Pearson Edexcel

Pearson Edexcel
Level 3 Advanced GCE in Statistics (9ST0)

May/June 2022 Exemplar
9ST0-02 A level Statistics
Paper 02: Statistical Inference
Senior Examiner's feedback on student responses

Contents

About this booklet 2

How to use this booklet 3

General Examiner Feedback 4

Question 1 5

Question 2 20

Question 3 36

Question 4 46

Question 5 58

Question 6 69

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

© Pearson Education Ltd 2023

Skip to Main Contents

About this booklet

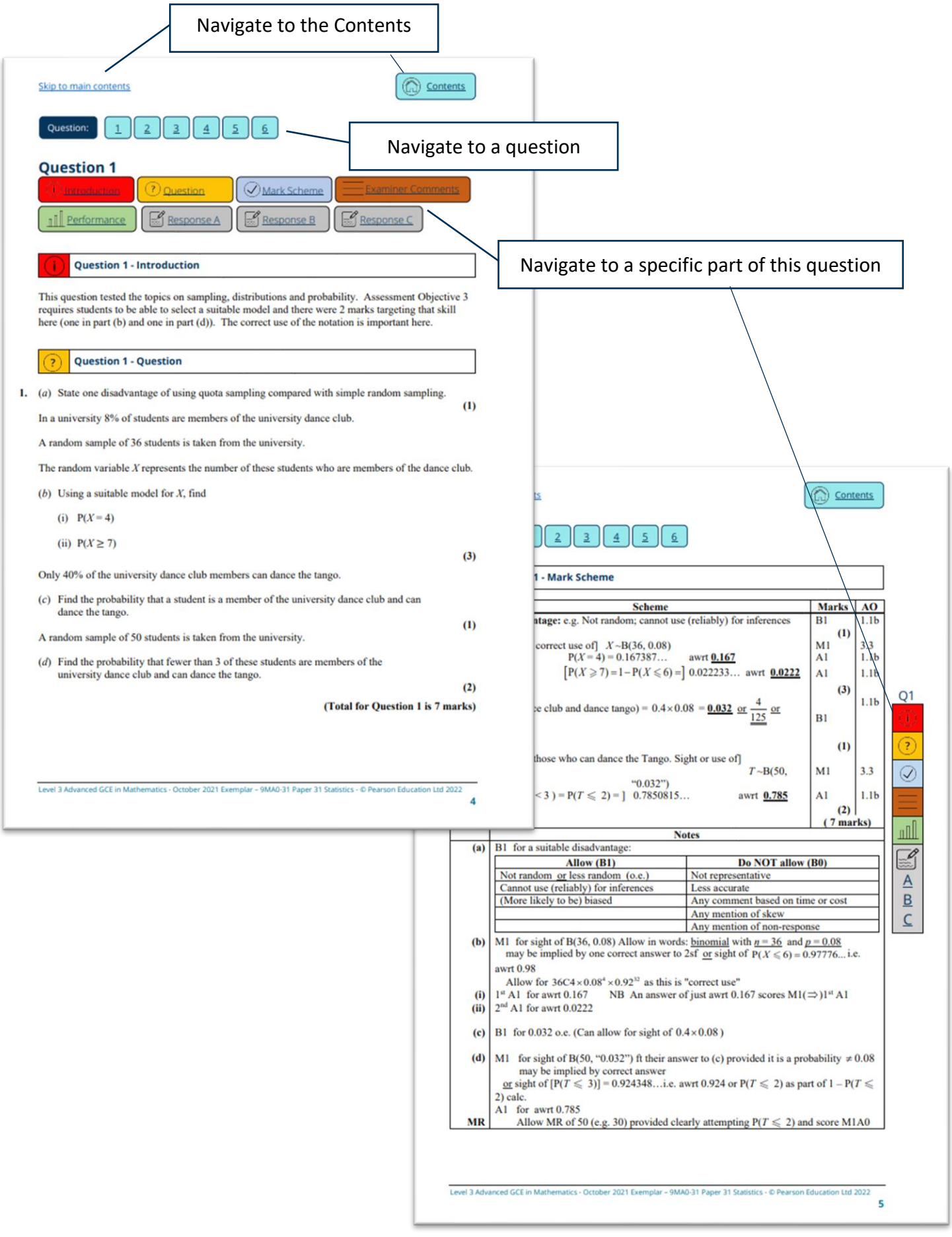
This document has been produced to support mathematics teachers delivering the GCE Statistics specification.

