Examiners' Report
Principal Examiner Feedback

Summer 2022

Pearson Edexcel International Advanced Level
In Statistics S1 (WST01) Paper 01

There were opportunities on this paper in all the questions for all candidates to make some progress but questions 2(f), 5 (e)(f) and 6(c) proved to be more challenging. The questions requiring a comment or explanation in words were often not answered very well and sometimes not even attempted. Candidates often assume independence when an appropriate conditional probability should be used instead.

Question 1

(a) Candidates generally were successful in finding the values of x and y, but the majority struggled at finding $w = 8$ for the first mark. Common mistakes included: obtaining an answer of 28 and not recognising that w was just the units part of it; calculating w as the median (26); or a combination of the 2 and finding $w = 6$. It was also quite common for students to not find a value for w at all.

(b) This part scored very well for most candidates. A minority lost the accuracy mark by failing to state the outliers having correctly calculated the outlier limit. A small number of candidates opted to apply the more commonly used outlier formula of $Q_3 + 1.5 \times (Q_3 - Q_1)$ rather than the one quoted in the question and some also calculated a lower limit for outliers which was unnecessary here.

(c) The diagram tended to be well drawn with most scoring all the marks, in some cases following through any errors from the outlier calculation. On some occasions there was no lower whisker drawn and sometimes it was extended well below the smallest data point.

(d) The substitution was handled well. Almost all candidates scored both marks for the calculation.

(e) Very few candidates scored this mark with quite a few just saying because it was a positive skew or attempting to discuss outliers but stated that the median was affected by the outliers. A very common misconception was that the mean is more accurate than the median. A large number of responses simply explained how to calculate a mean or said that it was because the mean is the average, showing no appreciation that the mean is just one measure of average and the median is another. Those who did score the mark here either used the 'mean uses all the data' or 'mean includes the outliers'.

Question 2

(a) The statistical calculation of the product moment correlation coefficient is clearly a topic which is well understood by the vast majority of candidates with full marks being achieved in a high number of cases. The most common error concerns inaccuracy due to candidates rounding off their calculator answers to fewer than the required 3 significant figures without stating a more accurate answer first.

(b) A large number of candidates failed to contextualise their answer by merely stating 'positive correlation' or 'strong correlation'. It is important for candidates to realise that their responses should clearly state the variables when they describe the relationship between them.

(c) This part of the question was again well attempted by the majority of candidates. Occasionally for the calculation of the value of a some used the given sigma values rather than the two mean values. In a small number of cases a and b were found but the final equation of the line was not stated, hence losing the final mark. It is also important to note that fractions are not accepted in a final regression equation; the values of a and b are both estimates and so a fractional answer is not appropriate.

(d) The question asked for an interpretation and again this indicates that a comment in context is required. Only the most able candidates successfully managed to write that the GDP increases by 31.2 billion dollars for every 1 million increase in the population.

(e) Here most candidates correctly substituted $t = 7$ into their regression equation and many gained both marks even with slightly inaccurate regression equations. A significant number of candidates were confused by the units and substituted $t = 7\ 000\ 000$ into the equation. It was rare to see responses which accurately assessed the reliability of the estimate found. There were inaccurate or incomplete statements used. Many stated that the estimate was unreliable, but they were unable to refer to the correct variable or value that is not in the range. It was also no surprise to find responses saying that the estimate was reliable, even with correct working earlier in the question.

(f) The final part of this question also discriminated the most able candidates. The most common mistake was to substitute $g = 0.1$ into the whole regression equation and solving for x. It was rarely identified that it was the rate of change that was the relevant part of the regression equation and so what was required was to find the value of x by solving $0.1 = 31.2x$. A few candidates used the whole equation and found the correct value of x by the substitution of two values for g with a difference of 0.1.

Question 3

(a) Most candidates were able to find the correct width of the new bar using the given information. As is common there was more of a problem in finding the height, dealing with both the change in frequency and width. A common incorrect answer was 9.33… as the result of scaling the height alone using frequencies but ignoring the change in class width. The successful candidates either scaled the height using frequency densities or scaled the area of the bar using frequencies.

(b) Finding the median by linear interpolation is well known by most candidates and there were many correct answers. Although most added to the lower class interval there was a minority who took a different approach (often helped by a sketched number line) working down from the upper class interval.

(c) Many candidates did not spot they could simply add half the 3rd class frequency to the class frequencies for shorter logs. The majority who gained the mark set up full interpolation equations but generally sufficient correct working was shown to gain the mark.

