in ascending order.

In part (b) candidates were presented with a table of Spearman's rank correlation coefficients and asked candidates to interpret in context the correlation for the characteristic that was most closely related to 'customer overall liking' scores.
The majority of responses were able to gain at least one mark for identifying that salt content was the most closely related to the 'customer overall liking' scores. The strongest responses went on to comment that as salt content increased the customer overall liking scores increased or, more commonly, that customers liked saltier cheeses.

Few attempted to interpret the correlation but of those who did try, few could interpret in context, and instead would comment on the value of the coefficient being positive or closer to 1. A common response was to say that 'people like the salt content' which received B1B0 as it didn't indicate whether it was increasing or decreasing salt content that was preferred.

In part (c) candidates were asked what the Spearman's rank correlation would be for perfect agreement. The majority of candidates were able to identify that this would have a correlation coefficient of 1. A range of incorrect answers were seen including a number of decimal values between 0 and 1.

Question 7

The first part of this question, on using Petersen capture recapture to calculate an estimate for the number of black bears in Pennsylvania in 2018. This was generally well answered. Where responses did not gain 2 marks there were a small proportion who set up a correct equation but were not able to rearrange this correctly to obtain the estimate. Other responses that gained 1 mark were those where the calculation was performed in stages leading to a loss of accuracy, in these cases the working often began with a calculation such as $\frac{690}{73} = 9.5$, followed by multiplying this value by 2017.

Part (b) involved candidates considering what would happen to their estimate if tags from previous years were included in the count of the tagged bears in the recapture. This was generally not answered well, with a minority of responses gaining full marks. The strongest responses identified that the number of tags in the recapture would increase and that this would lead to a decrease in the estimated population. Where two marks were awarded this was most commonly for a response that identified that the estimated population would decrease and explained that the denominator of the calculation would increase, this showed clear understanding of the impact of the greater number of tagged bears. Slightly more responses were awarded one mark. This was usually for saying their estimate would decrease, with no explanation included or with a vague unclear reason.

In part (c) candidates were asked to give an assumption of the Petersen capture recapture method, other than no loss of tags, that should have been considered in the research. Around a third of the responses gave a correct assumption, most commonly referring to the population not changing, there being no births, no deaths or no migration. A common error was to give a factor that would violate the assumptions, such as 'bears were born or died' rather than no births or no deaths. Those candidates that references tags falling off scored zero as this was the one referenced in the question.

Question 8

The first part of this question (a) asked candidates to use the graphs and summary statistics provided to show that there were no outliers in the times for the Ironman Canada in 2022. This involved identifying the appropriate outlier calculation based on the available summary statistics, using mean and standard deviation, and using this to find outlier limits before comparing these to the limits of the data shown in the histogram. This was found to be one of the most challenging question parts on the paper with a significant minority not attempting the question and many other responses that were incorrect.

The strongest responses to part (a) extracted the mean and standard deviation for Ironman Canada 2022 from the summary statistics before calculating $\mu \pm 3\sigma$ to obtain limits of 8.013 and 19.047 before stating that 8.5>8.013 and 18<19.047. Fully correct responses were uncommon, but there were more responses that gained 2 marks for correctly calculating the outlier limits without identifying the limits of the data presented in the histogram or identified values from the histogram that did not

demonstrate that there could be no outliers. A small proportion of responses gained 1 mark for either calculating only one limit or for correct calculations shown but with errors in evaluation. A common error was to use the median rather than the mean in the attempted calculations.

Part (b) of the question asked candidates to compare, in context, the distribution of the Ironman Canada 2022 men's results with the distribution of the Ironman UK 2022 men's results. This question differentiated well between candidates with the full range of available marks being seen. The strongest responses gave a well-structured set of comparisons generally starting with a comparison of average (often comparing both the means and the medians) and contextual interpretation, then a comparison of spread (often the standard deviations) and contextual interpretation, followed by a comment identifying the skews and interpreting these in context. Where responses scored 5 marks it was most common to see the omission of, or an incorrect, contextual interpretation of skew. If one of the comparisons was omitted or incorrect it was most commonly that of skew and these responses were often able to gain 4 marks for correct comparison of average and contextual interpretation (2 marks) together with correct comparison of spread and contextual interpretation (2 marks).

