Pearson

Examiner's Report
Principal Examiner Feedback

Summer 2018

Pearson Edexcel International A Level
In Statistics S1 (WST01/01)

edexcel

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 2018
Publications Code: WST01_01_1806_ER

General introduction

Overall this paper allowed all students to demonstrate their ability and knowledge of the WST01 specification. Questions 3 and 4 provided a good source of marks through routine calculations whilst questions 6 and 7 proved challenging for students at the top end. Though many did display good statistical knowledge throughout, quite a large number of students were clearly not prepared for some topics on this specification, notably the normal distribution and how to use frequency density to calculate the height and width of a bar on a histogram. Students found questions 1(c), 1(f), 2(d), 5(b) and 5(f) challenging as these required statistical reasoning. Students should be advised to read questions carefully as marks were dropped carelessly for not meeting the required demands.

Report on individual questions

Question 1

Students made a strong start to the paper and part (a) of this question was answered well, with most showing full working for finding $\sum y = 141.5$ and writing a clear and correct expression for S_{yy}.

Almost all students knew how to calculate the product moment correlation coefficient in (b) although there are still some who lose the accuracy mark by only giving their answer to 2 significant figures. Again part (c) was generally well answered. Some students believed that it was simply the positive value that supported the fitting of the model rather than the strength of the coefficient, being close to 1. A small number of students chose to answer their own question by interpreting the coefficient instead of saying why the coefficient supports the regression analysis.

The calculations for the regression line in part (d) were usually carried out well and an encouraging number arrived at a correct equation in x and y. The gradient was mostly correct although a few had the fraction the wrong way up. In calculating the intercept errors were made in finding Σx. In a few cases a and b were found but the final equation of the line was not stated.

Part (e) was more challenging as some students struggled to deal with the coding. It was pleasing, however, to see a large number of correct solutions. The most common involved substituting 4.4 into the equation of y on x and then adding 25. Errors here included forgetting to add on the 25 and slips in accuracy. Some used the equation of m on p and successfully negotiated the fraction $\frac{p}{10}$ to achieve a correct final answer. Though some incorrectly substituted $p = 4.4$ or even worse $m = 44$. An answer out of range should have been a sign to students that a mistake had been made not that the estimate was unreliable.

Marks were commonly lost in part (f) where inaccurate or incomplete statements were used to justify the reliability of the estimate found. Many stated that the estimate was reliable but they were unable to refer to the correct variable or value that is in the range, often referring to the y or m value instead of the x or p value. The answer "it is reliable because it is the range" scores zero as it is not sufficiently clear to imply that interpolation is being used. Other answers which scored no marks included: ‘not reliable because data is coded’ and ‘reliable because *r* is strong’.

Question 2

There were accessible marks for all students on this question with parts (a) and (e) discriminating the most able. Part (a) was the part of the question that was often answered incorrectly with the most common incorrect answer being 33. Students did not show understanding that 75% exceed the lower quartile.

Most students answered part (b) well although there were a few, who, having correctly written down the values of Q_1 & Q_3 then failed to subtract them to find the interquartile range.

There were many correct box plots seen in part (c), with most calculating the correct upper and lower boundaries of 38 and 22 and then usually marking in the two correct outliers. There are still some students who seem to have difficulty in interpreting the boundary instructions for identifying the outliers and even some who did the correct calculation but then marked in 2 or 3 outliers at the lower end. It was clear that some students did not know if 22 itself was an outlier (being on the boundary) and sometimes plotted their minimum value as 23. A significant number of students drew two upper whiskers and this was penalised.

Most students demonstrated good knowledge of how to use quartiles to justify the skewness of the distribution. It was very common for students only to identify that *Eastyou* was symmetrical with correct justification in part (d). They did not read the question properly and said nothing about *Westyou*. A handful spent time calculating the mean for *Eastyou* and then used this as a justification for (slight) positive skew.

Question 3

This question was well attempted with nearly half scoring full marks. Most scored full marks for parts (a) (b) & (c), but rather less answered (d) correctly. Conditional probability is a topic that many students still find problematical.

Most students know what a tree diagram looks like and at least drew the first 3 branches correctly with probabilities and labels. Whilst the majority also drew the second 6 branches correctly as well, a fairly small minority came up with some inaccurate probabilities on their branches. Perhaps the most common error on the second branches was to put the product of probabilities instead of the single probability. Some also had branches where the probabilities did not add up to 1.

Those with a correct tree diagram generally scored the mark in part (b). Again, in part (c), the majority managed at least the method mark even with an incorrect tree but most also got the correct answer.

Part (d) discriminated the most able students and was clearly the most challenging part of this question. Conditional probability is a topic that many find difficult, particularly obtaining a correct numerator. The most common mistake here was an attempt at P(S') × P(F) in the numerator.

