Pearson Edexcel

Examiners' Report
Principal Examiner Feedback

Summer 2022

Pearson Edexcel GCSE
In Statistics (1ST0)
Higher Paper 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 2022
Publications Code 1ST0_1H_2206_ER

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

Introduction

General comments

Most students 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. students found questions requiring interpretation in context and evaluation of approaches or techniques more slightly more challenging.

Students should be reminded that when drawing any form of graph or diagram they should use a ruler and a sharp pencil to ensure accuracy. They should also 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.

Question 1

As would be expected for the first question on the paper, this question was answered well, with the majority of students achieving full marks. students were usually proficient at calculating the averages correctly although it should be noted that although calculating the mean was a lot more work than the median or the mode, many students didn't realise this easier option – this was more likely to lead to calculation or rounding errors. Of the three averages the median was the most common, with mean next. Few students compared the mode.

Other than errors in calculation of the chosen average, common errors included calculation of the averages without a comparison in context and students who attempted to make comparisons using the range or highest/lowest values.

Question 2

Part (a) of this question required students to interpret a population pyramid. The majority of students were able to correctly identify the percentage of the population who were female and in the age group 50-54. Where incorrect responses were seen this was generally due to reading the wrong bar from the population pyramid (either the male bar or a different age range).

Part (b) of the question required students to identify the two relevant bars on the male side of the population pyramid and add these together. For the vast majority of students this proved straightforward and there were a high proportion of fully correct answers seen. A minority of students had correctly identified the two bars but made an error in adding the two, however this was often accompanied by the correct addition

being shown so some credit could be awarded. Where incorrect answers were seen these included adding the values for the female bars rather than the male bars $(2.7+2.7=5.4)$, finding the mean of the two appropriate percentages $((2.8+2.9)/2=2.85)$ or adding both male and female percentages for the appropriate age groups $(2.9+2.8+2.7+2.7=...)$

For part (c) of this question students need to explain why the 100+ age group was given as 0.0%. The majority of students were able to divide 13310 by 65511097 but a common error was to omit the multiplication by 100 to obtain a percentage. students who did obtain the correct percentage were often able to go on to explain that this would round to 0.0% or refer to rounding to 1 decimal place, however there were also those who did not give sufficient detail in their justification at this stage and merely referenced rounding to 0. Those trying an alternative method were less successful, comparing with 1% or 1% of the population with only a few comparing with 0.05% as would be necessary for the justification if following this approach.

A common issue in responses to this question was that students attempted to justify the 0.0% without use of the number of people in the age 100 and older group and total population, often by adding up all of the other percentages or by making broad statements. It is important that students read the instructions given in the question carefully – in this case there was an instruction to use the information above and from the population pyramid and show working.

In part (d) students were asked to comment on a given conclusion comparing the number of males older than 40 with the number of females older than 40. It was anticipated that students would identify that there was a higher percentage of females rather than males for each of the age groups from 40 upwards, however this was only identified by a minority of students. Some students commented on more females than males over 40 but did not make a justification for this or make it clear that they had compared the individual age groups.

The most common approach taken was to add all of the percentages for males older than 40 and add all of the percentages for females older than 40 and then to make a comparison. It was a little disappointing to note that arithmetic errors were relatively common when taking this approach, despite the availability of a calculator. There were some students who attempted the approach of adding all of the percentages but did not obtain the correct values because they omitted the 40-44 category in their calculations or only compared the 40-44 age group.

A common incorrect answer for part (d) was to indicate that a comparison could not be made because there were only percentages and not the numbers of males and females.

Question 3

Part (a) of this question asked why the given data would need to be cleaned. This was answered extremely well, with most responses commenting on missing data and/or different formats (usually by identifying that “two” was should have been written as “2”). There were a few responses that wrongly talked of outliers, making the data have the same or made generic comments such as making data easier to read or process.

In part (b) students were asked why the value of 124 in the spreadsheet must be wrong. Most students realised that the value of 124 was greater than the total though their answers were often expressed poorly. Few students reasoned that 124 should have been 24. The most common incorrect responses referenced anomalies or outliers.

