Examiners' Report
Principal Examiner Feedback

Summer 2022

Pearson Edexcel GCE
In Statistics (9ST0)
Paper 02: Statsitical Inference

Summer 2022
Publications Code 9ST0_02_2206_ER

General Comments

This was the second full sitting of this new A level specification and candidates appeared to be much better prepared than in the first 2019 sitting. The paper was accessible to all the candidates and there was no evidence that they could not complete it in the time allowed.

Questions requiring candidates to describe or comment on their findings proved to be particularly discriminating at the top end. Work on the paired *t*-test also proved a discriminator.

The new assessment objective A03 requires students to assess the reliability and validity of statistical methodologies used, and the conclusions drawn through the application of the statistical enquiry cycle. As such, it is important that students can comment on reliability by stating, with a reason, whether or not a suggestion is appropriate, as in questions such as 1c.

Candidates should be reminded that the number of marks should indicate the depth of response required.

Report on Individual Questions

Question 1

This was generally well answered, with many of the candidates scoring full marks in part (a). There was evidence that candidates had been well prepared in terms of reading all of the given information before identifying the required test to carry out. Candidates quickly realised that the statement about the percentage errors having a skew distribution was a clue towards carrying out a non-parametric test. The data given was from two independent samples so the Wilcoxon rank-sum test was the required test.

Candidates were given full credit for ranking either the magnitude of errors or the given data values – both approaches were credited. The main errors in part (a) were caused by arithmetic errors, using the incorrect formula for *U* (it is worth reminding candidates that the formula for *U* can be found in the formula book) or treating the data as paired leading to differences between errors being found and then, incorrectly, a paired sign or Wilcoxon signed-rank test used. It was pleasing to see that the majority of candidates are generally giving their conclusions in the full context of the question.

Part (b) was less well answered with many candidates simply stating the general conditions for carrying out a *t*-test rather than relating their answer to the given context and the information given within the question.

Although part (c) was generally well answered, some candidates missed the fact that this question required an assessment and many did not include ‘I agree’ or ‘I disagree’ in their response. Where candidates decided that this suggestion would be suitable, some found it difficult to give an appropriate supporting reason.

Candidates should be encouraged to ensure their response fits the given number of marks in a question. Part (d) had two marks available: the first for realising that the required test was now paired, and the second for being able to identify either the sign or Wilcoxon signed-rank test as a suggested test. Candidates generally scored at least 1 mark in this part.

Question 2

Overall this question saw a mixed response. Candidates are reminded of the need to read through all of the given information carefully before deciding which type of hypothesis test to carry out. Many candidates failed to realise that the study in India was a large-scale study, so the statistics quoted could be treated as population values.

Part (a) required the use of an **exact** test to decide whether the sample provided evidence to support Hamish's suspicion. The question stated that Hamish suspects that the **proportion** of students who sleep for less than 5 hours a night is smaller in the UK than in India. The second bullet point given at the start of the question identified the proportion of students in India who slept for less than 5 hours a night as 20%. These were the clues to identify this as a test for a single proportion using the binomial distribution. Common errors included using a normal approximation to carry out the hypothesis test or attempting to carry out a hypothesis test for a mean using the value of 6.45 from the first bullet point.

Part (b) was well answered by most candidates, with the largest proportion mentioning that Hamish only used responses for one night.

Part (c) turned out to be one of the most challenging questions on the paper.

Where students were able to identify 0.626 as the proportion of poor sleepers in India, the majority did correctly go on to use either an exact binomial distribution or a normal approximation to conduct the test. The most common error was to attempt to conduct a hypothesis test for the difference between two binomial proportions, as there was no sample size given for India this should have been an indicator that this was not a sensible approach. Many candidates carried on with their two proportion hypothesis test simply deciding the sample size for India was the same as that for the UK or using 40 from part (a). Some credit was given when this approach was attempted.

Part (d) of this question was well attempted by the vast majority of candidates who successfully calculated the 95% confidence interval for the population mean PSQI score.

Some candidates realised that the whole confidence interval for the UK was totally within the confidence interval for India in part (e) and could then comment on there not being enough evidence to suggest a significant difference in the mean PSQI scores for students in India and the UK. A small number of these candidates who successfully compared the two confidence intervals were then unable to explain the relevance of this fact. Some candidates simply compared their sample means with each of their confidence intervals rather than comparing the two confidence intervals for students in the UK and India.

Question 3

In (a) it was rare to see full marks and many candidates did not seem to understand the relevance of the given *p*-values. Common errors included interpretations about the given *p*-values indicating a very small probability of posting pictures of people being active, rather than identifying that a very small *p*-value meant that such an extreme observed outcome would be very unlikely under the null hypothesis. Candidates were required to comment on evidence of there being a difference between the proportions of pictures of people being active on Instagram and on Flickr.

Candidates gained far more marks in parts (b) and (c) of this question.