This document looks at questions from the 9ST0-02 A level Statistics Paper 2: Statistical Inference June 2022 examination paper. It shows real student responses to these questions, and how the examining team follow the mark schemes to demonstrate how the students are awarded the marks.

* The question level performance data is there to give an indication only of how students performed, on each question, in the context of sitting the entire exam paper and is not an indication of how students may perform sitting a question in isolation.

Skip to Main Contents

How to use this booklet



General Examiner Feedback

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.

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.

Question 1

Question | Mark Scheme | Examiner Comments | Performance | Response A | Response B | Response C

Question 1 - Question

1 The breaking strength of cables has usually been assessed by testing samples of cables until they break (testing to destruction).

A manufacturer of carbon-fibre-reinforced polymer (CFRP) cables wants to replace testing to destruction with a method for predicting breaking strength based on known characteristics of a cable.

The methods to be compared are a computer simulation model and a mathematical model.

The investigation involves selecting a sample of 18 CFRP cables. Nine of them have their breaking strengths predicted by a computer simulation model and nine by a mathematical model.

Each cable is then tested to destruction to determine its actual breaking strength.

The following are the percentage errors made in the predictions for each model.
A negative error means that the prediction is too low.

Previous research suggests that percentage errors such as these have skew distributions.

Computer model	Mathematical model
–7.5	4.4
–3.2	8.3
–3.3	8.2
–5	6.9
–1.5	10.1
–0.6	10.9
0.7	12.2
4.8	15.9
5.2	15.9

[Data source: https://doi.org/10.1590/1679-78254177]

(a) Perform a non-parametric test to decide whether there is evidence of a difference in the average prediction error of these models.

(9)

Khalid, a junior researcher, instead suggested evaluating the methods by comparing mean percentage error using a *t*-test.

(b) Make **two** comments on the suitability of this suggestion.

(2)

Lauren, the chief engineer, suggests that, to make the comparison, it would have been better to run both the computer and mathematical models using the same cables, and then testing each to destruction.

(c) State, with justification, whether or not you agree with Lauren's suggestion.

(2)

(d) Suggest a test that Lauren could use to compare the two methods.

(2)

(Total for Question 1 is 15 marks)

Q1

Question 1 - Mark Scheme

Qu	Scheme	Marks	AO	Notes
1(a)	[Assuming distributions are the same shape]			
	(Ranked for magnitude of errors)			
	H_0: no difference in (population) medians H_1: difference in (population) medians	B1	1.3	oe in words (samples from identical populations) or using $\eta_C = \eta_M$
		M1	2.1a	Clear attempt at ranks
	Rank (Comp) 11 4 5 8 3 1 2 7 9 **Rank (Math)** 6 13 12 10 14 15 16 17.5 17.5			
		M1	1.3	Attempt at totalling their ranks **or** at least 9 correct ranks.
		A1	1.3	All ranks correct (ignore ties) **or** either total correct.
	$T_C = 11 + 4 + \cdots + 9 = 50$ $T_M = 6 + 13 + \cdots + 17.5 = 121$			
	$U_C = 50 - \frac{1}{2}(9 \times 10) = 5$	M1	1.3	Attempt at U_C or U_M, using their totals and their n

Qu	Scheme	Marks	AO	Notes
	$U_M = 121 - \frac{1}{2}(9 \times 10) = 76$			
	ts = 5 (or 76)	A1	1.3	**or** 76
	[For 2-tail test, $\alpha = 0.05$] cv = 18 (or 63)	B1	1.3	**or** 63
	$"5" < "18"$ so reject H_0	M1	2.1b	**or** $76 > 63$ Comparison of their ts and their cv (dep use of WRS) in **same tail**.
	There is significant evidence (at the 5% sig. level) that… …the median prediction **error** for the computer model is **different** from the median prediction **error** for the mathematical model.	E1dep	2.1a	dep on ts & cv both correct Must be in context (at least difference and error/model) and conclusion not definite. **or** …that the computer simulation is more accurate.
	(Ranked using given data)			
	H_0: no difference in population medians H_1: difference in population medians	B1	1.3	oe in words (samples from identical populations) or using $\eta_C = \eta_M$
		(M1)		Clear attempt at ranks
	Rank (Comp) 1 4 3 2 5 6 7 9 10 **Rank (Math)** 8 13 12 11 14 15 16 17.5 17.5			
		(M1)		Attempt at totalling their ranks **or** at least 9 correct ranks.
		(A1)		All ranks correct (ignore ties) or either total correct.
	$T_C = 1 + 4 + \cdots + 10 = 47$ $T_M = 8 + 10 + \cdots + 17.5 = 124$			
	$U_C = 47 - \frac{1}{2}(9 \times 10) = 2$ $U_M = 124 - \frac{1}{2}(9 \times 10) = 79$	(M1)		Attempt at U_C or U_M, using their totals and their n
	ts = 2 (or 79)	(A1)		**or** 79
	[For 2-tail test, $\alpha = 0.05$]	(B1)		**or** 63