Part (c) of the question asked for the use of linear interpolation to work out an estimate for the median speed of the motorcycles. There were some good correct responses to this showing fully correct working, but equally there were also many who just identified the group containing the median. Of the correct responses the vast majority found 73.04 rather than 73.125. A minority of responses calculated the estimated mean rather than the median.

In part (d) students were asked to draw a frequency polygon. The vast majority of students were able to correctly draw this frequency polygon adding to the one that was already provided. A significant minority of students failed to plot their points at the midpoints of each interval, but many were allowed one mark for plotting consistently within the intervals.

In the final part of the question (part (e)) students were asked to use the two frequency polygons to compare the skew of the distribution of the speeds of the cars with the skew of the distribution of the speeds of the motorcycles. The question also indicated that the comparison should be interpreted in context. Many students correctly identified skew, however few were able to correctly interpret that skew, and especially not in context. Most students would say that ‘motorcycles were faster as the speeds are towards the right’ or something to that effect. The most common response to be awarded the final mark was stated that the motorcycle speeds less than the median were more spread out than those greater than median. Many attempts to interpret the skews were confused, inaccurate or contradictory, for example interchanging mean with median in the interpretations.

Question 4

This question was an extended response question where students were given descriptions of three possible sampling methods, asked to identify the type of sampling and discuss the appropriateness of each. Most students attempted this question and the vast majority of them were able to achieve some marks.

In terms of identifying the types of sampling described in the question there were only a minority of students who could correctly identify all three types. Those who were able to name the sampling methods were most likely to name Systematic and least likely to name Stratified. A significant majority of students incorrectly named Stratified sampling as Random. Some students did not realise that the sampling methods needed to be named.

Students generally made attempts to comment on the appropriateness of each sampling method, commonly commenting on these in turn. The most common correct comment on sampling method A (quota) was to identify that the method of selection was likely to be biased. A common incorrect comment on method A was where students indicated that they believed that the method would lead to a sample which was in proportion to the sizes of the shops.
Commenting on the appropriateness of method B proved more challenging for students than commenting on either method A or method C. In commenting on method B students often described the method again or explained how it would be carried out in practical terms. The attempted comments on method B were often the briefest of the three.

Students often commented in detail on method C, with many responses identifying more than one of the points identified in the mark scheme. Common correct comments included identifying that the method is random and that the sample is representative. Method C was (correctly) the most common method to be named as the most appropriate, however there were instances of both method A and method B being selected. In a significant minority of cases students did not identify one of the three methods as the most appropriate.

Overall, students were good at highlighting the lack of randomness and the potential bias in some of the sampling methods. Whether methods gave good representation (either of stores or employees) was often confused and sometimes contradictory. When writing a number of comments about a method students must be careful not to contradict themselves by saying something is both biased and unbiased, or representative and unrepresentative for example.

Question 5

In part (a) of this question students were asked to draw a cumulative frequency curve for the data provided. Most students knew how to draw a cumulative frequency diagram and achieved both marks for this question. The most common error was to plot the points at the midpoint of the intervals given in the table rather than the upper bound. The most common incorrect answer here (apart from incorrect plotting) was students producing bar charts rather than Cumulative Frequency curves.

In order to answer part (b) of the question students needed to read appropriate values from their cumulative frequency graph and find the difference between the two. This part of the question was answered well, even by those students who made mistakes in part a. A small but significant number of students gave answers that were not integers, clearly not appreciating the context of the question. It was apparent from the figures seen in working that some students had been inaccurate in reading from their graphs. A common mistake was just taking the value of £350.

Occasionally students used interpolation to work out the answer for part (b), which scored 2 marks as long as they remembered that their answer must be an integer. Part (c) of the question asked whether the data on house prices in Streetly could be used to predict the number of houses in Central London that have a price between £300 000 and £350 000. The question was answered well by most students with a minority of incorrect answers which often discussing the sample size. students who thought that the method was appropriate often explained how interpolation was being used and missed idea that the locations were not comparable. Several students failed to indicate whether or not the method was suitable.

Question 6

In part (a) students were asked to identify an appropriate diagram. The majority were able to indicate that this was a scatter diagram, although a range of other graphs were given incorrectly.