(d) This part was not answered well with only a minority realising the situation was sampling without replacement and scoring both marks. The most common response was scoring 1 mark for the special case using $\left(\frac{62}{88}\right)^4$. A small minority had their numerators decreasing (62, 61, …) but kept the denominator as 88, whilst others cancelled $\frac{62}{88}$ first and then found $\frac{31}{44} \times \frac{30}{43} \times \ldots$ Some candidates rounded to 2dp directly from the fraction product losing the final accuracy mark.

(e)(i) When attempted, most candidates were able to score marks here. Finding the correct mean was quite successful with candidates decoding correctly, although there were a number misreading 255 in the coding as 225, leading to a fairly common incorrect answer of 471. A small number attempted to decode using the $\sum y$ value, usually without success as they omitted the corresponding '×88' needed with the assumed mean of 255.

(e)(ii) Although the variance of *y* was often found, correct answers for variance of W in part (ii) were much rarer, with candidates not decoding, commonly including the assumed mean of 255, scaling using 0.5 instead of 0.5^2, or knowing they should use 0.5^2 but multiplying instead of dividing.

Question 4

(a) This part was fairly well attempted with many candidates clearly setting out their working using appropriate probability notation and showing all stages leading to the given answer. A number of candidates tried to use the given information to prove the given information and were unable to earn more than the first two marks in this part.

(b) Many candidates realised that they should divide by $\frac{3}{8}$ but were not quite sure what the numerator should be. It was concerning, as always, to see answers greater than one with no hesitation or doubt shown. Common errors were to assume independence or to do $1-\frac{1}{15}$ before dividing by $\frac{3}{8}$.

(c) Even those making little or no progress in parts (a) and (b) were able to make a start at this part of the question. Some candidates left out the 0 in the outside region of *N* and should be reminded that blank spaces are not assumed to be 0s in Venn diagrams.

Question 5

(a) This was a very accessible start to this question for the overwhelming majority of candidates.

(b) This part was also very well attempted but many candidates lost both calculating the variance instead of the standard deviation. Other errors include: dividing the 15.8 or 3.7^2 by 5 in the calculation; omitting the power of 2 on the 3.7.

(c) Well answered overall. The most common incorrect answer was 0.6.

(d) Well attempted by most, and for those that recognised the need to set up 2 equations in *a* and *b*, it almost always led to full marks. However, many candidates scored the 2nd M mark but failed to write down the equation for the sum of probabilities = 1 for the first M mark, which meant they could not solve for *a* and *b* and therefore couldn't proceed to find *c*.

(e) This was perhaps one of the most challenging parts of the paper with very few realising that the product of 3 probabilities was required. The majority multiplied only 2 probabilities and calculated $0.1 \times 0.9 = 0.09$.

(f) Again, a discriminating part which was well attempted but the vast majority but rarely fully correct. The most common incorrect answer of 0.195 came after failing to realise that if Jessie spins a 2, the game stops and Pabel doesn't spin.

Question 6

(a) Most candidates responded well to this familiar type of question on the normal distribution with many scoring full marks in this part. Some candidates found the left hand tail but forgot to subtract from 1 to find the right hand tail. There was a smaller number who failed to score as they attempted to standardise using variance rather than standard deviation.

(b) The majority of candidates realised they simply had to multiply their answer to part (a) by 150 to find the expected number, but very many of these thought they must give an integer value, either rounding up to 4 or truncating to 3. Whilst this was not penalised, candidates should be made aware that ‘expected number’ is a mean and need not be a whole number. There were a small number of candidates unaware of the connection with part (a) and either left this part blank or attempted to start a new probability calculation.

(c) The final part of the paper was also the most discriminating part. Many candidates were able to form a conditional probability statement for the first mark but a fair number then had trouble establishing the numerator as $\mathrm{P}(v < V < 104.9)$, some instead having simply $\mathrm{P}(V > v)$, which if followed through would have led to $z = 0.603$ ... and a final common wrong answer of 101.51

Dealing with the double inequality was a problem for others who erroneously had
$\mathrm{P}(v < V < 104.9) = \mathrm{P}(V < 104.9) - \mathrm{P}(V > v)$.
This should have led to $z = -0.53$ and a final common wrong answer of 98.68, but some compounded their first error with an inconsistent sign error for their z-value, fortuitously leading to the correct answer from incorrect working.

Also, there were many instances of P(V < 104.9) – P(V < v) = 0.2801 as the first line, suggesting that the conditional probability was not understood.

The most able candidates scored full marks here.