Question 4

Students had the highest success rate on this question as again nearly half of students scored full marks here. Students were able to understand function notation for a discrete random variable very well and the routine calculations of E(X) and E(X^2) were also correct most of the time. In part (a) the majority gained both mark as students are well prepared for the idea that the sum of the probabilities is 1. Most gave a sufficient amount of working to show the given result.

It was part (b) that caused the greatest difficulty with a range of incorrect answers but $\frac{7}{9}$ being the most common. Many attempted a partial expectation calculation rather than a probability.

Parts (c) and (d) were nearly always correct though occasional slips meant loss of accuracy marks for some students. Those who neglected the instruction to find the exact value were penalised.

Whilst the majority gained all 3 marks in part (e), usual mistakes included forgetting to square the E(X) term in the calculation of Var(X) and, to a lesser extent, believing that Var(3X + 1) = 3Var(X) +1.

Question 5

Question 5 of this paper was significantly more demanding with only around 5% of students successfully achieving full marks. In part (a) most identified the width as 0.5 cm but, as usual, the height of the bar caused some difficulties as not all were able to use frequency density for the calculation. Some seemed unaware of the relationship between frequency and area, a few still seeming to think that the height represents the frequency.

Most realised the requirement for the data on a histogram to be continuous and scored in part (b) though quite a few students answered that it was a 'good representation of the data'.

Many found the correct mean in part (c) although in a small number of responses a less accurate mean of 1.4 was seen. The standard deviation was often correct but omitting the square root and hence leaving the variance as the final answer was again seen in a small number of responses. Common mistakes $\left(\frac{101.56}{48} - \frac{67.8^2}{48}\right)$ or $(101.56 - \bar{x}^2)$ were seen. Students should be aware that the formula for standard deviation is very sensitive to rounding errors and an accurate value for the mean (stored on their calculator) should be used rather than a rounded answer.

In (d) calculation of the median was well done with the correct fraction often being added to a correct lower-class boundary.

Part (e) proved to be a very good discriminating question with very few correct answers. It was poorly done with $\frac{31}{48}$ the most common incorrect answer.
Statistical reasoning in part (f) was less well attempted. Most stated that the mean would increase and often gave a correct reason for this but fewer knew that this would cause the standard deviation to decrease. A very common misunderstanding demonstrated by students was to say that the standard deviation would not change as it is not affected by addition.

Question 6

Whilst 20% of students made no progress on this question, another 20% went on to score full marks as once again the normal distribution is a discriminating topic on this specification.

In part (a), students were able to set up a pair of simultaneous equations in μ and σ on almost all occasions. Sometimes the probabilities of 0.9 and 0.05 were used in them, or 0.5199 and 0.8159. Of those who correctly used *z*-values from the second table (or from their calculator) in their standardisations, (–)1.6449 was seen much more often than 1.2816. Another common error was to use incompatible signs in the second equation. Students should be encouraged to draw a small sketch which should enable them to decide on whether the *z*-value is positive or negative. A minority of students incorrectly attempted to standardise with σ^2. The solving of simultaneous equations was done correctly by most students.

After finding the mean and standard deviation, students then needed to find the expected number from 23 patients who would be waiting longer than 12 minutes to see the doctor. In part (b), many knew how to find the probability of a given person waiting more than 12 minutes but finding the expected number often wasn't attempted.

Question 7

This was the most demanding question on the entire paper and only the most able students successfully calculated the probabilities required in parts (a), (b) and (c).

Those who did get all the marks in (a) usually went on to get full marks for the whole question. The key to part (a) was to understand that $\mathrm{P}(A \cap B) = 1 - \mathrm{P}(A' \cup B')$. This rarely happened and there were many unsuccessful attempts at the conditional probability resulting in numerous incorrect results for P(*B*). Weaker students incorrectly assumed independence and used $\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B)$. A sketch of a Venn diagram in part (a) would have helped out many who were struggling to rely on the given formulae.

Part (b) was also a challenging conditional probability where most students were adrift in the numerator but used correctly 1 – their P(*B*) in the denominator. Those who had a correct answer in part (a) almost always scored both marks in this part.

The method mark was generally achieved for the use of their (a) × 0.15 and again those who had a correct answer in part (a) scored both marks here.

Finally in part (d) students who drew 3 circles side by side with *B* intersecting both *A* and *C* were more successful than those who drew 3 overlapping circles, as they usually left blanks rather than writing zeros in the two sections where $\mathrm{P}(A \cap C) = 0$. Many scored the second mark by inserting their answer from part (c) into $\mathrm{P}(B \cap C)$ and then ensuring that P(C) had a total probability of 0.15. The third mark for having the diagram such that the probabilities in event A added up to 0.5 was often gained. Most students were able to complete the Venn diagram with their own probabilities such that all of the probabilities summed to 1 and it was pleasing to see that in virtually all cases a box was drawn.

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