Part (b) presented information about the data for some football teams giving the mean value of all players in the team and the final position of the team. students were asked to calculate Spearman's rank correlation coefficient for the information in the table. The majority of students made a start at calculating this by completing the values of d and also d^2, however the extent of progress beyond this point varied quite considerably. Some students made mistakes subtracting ranks, others made errors when squaring and obtained negative numbers as a result of their attempts to square. Where an attempt to add the values of d^2 was made this was often followed by use of the correct formula in a attempt to find Spearman's rank correlation coefficient. students generally used the formula given with a high degree of success but some forgot to subtract from 1, others tried to square the total sum and some had an incorrect value for n.

For part (c) students were asked to interpret their answer to part (b) in context and comment on the effect of any anomalous data. Many students gave a correct interpretation in context as was required. However, a similar number stated a positive correlation or agreement in ranking between position and mean value without interpreting it. Very few students were able to correctly describe the impact of the outliers; most either ignored that part of the question or described which teams were outliers or said outliers had an effect making the result less accurate/reliable.

The final part of this question (part d) required a comment upon whether Pearson's product moment correlation coefficient should be used instead of Spearman's rank correlation coefficient to measure the correlation in the given data. This question was attempted by most students, who referred to the appropriateness of Amelia's suggestion. This was not awarded a mark, unless one of the coefficients were referred to correctly. In the majority of cases, students referred to Pearson's correctly more times than Spearman's. The majority of those that attempted this question understood that Pearson's linked to a linear association, referring to a line or linear correlation; however, students seldom understood that Spearman's was connected to ranked variables.

Question 7

In part (a) most responses correctly identified that the use of 4-point moving averages for the data in the table was appropriate because the data is in quarters. Those who didn't answer correctly often wrote a spurious reason that did not link to the number 4.

For part (b) students were asked to plot the final 4-point moving average onto the time series graph. Most who attempted this part plotted the point correctly but many did not plot a point at all. Where incorrect plotting was seen the most common error was to plot the plot at quarter 3 of 2018 rather than between quarter 2 and quarter 3 of 2018.

For part (c) students were asked to draw a trend line for the time series graph. The majority of students were able to draw an acceptable trendline. A few students joined the data pointwise but this was a minority.

Part (d) was answered poorly. In this part of the question students were asked to describe how the time series graph, the trend line and average seasonable effect could be used to obtain a prediction for the 1st quarter of 2019. There were many vague descriptions that lacked the detail. There were also a number of completely incorrect suggestions such as drawing a trend line based solely on Q1 figures. In describing how to calculate the seasonal variation, many students only talked of the difference and didn't indicate which way round the subtraction should be. In describing how to use the mean seasonal variation to obtain an estimate, students often said 'subtract' from trend line or 'add/subtract' rather than simply 'add'.

Question 8

Part (a) of this question required the student to describe what the chain base index numbers how about the average price of houses in January for the years 2016 to 2019. About half the students recognised that the house prices were increasing. Of these students about half of them again then recognised that the rate of increase was decreasing.

Part (b) of the question was slightly unusual as it required the student to work backwards using a chain base index number to find the average price of houses in January 2016 based on the average price of houses in January 2017. There were a significant number of students who were able to identify the correct chain base index number to use and successfully work backwards to find the average price of houses in January 2016. There were, however, a significant number of students who divided by 104.76 and then failed to multiply by 100. Some students worked with the chain base index number for January 2016 rather than January 2017 which led to the special case answer. The most common error was to divide by 104.76 and then multiply that answer by 107.76.

Question 9

This probability question was often attempted by use of a tree diagram and students working using this approach were often successful provided they could identify the correct probabilities to include on the branches. Other students identified that they needed to calculate $3/8 \times 5/9$ and $5/8 \times 4/9$.

Common errors included adding probabilities that should have been multiplied (and) and multiplying probabilities that should have been added (or). Some students multiplied all four of the relevant fractions together. There were also instances seen of students with correct methods obtaining incorrect answers due to errors in arithmetic when working with the fractional probabilities. A small number of students chose to work with decimal values rather than fractions and in these cases there were instances of loss of accuracy in intermediate calculation steps which led to an inaccurate final answer.

Some students only worked out the probability of two red discs or the probability of two yellow discs (rather than the probability that both discs were the same colour).
In a minority of cases students merely combined all of the discs in bag X with those in bag Y and gave a probability out of 17 as their answer.