Qu	Scheme	Marks	AO	Notes
	cv = 18 (or 63)			
	"2" < "18" so reject H_0	(M1)		**or** 79 > 63 Comparison of their ts and their cv (dep use of WRS) in **same tail**
	There is significant evidence (at the 5% sig. level) that… …the median prediction **error** for the computer model is **different** from the median prediction **error** for the mathematical model.	(E1dep)		**or** …that the computer simulation is more accurate. Must be in context (at least difference and error/model) and conclusion not definite dep on ts & cv both correct
1(b)	[The *t*-test would be unsuitable as…]			
	…the distribution of percentage errors cannot be assumed to be normal …			**or** distribution of percentage errors is skew
	…and the sample may be too small for the Central Limit Theorem (or CLT) to apply.			Need some reference to CLT
	(Population) variances may not be equal.			
		E1, E1	3.1a, 3.1a	E1 for each comment (Max E2) context not required.
1(c)	I agree…	E1dep	3.1a	dep on reasonable effort at explanation
	…because any differences between the cables used for each model would be accounted for.			oe
	…reduces experimental error			oe
	…so that any difference found is due to the model rather than the cable.			
		E1	3.1a	For explanation
	I disagree…	(E1dep)		dep on reasonable effort at explanation
	…because the cables used for each test could become damaged which could have an effect on the following results. **or** …that the computer simulation is more accurate.	(E1)		oe

Qu	Scheme	Marks	AO	Notes
1(d)	Analyse with **paired**…	E1	3.1a	
	…**Wilcoxon signed**-rank test.	E1	3.1a	**or** sign test
	Total	**15**		

Question 1 - Examiner Comments

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.

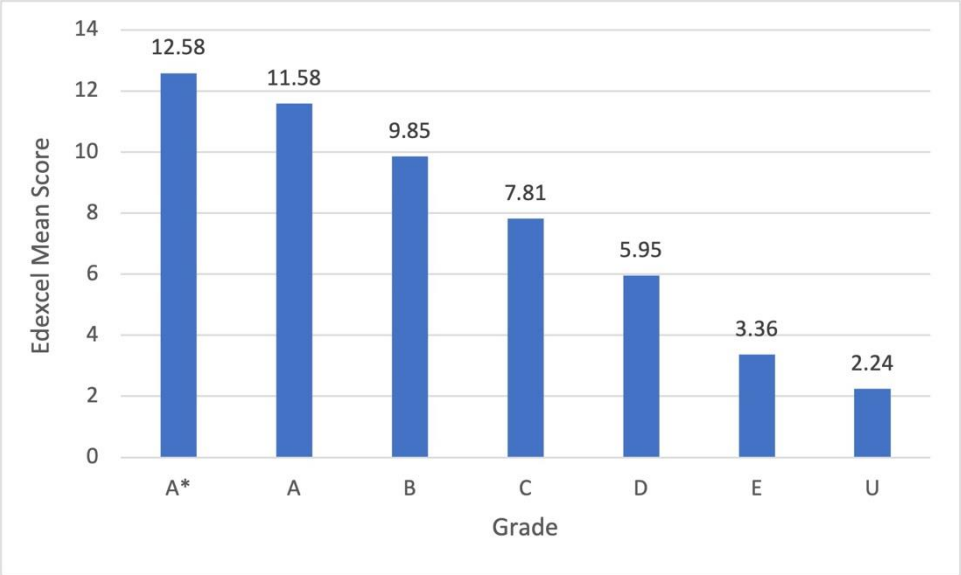
Part (b) was less well answered with many candidates simply stating the general conditions for carrying out a *t*-test rather than relating their answer to the given context and the information given within the question.

