Examiners' Report
Principal Examiner Feedback

Summer 2024

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

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 2024
Publications Code 9ST0_02_2406_ER

Principal Examiner Feedback Report 9ST0 02

General Comments

This paper was the fourth full sitting of the new A level specification. Throughout the marking it was shown that the paper was accessible to all candidates and that the paper could be completed within the time allowed.

It is important to make candidates aware that working in the tables will be seen and marked by examiners. Therefore, they do not need to use time in the exam by rewriting the tables to then rank or calculate expected values etc.

It is important for students to be able to comment on reliability and validity to acquire the A03 marks in this specification. There were several opportunities in the paper to do this, it was pleasing to see more candidates answering these questions and giving reasons to support their choices.

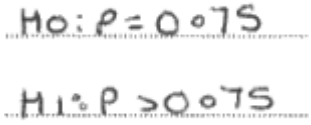
Candidates should continue to be reminded that **explanation answers should be given in the context of the question**. This comment is raised regularly in examiner feedback and can cost candidates several marks across a paper. For instance, statements such as 'the sample is random' will not gain credit in this specification as there is a focus on candidates answering the question within the context of the data given.

With regards to hypotheses, it is important that candidates know the difference between hypotheses for 1 sample, 2 sample and paired tests. Lack of understanding on this point can see candidates regularly penalised for an easily obtainable mark. Also, if a candidate is not using the required Greek letter for the test, then there is the expectation that the word 'population' is used in the hypotheses instead, there were several instances of candidates just rewriting the wording of the question which is not adequate to gain credit.

Throughout the paper the final method mark was only awarded for a correct comparison of the candidates test statistic and critical value so long as a **reasonable** attempt at a **reasonable** hypothesis test had been attempted.

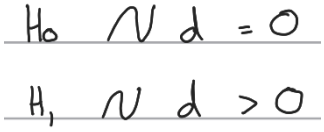
Handwriting is becoming a growing concern when it comes to marking. It is important to raise two key issues that were seen in the 9ST0 02 paper this year:

Example 1:



A few candidates were seen using a small circle as a decimal point, as shown. This example is not particularly bad as it was known what value was required. However, there were occasions when this small circle was larger and appeared as another 0 which in this course will cause difficulty due to the need to use decimals frequently across all papers.

Example 2:



Greek letters need to be easily distinguishable in hypotheses. This example is from question 1 on the paper where the test uses μ in the hypotheses, however you could equally argue that this could be an η. This will be something that could be penalised in the future so it is important that this is highlighted to candidates to avoid losing their hypotheses marks.

Question 1

It was pleasing to see that many candidates used the prompt in Q1(a) to carry out a reasonable attempt at a paired *t* test. Candidates should be aware that there will be a mixture of named and unnamed hypothesis tests in the 9ST0 02. Hypotheses were marked generously in this case, but should technically be of the form H_0: $\mu_d = 0$ as this is a hypothesis test on the differences.

A common misconception was whether the value of 0 is included in the test statistics and critical value calculations. It is important that candidates are aware that the **0 is included in a *t* test**. Those who did not include it lost accuracy and the final explanation mark.

Another common error was hypotheses that did not agree with their test statistic values. In these cases, it was the hypotheses mark that was uncredited to allow candidates to gain more marks later on in the question.

Although the hypothesis test was named in the question, it was frequent that candidates chose to treat this as 2 sample *t* test instead. In this case candidates could achieve a maximum of 3 marks for a 'correct' attempt at this hypothesis test.

Part (b) was attempted by the majority of candidates. Good exam technique was shown by candidates **looking back at the information provided in the question** and using this as the way to form their answers, this is the best way to gain full credit on questions of this style as there will always be enough information provided within the question.

As mentioned in the general comments, context is required on these explanation answers and should be encouraged whenever a candidate is responding in writing in this specification. It is important for candidates to be aware that their responses must be **specific and in context**.

A common error in exam technique was not commenting on what affects bias, which is what the question specifically asked for. Therefore, answers of the following form did not gain credit as they were not specifically answering bias, more the reliability of the result.

- Placebo effect
- Other factors such as sleep or wearing glasses.
- Correlation does not imply causation.
- Sample size is small.

Another error was candidates making assumptions about the sample of volunteers, for instance:

- Only testing on females
- All volunteers are of the same age

We are not told this in the question so we cannot give credit for these responses.