Question 10

This was well done by the majority of students. Most knew how to calculate standardised scores and then interpret them correctly. Almost all students who used the correct method, then achieved the accuracy marks by arriving at the correct answers. Some students thought that the value closer to 0 was the better of the two, students also confused the idea of standardised score with that of skewness or correlation. Occasionally a student didn't realise that they must assess the conclusion and didn't state that it was wrong. A minority of students did not use statistical calculations in their attempt to evaluate the conclusion given in the question – students should read the question carefully to ensure that they are following the guidance provided.

Question 11

In part (a) of this question students were asked why it was only possible to calculate an estimate of the standard deviation for the data presented in the table. This was answered correctly by most students, generally by referring to not knowing exact values. Knowledge of the method which would be used to estimate was shown by the good few responses which referred to the need to use midpoints. Incorrect responses were either vague or did not appear to understand the question. A minority of students referred to outliers or the accuracy of the data in their attempt to answer this part.

For part (b) of the question students needed to use the estimated summary statistics presented to them to calculate an estimate of the skew of the distribution of the times taken by the 50 boys. There were a pleasing number of fully correct responses where students had been able to use the estimated summary statistics to find an estimate of the standard deviation of the data and then calculate the skew. The most common error was in attempting to calculate an estimate of the standard deviation for the data with a significant proportion of students able to do this correctly, although some were still able to gain some credit for use of the correct formula for skew with the given values and an attempt at an estimate of standard deviation.

In part (c) students were asked to interpret the skew that they had found in (b) in context. This was poorly answered. Many students did not score in part c as, although it was common to see references to negative skew, there were very few correct interpretations of the skew in context.

Question 12

This question related to the use of the binomial distribution as a model for the number of sixes recorded when dice are rolled.

In part (a) students were asked to write down two conditions needed so that a binomial distribution is a suitable model for the number of sixes recorded. In order to answer this fully students should consider how the conditions for the binomial distribution to be used apply in the context of the question and therefore in order to be awarded full marks for (a) at least one of the conditions needed to be given in context.

In this first part of the question some students gave two (or more) conditions and included context giving a fully correct answer. Those who did answer in context were generally able to explain that the probability of getting a six needed to remain constant and that the only outcomes should be a six or not a six. There were also lots of answers which correctly listed binomial conditions but didn't actually write them in the context of the question. Common incorrect answers included those who stated conditions to ensure a fair experiment e.g. all dice thrown the same way rather than considering that the dice were already stated as "fair".

Part (b) required calculation of the probability that all of the 4 dice would land on a six. This was generally answered very well. A small number of students made things more difficult for themselves by using decimals and rounding too early. A common incorrect answer was to multiply 1/6 by 4 rather than raising to the power of 4.

In part (c) students were asked to find the probability that at least 2 of the 4 dice land on a six. As would be expected, this proved more challenging than part (b). Significantly fewer students were able to obtain a fully correct answer to this part of the question. The most common error was to omit the nCr aspects of the binomial calculations.

Question 13

This question required the use of geometric means and proved to be very challenging. Although many students understood the process for calculating a geometric mean, only a small minority knew to convert to the overall percentages (or multipliers) prior to performing the calculation.

Students who found the geometric mean of the values as presented in the table were able to gain partial credit for knowledge of geometric means. The most common incorrect method was calculation of the arithmetic mean which no credit was given for. Regardless of the methods used, students were good at interpreting the values of their answers to make appropriate conclusions.

Question 14

This was an unusual question about probabilities which required students to work backwards from the probabilities given, combining these with the information in the question and the partially completed Venn diagram in order to complete the Venn diagram.

Most students found this question challenging and very few students were able to give a completely correct answer. Many did not know how to start the question with the information provided.

Summary

Based on their performance in this paper, students are offered the following advice:

- 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.
- Develop their understanding of when Spearman's rank correlation coefficient is appropriate compared to when Pearson's product moment correlation coefficient.
- Develop their understanding of the meaning of skew.
- Develop their understanding of chain base index numbers and practise answering a mixture of questions using these.
- Practice calculating geometric means for a range of different contexts.

Pearson Education Limited. Registered company number 872828
with its registered office at 80 Strand, London, WC2R 0RL, United Kingdom