Examiners' Report
Principal Examiner Feedback

Summer 2025

Pearson Edexcel Advanced Level
In Statistics (9ST0)

Paper 2: 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 2025
Publications Code 9ST0_02_2506_ER

General Comments

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

It is important to make learners 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.

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

Learners should be reminded that explanation answers should be given in the context of the question. This comment is raised regularly in examiner feedback and could cost learners several marks across a paper. For instance, statements such as 'the responses may not be independent' will not gain credit in this specification as there is a focus on learners answering the question within the context of the data given.

With regards to hypotheses, it is required that learners know the difference between hypotheses for 1 sample, 2 sample and paired tests. Lack of understanding on this point can see learners unable to achieve an easily obtainable mark. Also, if a learner 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 learners just rewriting the wording of the question which is no adequate to gain credit.

Throughout the paper the final method mark was only awarded for a correct comparison of the learners test statistic and critical value so long as a reasonable attempt at a reasonable hypothesis test had been attempted. On several occasions those learners solely relying on calculator outputs were not credited this mark as there was no evidence of a reasonable attempt. There were also a few instances seen where learners wrote down a set of hypotheses and immediately went to a conclusion without an attempt at finding a test statistic or critical value, in these cases the candidate can only score a maximum of 1 mark for their hypotheses.

Follow through marks appear on several occasions in the 9ST0-02 paper this series. For instance, as mentioned above a reasonable attempt at a hypothesis test can allow learners to receive the final method mark. There were also opportunities in 3(b), 5(d) and 6(c) where learners could receive full credit even if they had made an error in previous parts of the question. Therefore, it is very important that learners are actively encouraged to attempt all parts of a question so that they can awarded the follow through marks which are available.

It is important to note that responses of the form 'the sample is outdated' are unlikely to be credited. The only time where this would be considered as an appropriate answer would be if the candidate gives a specific and valid reason as to why the sample being from that period may not give the full picture. For instance, using data on glacial melting from 100 years ago, the candidate could say that using this outdated research is unreasonable due to the impact of climate change over the last century.

Question 1

Learners answered part (a) well. Common errors that occurred were when learners made mistakes with the maths trying to find relevant multipliers to get from 1.2 to 2.0, which was not what the question was looking for.

Part (b) was attempted by the majority of learners. However, a number of learners appeared to be unsure about the use of summary statistics in the question. It is important learners are comfortable with using the formulas provided in the formula booklet; as well as selecting the right formula for samples and populations.

The most common incorrect response was to treat the question as a one tailed test. Learners should recognise that when there is no specific change given i.e. increase/decrease then the test should be treated as two tailed.

A significant number of learners made errors with the conclusion of the test. Those who wrote their response with regards to the reporter's claim often made an error by saying that the reporter's claim is true. It is often the case that the alternative hypothesis represents the claim/point of investigation, however, for this question the claim that the final was 2.0 was represented by the null hypothesis. Learners should make sure they have practice of cases like this so they do not fall into assuming that 'Reject H_0' automatically implies a claim is correct.

Question 2

The Wilcoxon signed-rank sum test is often a well performed question in a series and this year was no exception. Those who recognised that the data was unpaired and picked up on the phrase 'distribution free test' performed well on this question. Many learners annotated the information provided showing a good statistical process in their recognition skills (which are often the most challenging to teach).

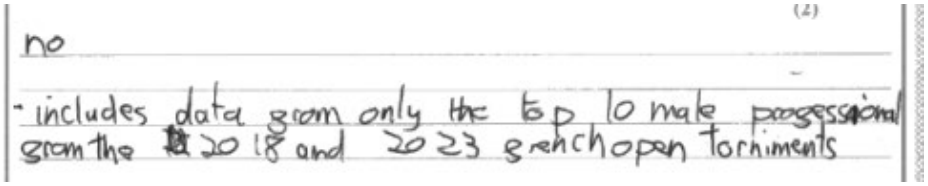
A few learners misread the critical value as they subtracted one from each sample (like a *t* test). Again, there were instances of learners rewriting the original table to then rank it; as a reminder this is not necessary, and learners should be encouraged to use the table provided in the question where possible.

There were a number of learners who completed the question as a Wilcoxon Signed Rank test. Although this was an incorrect method, partial credit could be awarded. Several responses stated that equal sample sizes implied paired data (more on this point in question 4a). It should be stressed that paired data requires a variable to connect the 2 pieces of data to each other. At no point in the question was such a variable identified and therefore the sample could not be treated as such.