The contextual interpretation aspect of the comparisons was an aspect of the response that was commonly incomplete, incorrect or omitted. Some responses attempted to interpret the comparisons made but did so in general terms or lacked the contextual aspect. Others made incorrect comments that interpreted the comparisons in the wrong direction, for example stating that the average competitors were slower in the Ironman Canada or the times were more consistent for Ironman Canada. There were also some responses where a correct contextual interpretation was given which was clearly linked to the wrong statistic and in these cases the interpretation was not awarded the mark.

It was reasonably common to see responses that scored 2 or 3 marks. This was generally for correct comparison of either means or medians and correct comparison of spread (2 marks) which was sometimes accompanied by one correct contextual interpretation which was most commonly related to the comparison of average.

In part (c) candidates were asked to use standardised scores to compare performances relative to other competitors in each stage of the race. This was a relatively challenging question within the paper. Only the strongest responses scored full marks, showing the correct standardised score for the run for Dan together with clear comparisons of standardised scores indicating that swim had the best performance as it had the most negative standardised score and the run had the worst performance as it had the most positive standardised score. It was more common to see responses that gained 2 marks for correctly calculating the standardised score for Dan for the run, these responses often attempted to comment on the relative performances but did not do so correctly due to incorrectly interpreting the negative standardised scores as meaning that the performance was worse than average and positive standardised scores as meaning the performance was better than average which is not the case when

discussing times in a race. Other responses attempted to compare the performances relative to the other competitors but did not use the standardised scores.

A minority of candidates scored just 1 mark for a correct calculation for the standardised score which was not accurately evaluated or where the subtraction in the calculation was performed the wrong way around. A common error was to attempt to find the standardised score but to use an incorrect formula and this did not gain any marks.

Question 9

The first part of this question required candidates to complete the probability tree diagram. Many candidates gained this mark. Responses that did not gain the mark incorrectly filled in the second branch either with incorrect numerical values instead of p and 1-p (not 0.45 and 0.55 which were allowed as these were correct). A common incorrect response was p and q on the 2nd branch or 0.3 and 0.7.

Part (b) involved use of conditional probability. This was found to be challenging with few fully correct responses seen. The most common incorrect response was $0.3 \times 0.4 = 0.12$. Some candidates made an attempt at forming an equation but this was of often done incorrectly or not equated to 0.3. There were a small number of responses where $p = 0.45$ was obtained but then spoilt by further working, in this case 1 mark was awarded.

In parts (c) and (d) candidates needed to find probabilities using the Binomial distribution. As would be expected for the final parts of the last question on the paper this was found to be challenging by candidates.

The strongest responses in (c) showed clear working and were able to find the required probability, often by using $1 - {}^4C_0 \times 0.05^0 \times 0.95^4$, although some responses that showed a correct method lost the accuracy mark by prematurely rounding off their answer.

In part (d) few responses gained full marks. Some responses showed use of the Binomial distribution to find one appropriate probability and gain a method mark. This was often followed by award of the B mark for a correct conclusion based on their answers to part c and d as the award of this mark required that there was an attempt at the Binomial distribution. There were some responses that were awarded 1 mark where probabilities using the Binomial distribution were attempted but were incorrect and this was then followed by the correct decision following on from the answers given to (c) and (d).

Summary

Based on their performance on this paper, students should:

- show working for statistical calculations,

- practise writing clear explanations, bearing in mind exactly what is asked in the question and what evidence you should give to support your answer,
- practise interpreting statistical calculations in the context of the question,
- practise giving statistical reasons for or against statistical approaches suggested,
- practise calculating limits for outliers,
- practise using the binomial distribution to calculate probabilities,
- practise working with and interpreting standardised scores in a variety of contexts.