Question 2

This question proved to be more challenging for some candidates. It was good to see students recognised the need for the normal approximation to be used. However, in some cases this stopped candidates from attempting the question. Although the question required a normal approximation for full credit, partial credit was given for using the exact binomial test for single proportions. In terms of exam technique, it is **better for candidates to conduct an appropriate test** which may not be exactly what was asked for but gain some credit, rather than leaving their response blank.

It was particularly pleasing to see a significant number recognised that they needed to calculate the test statistic as 249 themselves. With regards to the use of a continuity correction, it was not required in this mark scheme, though many candidates did find this. However, there were a number of responses where the incorrect correction of 249.5 was given instead, these candidates only lost the final explanation mark.

A common misconception was using $\frac{249}{312}$ to calculate the mean and variance for the normal approximation that was being used. Several candidates calculated $\frac{249}{312}$ and then rounded it to be 0.80 this **early rounding then affected later accuracy marks** in the question.

With regards to hypotheses, this is a test on a proportion, therefore the hypotheses should be written as such. Although I understand that testing for $H_0: \mu = 234$ is comparable, the **hypotheses should reflect what is being asked in the question**.

Also, in terms of hypotheses, as mentioned in the general comments, candidates should be using the relevant hypotheses for the test. Therefore $H_0: p = 0.75$ (or $p = 0.25$ when using the test statistic as 63) are what would be expected. Not $H_0: p_{adults} = p_{under\ 18s}$, again although it does show an understanding of the question these are incorrect hypotheses for a single sample test.

Question 3

Again, the contingency table method has proved to be accessible to many candidates.

Part (a) was attempted by almost all candidates and the majority scored these marks. Those who didn't obtain the final mark either: incorrectly rounded or left their answers to more than 1 decimal place. Candidates are reminded that answers should be to **3 significant figures, unless stated otherwise**.

In part (b) many candidates gained credit. However, those who didn't were on the right lines but were **not specific in their answers** for instance 'values are less than 5', 'expected values are small'. We did not penalise 'observed and expected values are less than 5' however, this could be something that we look at penalising in the future as observed values do not affect the appropriateness of using the chi squared test.

The completion of the hypothesis test in (c) was done well by many candidates using a variety or methods. For those who are using the calculator it is important that they **write down the test statistic contributors**. This is important so that method marks can be awarded if errors have been made in the calculator work, but also so that contributors can be identified for association questions which may follow. Those who showed no working and obtained an incorrect test statistic automatically lost 4 of the 8 marks for this part of the question.

As mentioned in last year's report the main errors of candidates not gaining full credit were for hypotheses being reversed or selection of the incorrect critical value.

The most challenging part of the question was (d). As already mentioned for Question 1(b) good exam technique was shown by candidates **looking back at the information provided in the question** and using this as the way to form their answers, this is the best way to gain full credit on questions of this style as there will always be enough information provided within the question.

Context is required on these explanation answers and should be encouraged whenever a candidate is responding in writing in this specification. It is important for candidates to be aware that their responses must be **specific and in context**.

A common error in exam technique was not commenting on what affects bias, which is what the question specifically asked for. Answers as shown below did not gain credit as they were not specifically answering bias, more the reliability of the result.

- Families may have answered the questionnaire together.
- Increasing the sample size
- Adding more supermarkets

Other candidates stumbled on either only giving sources of bias or only improvements rather than ensuring their answer contained both aspects. Another common issue, again as seen in Question 1(b), was making assumptions about how the data was gathered and the sample; for instance, 'the questionnaire could be next to an advert for a specific supermarket'.

The reasoning for the explanation marks needed to show more understanding of the context. For instance, 'sample other towns so it's more varied/representative' would not be enough to get the explanation mark for the source of bias.

'Introduce a blocking factor' was a very common incorrect answer for (d), as was mentioned in last year's report.

Question 4

One of the more challenging questions on the paper. Though it was pleasing to see many candidates attempting a response.

For (a) it is important that candidates using hypotheses like $H_0: p_A = p_B$ that the subscripts are clearly identified in their method. Hypotheses like $H_0: p_L = p_S$ gained full credit as it was identifiable by the context.