Part (b) was answered very well and most learners provided their responses in the context of 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 learners to be aware that their responses must be specific and in context.

One key point for improving exam technique is shown in the response attached



Responses of this nature could not gain credit as the candidate has not identified why this is an issue they have simply copied a phrase from the question. Questions are designed so that the information is easily accessible for the learners, as we do not expect every learner to be an expert in all of the context given in our papers. Because of this we do expect them to explain in their own words why only selecting Top 10 male professionals is an issue for generalising the results.

Another reasonable common response which gained no credit was to say that the sample information is out of date. As mentioned in the general comments section this response will only be given where appropriate and when backed up with a valid reason for why the gap in the data may cause issues with generalisation.

Part (c) was accessible to the vast majority of learners with the most common responses being related to before 1989 and outside of major tournaments. A few learners gave an incorrect response on 'a woman may have served faster' confusing the information with the context earlier in the question.

Question 3

The majority of learners recognised this to be a single sample proportion test and although learners were asked to complete the question using a critical region method partial credit was given for those who completed the question using *p*-values instead. Those who conducted a normal approximation or 2 proportion test could only gain credit for hypotheses (and possibly quoting the correct binomial distribution).

As mentioned in the general comments, learners should be using the relevant hypotheses for the test. Therefore $H_0: p = 0.574$ are what would be expected. Not $H_0: p_{2013} = p_{2017}$, again although it does show an understanding of the question these are incorrect hypotheses for a single sample test.

There were also learners who relied solely on the use of the 'Inverse Binomial' function on their calculators. This was given full credit. However, it should be noted that we can require that all learners show the full method of 'searching' for the critical value. Therefore, it is important that learners know this method for future series so they do not rely on this function of their calculator to complete the question but can use it to check their method instead.

In part (b) follow through marks were given for the probability of a Type I error, so long as the value had come from a critical region method using the binomial distribution (and not the approximation). In a few cases learners had performed the correct test in (a) but stated the probability as 5%. The explanation in context was completed well by a large proportion of learners, those who missed out on this mark either gave a Type II error explanation or did not explain the error as a one tailed test and said phrases such as change/different.

Part (c) did not require any explanation as to why the power increased and so long as the candidate's explanation did not include any incorrect statements it was ignored. However, there was a significant number that used phrases such as 'the test will be more accurate/reliable'. This is a common misconception about what power represents and is not acceptable. Examiners cannot always ignore incorrect statements so it is important that learners are made aware that their statements are succinct and correct.

Question 4

The majority of learners recognised the test correctly for this question. The unequal sample sizes implying an unpaired sample was highlighted and recognised by a significant number of learners. Though there were still several who used a z critical value method instead or created a paired design be either adding in values of 0 or crossing out the final 2 values for the female sample.

The formula for pooled variance is given in the formula booklet, so there is very little leeway given in marking with regards to substituting standard deviations into the formula rather than the variance.

Again, it was common for those who relied solely on calculators to do the majority of the hypothesis test for them were more likely to not gain the marks for this question. For example, several learners quoted simply a probability with no method to show that they had needed to pool the variances and, in these cases, marks would not be awarded. We encourage the use of calculators but we do expect to see more than just a set of hypotheses and a p-value for learners to show that they understand the hypothesis test being examined.

Part (b) proved to be a more discriminating question, with many answering it well. There were many instances of learners discussing assumptions around the collection or the sample rather than the distributional assumptions as stated in the question. A common incorrect response for (b) was to say that 'women/men are normally distributed' the variable being measured in this case was time so we needed to see this **to credit the context required.** Also, several learners simply said that 'variances are unknown'.

The majority of learners were able to pick up at least one mark in (c), though some did not have distinct responses, therefore were unable to gain the other marks. As shown in the mark scheme it was deemed that only using long haul flight/only using one airline are not distinct so were only credited once. Another common reason learners did not gain marks was for writing standard answers with no explanations or context. For instance, 'responses may not be independent' as mentioned previously we expect the learners to explain why they believe this in their own words. There were also several instances of learners writing gender stereotypes for instance stating women packed for children and men didn't pack for themselves.

Question 5

This was the first time the uniform distribution appeared in the context of a Goodness of Fit test. Because of this part (a) was introduced to test their understanding of why this distribution was suitable for the context of the question. The first mark was easy to obtain as simply stating $\frac{1}{8}$ was the minimum required to gain this mark. Many learners also explained how the shape of the distribution had been arrived at.