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 1 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
7.94	15	53	7.94	12.58	11.58	9.85	7.81	5.95	3.36	2.24



Question 1 - Response A

Computer model	Mathematical model	
–7.5	4.4	+11.9
–3.2	8.3	+11.5
–3.3	8.2	+11.5
–5	6.9	+11.9
–1.5	10.1	+11.6
–0.6	10.9	+11.5
0.7	12.2	11.5
4.8	15.9	11.1
5.2	15.9	10.7

[Data source: https://doi.org/10.1590/1679-78254177]

Question 1 continued

(a) Perform a non-parametric test to decide whether there is evidence of a difference in the average prediction error of these models.

(9)

Wilcoxon signed rank

differences
+11.9
+11.5
+11.5
+11.9
+11.6
+11.6
+11.5
+11.7
+10.7

Ho: difference in average percentage error of models
H1: No difference in average percentage error of models

$\bar{d}$ = mean
$\bar{d}$ = 11.46
sd = 0.374

p(x

Question 1 continued

Khalid, a junior researcher, instead suggested evaluating the methods by comparing mean percentage error using a *t*-test.

(b) Make **two** comments on the suitability of this suggestion.

(2)

This is not a suitable test as it is not a non parametric

A test & test use used to compare standard derrations

Lauren, the chief engineer, suggests that, to make the comparison, it would have been better to run both the computer and mathematical models using the same cables, and then testing each to destruction.

(c) State, with justification, whether or not you agree with Lauren's suggestion.

(2)

I think this is a better comparison as you are able to find the difference in percentage error of each individual cable

(d) Suggest a test that Lauren could use to compare the two methods.

(2)

Wilcoxon signed rank test

Q1

2 /15

Part (a)

B0: These hypotheses are incorrect; the hypotheses have been stated in words, but they are the wrong way around. Reference to the population medians rather than just referring to the averages is preferable.

M0: The given data has not been ranked.

M0: There are no ranks and therefore they have not been totalled.

A0: Following M0, also no ranks or totals given.

M0: There is no attempt at either U_C or U_M.

A0: Following M0.

B0: No critical value has been attempted.

M0: There is no attempt to compare a test statistic with a critical value.

E0: No conclusion has been given and the correct test statistic and critical values have not been stated.

Part (b)

This question required the consideration of whether a *t*-test could be used to evaluate the methods by comparing mean percentage error. The required answer was to consider whether a test for two independent normal distributions with unknown variances was valid rather than a generic response referring to when the t-distribution can be used.

E0: Incorrect comment.

E0: Incorrect comment.

Part (c)

E1: There is a reasonable attempt at an explanation, even though this explanation is not correct.

E0: The given explanation is not sufficient, we wanted to see a reference to Lauren's suggestion reducing experimental error or leading to any difference found being due to the model rather than the cable.

Part (d)

E0: The word 'paired' has not been stated.

E1: Correctly stated the Wilcoxon Signed-Rank test, this mark would also be awarded for sign test.

Question 1 - Response B

Computer model		Mathematical model	
−7.5	1	4.4	8
−3.2	4	8.3	13
−3.3	3	8.2	12
−5	2	6.9	11
−1.5	5	10.1	14
−0.6	6	10.9	15
0.7	7	12.2	16
4.8	9	15.9	17.5
5.2	10	15.9	17.5

[Data source: https://doi.org/10.1590/1679-78254177]

Question 1 continued

(a) Perform a non-parametric test to decide whether there is evidence of a difference in the average prediction error of these models.

(9)

$H_0: \eta_d = 0$ $T_c = 47$

$H_1: \eta_d \neq 0$ $T_m = 124$

$47 - \frac{9 \times 10}{2} = 2$

$n = 2$

$124 - \frac{9 \times 10}{2} = 79$

$CV = 21$ $2 < 21$

Reject H_0

There is sufficient evidence to suggest a population median difference in the average prediction error of computer model and mathematical model.

Question 1 continued

Khalid, a junior researcher, instead suggested evaluating the methods by comparing mean percentage error using a *t*-test.

(b) Make **two** comments on the suitability of this suggestion.

(2)

would be suitable as the sample sizes are small. Also the population standard deviation is unknown.

Lauren, the chief engineer, suggests that, to make the comparison, it would have been better to run both the computer and mathematical models using the same cables, and then testing each to destruction.