In (b) the majority of candidates were able to interpret the effect size for each of the stated values for Cohen's *d* , as well as offer an overall interpretation. The specification gives a rough guide for these definitions.
In (i), as $d > 0.8$ this should be interpreted as a large effect size.
In (ii), as $|d| < 0.5$ this could be interpreted as a small or medium effect size.
A few candidates interpreted the negative Cohen's *d* values as indicating a very small effect size. The question did explain the relevance of a negative value of Cohen's *d* as this was the first time a negative Cohen's *d* value had been included in an examination paper. Most candidates were also able to interpret two Cohen's *d* values in the context of the question.

Part (c) was also a source of at least two marks for the top end of candidates. Attention should be drawn to the wording of the question, asking for an overall summary 'for a reader with limited statistical knowledge' indicating that non-technical language should be used in their summary. Comments referencing the *p*-values or values of Cohen's *d* could not score full marks in this part of the question.

Question 4

Part (a) of this question was well answered in general with the majority of students earning at least 6 of the available 11 marks. Candidates had clearly been well prepared for this style of question and were able to carry out a two-factor ANOVA test. Those students who made use of the statistical functions on their calculators were often the most successful, although it is worth noting that showing the divisions for the mean sum (MS) and final *F*-ratio calculations will enable method marks to be awarded even when there has been an error in the calculator input. Encouraging a quick manual check of the degrees of freedom is also advisable. Common errors when using the formulae were mistakes in the denominators for the SS row and SS column calculations. Most candidates spotted the requirement to compare the test statistic for rows and the test statistic for columns with the corresponding critical value for *F* leading to two separate conclusions in terms of oat variety and fertiliser concentration.

In part (b), candidates generally scored at least one of the two available marks for stating the assumptions necessary to carry out a two-factor ANOVA test. Although full context was not required in this part of the question, this does not mean it will not be a necessity in future series.

In part (c), most candidates could correctly identify Merlin as the recommended variety due to the highest total (or mean) oat yield. However, very few candidates made the connection between their conclusion in part (a), that there was no significant evidence of a difference between mean oat yield due to fertiliser concentration, and the required additional comment that there was no specific advice based on preferable fertiliser concentration.

Question 5

This question was another challenging question for the many candidates.

Candidates should be encouraged to take their time when deciding whether or not the data given is paired. The given data in Figure 4 had three ratings for each of the jokes.

In part (a) the jokes were all numbered 1 to 10, clearly indicating that the results for 'no laughter' and 'fake laughter', and for' no laughter' and 'real laughter' were paired. A common error was to compare the two given test values (t=1.93 and t=3.51) together rather than finding the critical value using the t-distribution with 9 degrees of freedom.

In part (b), a common error was to treat the given ratings as two independent samples and to carry out a hypothesis test on two means using a pooled estimator of the variance. (The advance information stated paired tests but made no reference to a hypothesis test for two means in this paper).

Candidates who calculated the differences between the ratings for fake and real laughter generally went on to score full marks in this part of the question. However, some candidates used a critical value of t =1.833 rather than the two-tailed value of t =±2.262.

It was pleasing to note that the vast majority of candidates clearly compared their test statistic and critical value, or obtained the correct critical region or p-value, and went on to state their conclusions in the full context of the question.

Question 6

This question was a standard contingency table question. Over 50% of candidates were able to score the full marks available in part (a) demonstrating that they were able to recognise and analyse a contingency table with ease. Many candidates appeared to be familiar with the use of the statistical functions on their calculator to obtain the test statistic and had clearly been encouraged to write down the contributions to this test statistic, thus enabling method marks to be gained even when a value had been mistyped. For those candidates who did not earn full marks, the most common mistakes were incorrectly combining rows due to the observed value of 2 for 'no' and probably yes', giving hypotheses the wrong way around, ‘an association’ for H_0 (rather than ‘no association’) or obtaining an incorrect test statistic, possibly due to mistyping a value into their calculator.

Candidates found part (b) more challenging. Candidates were specifically asked to give numerical justification to describe the nature of any association identified in part (a), yet few did. This is a relatively standard application of a chi-squared test and candidates should expect this type of question. Those candidates who had stated the contributions in part (a) were the most successful in part (b) although not all were able to fully explain the nature of the association in context.

Part (c) was generally well answered with many candidates being able to give at least one source of bias in the investigation. Candidates should be encouraged to return to the beginning of a question and highlight exactly what information has been given about how an investigation has been carried out before starting to formulate an answer to this style of question.

Summary

Based on their performance on this paper, candidates should be advised to:

- read all of the given information in a question carefully and fully before answering the question. In particular, look at all the given information before deciding on the most appropriate hypothesis test to carry out.
- look out for words shown in **bold** type.
- write conclusions to hypothesis tests in terms of ‘**evidence of**’, rather than stating a definite conclusion.
- keep their working to more than 3 significant figures accuracy, only rounding their final answer.
- use **bullet points**, each written in clear, specific, and concise sentences for explanation questions.

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