Of those who correctly found a test statistic of 0.460 there was a worrying number who then assumed this was a *p*-value and not a standardised *z*-value. Although not required in the specification, they should have a **basic understanding that a normal approximation has been performed in this test**, hence the use of a standardised *z*-value (or use of inverse normal to find critical value in the alternative method). This misconception lead those using a standardising method to compare 0.460 against the significance level and automatically lose the final 2 marks.

In (b) the question was assessing the candidate's ability to communicate findings effectively. **Follow through marks** could be achieved for this part (even without completing the test in (a)) it is therefore prudent to encourage candidates to use any information that they can to answer

these parts (those who just compared the proportions directly were awarded credit for a correct explanation in context).

Some candidates got confused about wording, using number instead of proportion although this was permitted in this series.

The audience mark was penalised for use of hypothesis test, test, quoting of the test statistic or critical value, use of fractional proportions, words such as significant/insufficient etc.

Although several candidates missed the opportunity for follow through marks on (b) it was pleasing to see that candidates did attempt (c) and (d).

Answers for (c) were technically correct but candidates lost marks for a lack of context e.g. 'sample is selected randomly' 'businesses are independent' did not gain credit.

Part (d) was generally answered well with understanding shown that one country within in the United Kingdom could not be used to be representative of all nations. Again, the sample being small/large was a common incorrect answer.

Question 5

The largest question on the exam paper, for the second series in a row.

As in Question 1, the hypothesis test was **specifically named in the question**. However, it was surprising the number of candidates who then proceeded with a different test (most commonly the sign test). Again, if the test is named then they will lose marks for conducting an alternative, particularly the sign test.

It was pleasing to see many candidates obtain almost full marks in (a) by using the space provided to find their test statistic and critical values. As mentioned in the general **comments candidates can write some working in the question**, hence the space that was given under the sample data.

As has been mentioned consistently throughout this report candidates need to be using **appropriate hypotheses for the samples given**, in this case we should have had $H_0\ \eta = 3.82$ as we are performing a single sample test.

If marks were lost, they were for: selecting the wrong critical value from the formula booklet or calculating values for U using the formulas for Wilcoxon Rank Sum.

Those candidates who had more precise exam technique picked up on the need to 'State' the assumption for the test, though **many were not specific with the context** required saying vague answers such as 'skydivers are symmetric'.

As a reminder if the question states '**Making** any appropriate assumptions' the candidates are **not required to write them down**. If the question states '**State** any necessary assumptions' the candidates **are required to write them down**, using the context of the question.

Part (b) was answered well by the majority of the students. Those who did not acquire the mark were due to repeating the information they were told, stating that they were two different readings but not explaining that this could therefore mean 2 different measures for cortisol. Another common incorrect response was saying that people may not produce enough

sweat/saliva so cortisol measurements would be lower, cortisol would still be present even in a small amount of sample.

Part (c) highlighted the need for candidates to **know the required assumptions for each test**, not only so they can state them if required but also so that they may pick an appropriate test based on information provided. In this case not having an underlying bivariate normal distribution meant the only appropriate test was a Spearman's rank correlation coefficient test. Those who spotted this performed very well on these 6 marks.

The final part of this question proved to be challenging for many candidates. As the question states 'Based on the result of the hypothesis test in (c)…' all responses had to refer to what had been found in (c) and could not refer to information given previously in the question (or their own opinions).

Those who did realise they needed to refer to the conclusion of the test, some focused on the idea of correlation does not mean causation rather than the misconception of Miku's that correlation meant that sweat and saliva had equal cortisol levels.

However, there were candidates who had not attempted/gain full credit on part (c) but did obtain all the marks in (d) for spotting Miku's misconception.

Question 6

Part (a) was answered well by the majority of candidates. Though they should be reminded that the **formulas for pooled variance and Cohen's *d* are given in the formula booklet**, so there is very little leeway given in marking with regards to incorrect substitutions. Common errors included: forgetting to use the variance in the pooled variance formula (not squaring their standard deviations) and not square rooting the pooled variance in the denominator of the calculation of Cohen's *d*.

Unfortunately part (b) was left blank by several candidates. Although Cohen's *d* is one of the later topics in the A-level content, there are some easily accessible marks for learnt terminology and methods.

As mentioned in last year's report, candidates who were the most successful on this question **broke their answers up** by dealing with the interpretation of the hypothesis test first and then interpreting the value of Cohen's *d* separately.