(c) State, with justification, whether or not you agree with Lauren's suggestion.

(2)

Lauren suggestion is correct as this would make the data paired which is much more reliable than unpaired data.

(d) Suggest a test that Lauren could use to compare the two methods.

(2)

[illegible] signed ranks test

Paired sign test

9 /15

Q1

Part (a)

B0: These hypotheses are incorrect, this is not paired data, so are not finding a difference between each pair of data values. Hypotheses written in words or using $\eta_C = \eta_M$ would have been accepted.

M1: The given data has been ranked (ranking of either the magnitude of errors or the given data is allowed here).

M1: The ranks have been totalled.

A1: All ranks are correct and at least one total is correct.

M1: There is an attempt at U_C (and also U_M). These are both correct although only an attempt using their totals and their *n* for this mark was needed for this mark.

A1: The answer for their U using this ranking method is correct, the mark was awarded for either 2 or 79.

B0: The critical value is incorrect (this is the one-tailed 5% value rather than the two-tailed 5% value from table 11 in the formula book).

M1: Their test statistic has been compared with their (incorrect) critical value.

E0: Although the conclusion has been given in context, this mark cannot be awarded as it depends upon both the test statistic and critical value being correct.

Part (b)

This question required the consideration of whether a *t*-test could be used to evaluate the methods by comparing mean percentage error. The required answer was to consider whether a test for two independent normal distributions with unknown variances was valid rather than a generic response referring to when the *t*-distribution can be used.

E0: Incorrect, 'small samples' refers to the *t*-distribution in general rather than a test for two independent normal distributions with unknown variances, reference to the sample being too small to apply the central limit theorem was needed.

E0: Incorrect, population standard deviations unknown refers to the *t*-distribution in general rather than a test for two independent normal distributions with unknown variances. Reference to the two population variances not being equal was needed.

Part (c)

E1: There is a reasonable attempt at an explanation, even though this explanation is not correct.

E0: The given explanation is not sufficient, reference to Lauren's suggestion reducing experimental error or leading to any difference found being due to the model rather than the cable was needed. The given explanation just refers to the data being more reliable.

Part (d)

E1: The word 'paired' is stated.

E1: They have stated the sign test (this mark would also be awarded for Wilcoxon Signed-Rank test).

Question 1 - Response C

Computer model	Mathematical model
−7.5 (11)	4.4 (6)
−3.2 (4)	8.3 (13)
−3.3 (5)	8.2 (12)
−5 (8)	6.9 (10)
−1.5 (3)	10.1 (14)
−0.6 (1)	10.9 (15)
0.7 (2)	12.2 (16)
4.8 (7)	15.9 (17.5)
5.2 (9)	15.9 (17.5)

Rank

[Data source: https://doi.org/10.1590/1679-78254177]

Diff. Models = Not Paired.

(a) Perform a non-parametric test to decide whether there is evidence of a difference in the average prediction error of these models.

(9)

Wilcoxon rank sum test

H_0: $\eta_C = \eta_M$

H_1: $\eta_C \neq \eta_M$

5% 2 Tail

η_C = Pop. Median of Computer model

η_M = Pop. Median of Mathematical model

Q1

A B C

$T_C = 50$ $U_C = 50 - \frac{9 \times 10}{2} = 5$

$T_M = 121$ $U_M = 121 - \frac{9 \times 10}{2} = 76$

$TS = 5$ $5 < 18$ so reject H_0,

$CV = 18$ Sufficient evidence at 5% to suggest a difference in the average prediction error of these models.

Khalid, a junior researcher, instead suggested evaluating the methods by comparing mean percentage error using a *t*-test.

(b) Make **two** comments on the suitability of this suggestion.

(2)

- Not suitable as the methods are Not normally distributed, thus a t test is inadequate
- It could be suitable as the methods are independent of one another

Lauren, the chief engineer, suggests that, to make the comparison, it would have been better to run both the computer and mathematical models using the same cables, and then testing each to destruction.

(c) State, with justification, whether or not you agree with Lauren's suggestion.

(2)

I agree as running both models using same cables will ~~reduce~~ reduce experimental/error of other factors which could influence the test.

(d) Suggest a test that Lauren could use to compare the two methods.

(2)

A Wilcoxon A paired Wilcoxon signed rank test.