It was fantastic to see so many learners used the information in the table provided to complete the final steps of the χ^2 test, though a few did get the conclusion the wrong way round when dealing with the double negative.

As has been said previously for the χ^2 test learners should show the calculation for at least one test statistic contributor. If a candidate reaches an incorrect test statistic value by using their calculator they are at risk of not gain any of the method marks that go towards the calculation.

As is often seen with the goodness of fit test the main source of inaccuracy was learners multiplying the probability by the individual group frequency and not the total frequency. In these cases, we could get M and B marks but not the A or E marks.

Part (c) was identified as one of the more challenging questions but many learners had a good and reasonable attempt at what some called the 'preference sign test'. This was the first time it had appeared in the assessment of the specification so it was encouraging to see that learners were familiar with the method.

For the first mark and accepted quite a large array of hypotheses. In situations like this where medians and probabilities do not make contextual sense those learners who wrote the hypotheses out in words often were the most successful. The most common mistake on this question was either forgetting to halve the significance level or assuming the significance level was 5%.

In the final part of the question this was assessing the candidate's ability to communicate findings effectively. Although it was challenging to gain full credit many candidates achieved 3 marks on this part. Follow through marks could be achieved for this part (even without completing the tests in (b) and (c)) it is therefore prudent to encourage learners to use any information that they can to answer these parts (those who just compared the expected/observed values directly or the proportion of disagreements in quality assurance were awarded credit for a correct explanation in context).

Several learners were confused by what their answer in (a) told them and how it supported evidence for Matt's promotion. Some learners, even after stating Matt had failed one of the targets due to earlier errors, then suggested Matt keep his promotion, even though the question clearly stated he needed to meet both targets.

This question also required the learners to summarise their findings using non-technical language. Some examples of phrases and words which would lose the learners the final mark were:

- $\frac{1}{8}$ or 0.125
- Significant/sufficient/insufficient evidence
- Hypothesis test
- Uniform/Rectangular
- Distribution
- χ^2
- Expected/observed values

Question 6

The ANOVA was answered well by most. Some learners did not gain the marks for their hypotheses as they either did not label the table of identify their subscripts when not using T, M or H. Another common issue with hypotheses was phrases such as 'At least two differ', credit was only given if learners referred to means/average or effect (this was also the case in their final conclusion marks.

Those learners who relied solely on the use of a calculator often had an all or nothing attempt at this question. As was highlighted in previous examiner reports it is important that learners show they understand where the values in their ANOVA table and do not just quote a *p*-value as their only working.

Several learners also conducted a one factor ANOVA test which saw them lose a considerable number of marks. Though it was pleasing to see that some learners conducted a one factor ANOVA and then having seen the question in (c) went back and modified their answer to a two factor ANOVA instead.

There were quite a few cases where learners decided to test both factors in (a), in this case this working was ignored. However, this could be seen as two attempts at the same test, therefore learners should only be carrying out a test on the factor identified, and only both when instructed to do so (as seen in 9ST0-02 in 2022).

The majority of learners managed to pick up at least one mark for 6(b). For the first mark we allowed them to just use the totals as a way of comparing the three font styles as the sample sizes were equal. However, since this is a test on the means the appropriate method is to calculate the sample means as a way of distinguishing between the different groups. Some learners stated Traditional with no numerical justification and therefore received no marks as this was a requirement stated in the question.

The final question of the paper was answered well by those learners who had completed (a) successfully. Though it was good to see that many learners who made a calculation error in their test still were able to pick up all 3 marks as they were all given as follow through for their previous work. However, it was required that the working and method was seen in (c) and therefore could not be awarded if the candidate relied on the fact they had carried out both factor tests in (a).

Quite a number of learners completed the question successfully to conclude to 'Reject H_0' but either did not state what this meant with regards to the effectiveness of the blocking factor or reached the incorrect conclusion.

Summary

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

- Give explanation answers within the context of the question.
- When commenting on generalising results and reliability learners should reword responses and explain why certain information needs to be considered e.g. women have not been included
- Look out for words shown in bold type in the question.
- 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 commonly not gain at least half of the marks for the question.
- If learners 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.
- Those learners 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.
- Learners relying on the use of a calculator for their answers will need to show more method than a *p*-value to gain credit if they make an error.