When interpreting the *p*-value, several candidates decided to assume the significance level as 10%. If a significance level is not given in the question, they are to take this to mean it is 5%, they cannot choose to select another value to ensure a significant result. Another common error with the significance level was just to say it's close to 0.05 therefore small enough to reject H_0, this showed a lack of appreciation for the level of accuracy required in this specification.

An unusual misconception, which has not appeared before, was comparing the *p*-value against the Cohen's *d* found in (a), this obviously gained no credit.

As mentioned in last year's report, when interpreting the value of Cohen's *d*, many students used incorrect terminology. The specification states the following:

21.2 Know and use Cohen's *d* in simple situations.	Students should be aware of the standard guideline boundaries for interpreting the value of Cohen's *d*. $0.2 \leq d < 0.5$ small effect size $0.5 \leq d < 0.8$ medium effect size $0.8 \leq d$ large effect size

This mark was given as a follow through for their calculated values in (a).

For the final mark candidates were expected to write 2 separate statements (one for the interpretation of the *p*-value and one for Cohen's *d*) that contained context as well as the word mean/average, it was not enough to see this in only one of their statements.

Although not required in this series, it was excellent to see several candidates picking up on the possible contradiction between the hypothesis test and Cohen's *d* and also stating how they would rectify this. This style of question was seen in the specimen papers and is allowable on the specification.

Question 7

Although the final question with a particularly difficult final three marks, it was excellent to see many candidates have realised that they should keep going to the end of the paper and **attempt all questions** set as there were some accessible marks.

Alongside this as general exam technique, candidates should understand that a question, and the paper itself, **may not always be in difficulty order**. For instance, there have been a few occasions in this particular paper where more accessible marks were found later on in a question as well as later on in the paper.

In (a) the first mark was awarded for evidence that the student understood that a FILTER function would be required, the second mark was for stating 'Branches/Column D' to 'Manchester/Liverpool'. Some candidates went on to describe how to find relevant means etc. this was not required as the question 'to help **select** appropriate data'. There were several candidates who answered this as a database rather than as a spreadsheet as specified in the question.

It was pleasing to see that the majority of candidates were not perturbed by the unusual way summary statistics were given in (b) and did attempt a 2-sample *z* test correctly. A few candidates did not recognise that the samples were both large and performed the *t* test version, this gained a maximum of 4 marks out of 8. Again, a few lost marks for hypotheses where **subscripts were not clearly identified** as well as comparing their standardised *z* test statistic against the significance level.

In (c)(i) the candidate was asked to 'Explain why…' many candidates just defined what a Type II error was which is not what was asked in the question. It was also fantastic to see several candidates who had Rejected H_0 in (b) go back to the previous part, find and correct their error based on the information in (c).

Part (c)(ii) was answered very well by the majority of candidates where clear statements were given about the conclusion of the test **and** the error made within the context of the question.

The final question on the paper proved to be one of the most challenging. Unfortunately, many candidates left it blank as they were not prepared for power/P(Type II error) to appear on this style of question. Of those who did attempt it many did not do enough to gain credit. In terms of teaching this content in the future the mark scheme should provide suitable methods for both the standardising method as well as the alternative method for *z* and *t* hypothesis testing.

To clarify the specification allows power and P(Type II error) to be found on any hypothesis test that is based on a *z* or single proportion test. Therefore, this could be asked for the following test on this specification:

- Single proportion test (exact binomial)
- Single z test
- 2 sample z test (unpaired z test)
- Normal approximation for the proportion test

Summary

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

- Give explanation answers within the context of the question.
- Look out for words shown in bold type.
- Consider the structure of an answer to make it clear issues and improvements, such as Q3d.
- Ensure that they identify key terminology given in the question to ensure answers are detailed enough e.g. 'explain why…', 'to reduce bias',
- Multiple attempts at answers will be marked separately, averaged, and rounded down. The examiner will not just mark the correct response.
- If a specific hypothesis test is named, the candidate is expected to carry out this test. Alternative responses will automatically lose at least half of the marks for the question.
- If candidates are using subscripts that are not based on the context of the question, then they must be stated or easily identified by labels in the question.
- Early rounding can affect accuracy marks later in an answer, encourage candidates to store values within their calculators.
- Over rounding was seen more frequently this series. Candidates are reminded that unless stated in the question they should be leaving answers to 3 significant figures.
- Those candidates who use the standardising method are to be reminded that their critical value must be clearly stated in their responses and to not rely on the examiner knowing what value they have chosen.

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