Q1

A B C

14 /15

Part (a)

B1: These hypotheses are correct using $\eta_C = \eta_M$ or the equivalent in words were both acceptable forms.

M1: The magnitude of errors has been ranked (ranking of either the magnitude of errors or the given data was acceptable)

M1: Their ranks have been totalled.

A1: All ranks are correct and at least one total is correct.

M1: There is an attempt at U_C (and also U_M). These are both correct although only an attempt using their totals and their *n* for this mark was needed.

A1: The answer for their U using this ranking method is correct, the mark is awarded for either 5 or 76.

B1: The critical value is correct.

M1: Their test statistic has been compared with their critical value. This comparison can be shown using inequalities or on a number line.

E1: The conclusion has been stated correctly and in full context, this mark can be awarded as it follows the correct test statistic and critical value.

Part (b)

This question required the consideration of whether a *t*-test could be used to evaluate the methods by comparing mean percentage error. The required answer was to consider whether a test for two independent normal distributions with unknown variances was valid rather than a generic response referring to when the *t*-distribution can be used.

E1: This mark was awarded for the comment about the methods not being normally distributed which was assumed to refer to the fact that the distribution of errors was stated to be skew. This mark could also have been awarded if there was any reference to the sample being too small to apply the central limit theorem.

E0: Reference to the two population variances not being equal was needed. The two samples being independent was not sufficient for this mark.

Part (c)

E1: There is a reasonable attempt at an explanation, so this first mark was awarded.

E1: The given explanation about reducing experimental error is sufficient.

Part (d)

E1: The word ‘paired’ is stated.

E1: They have stated the Wilcoxon Signed-Rank test (this mark would also be awarded for the sign test).

Question 2

Question | Mark Scheme | Examiner Comments | Performance | Response A | Response B | Response C

Question 2 - Question

2 A large-scale study of the sleeping habits of students in India was carried out.

Students in the study were required to fill out questionnaires about their sleeping habits, which were then used to give a score on a standard Pittsburgh Sleep Quality Index (PSQI).

A student with a PSQI score more than a given value was classified as a 'poor sleeper'.

Amongst the study's reported findings were

- the PSQI scores for students had mean 6.45 and standard deviation 2.85
- 20% of students slept for less than 5 hours a night,
- 62.6% of students were classified as 'poor sleepers'.

[Source: ncbi.nlm.nih.gov/pmc/articles/PMC4659578]

Hamish, a university researcher, is conducting a study to compare the sleeping habits of students in India with those in the UK.

He suspects that the proportion of students who sleep for less than 5 hours a night is **smaller** in the UK than in India.

To investigate this, Hamish asks 40 students at his university how many hours they slept the previous night.

He finds that 6 of them reported sleeping for less than 5 hours.

(a) Making any necessary assumptions, use an exact test to decide whether this sample provides evidence to support Hamish's suspicion.

(5)

(b) Identify **two** limitations of Hamish's study.

(2)

Hamish also obtains the PSQI scores for a randomly selected sample of 105 UK students.

He finds that 80% of these students can be classified as 'poor sleepers'.

(c) Conduct a test to investigate whether there is evidence for a difference in the proportion of 'poor sleepers' between students in the UK and those in India.

(6)

The PSQI scores for these 105 students have mean 6.48 and standard deviation 1.71

(d) Construct a 95% confidence interval for the corresponding population mean PSQI score.

(3)

The 95% confidence interval for the mean PSQI score for students in India was found to be (6.07, 6.83).

(e) Using this information and the confidence interval calculated in (d), comment on the mean PSQI scores of the UK and India.

(2)

(Total for Question 2 is 18 marks)

Question 2 - Mark Scheme

Qu	Scheme	Marks	AO	Notes
2(a)	$H_0: p = 0.2$ $H_1: p < 0.2$	B1	1.3	**Accept** π **Do not accept** $\hat{p}$ Condone $H_0: p_I = p_{UK}$ $H_1: p_I < p_{UK}$ oe well explained in words
	[X = number of students sleeping for less than 5 hours the previous night]			
	$X \sim B(40, 0.2)$	M1	1.3	PI Clear use of binomial distribution with n=40 and any p
	$P(X \leq 6) = 0.2859$	A1	1.3	0.285~0.286 **or** CR: X≤ 3 with p=0.028
	$"0.2859" > 0.05$ so do not reject H_0	M1	2.1b	PI Comparison of 'their *p*-value' with 0.05 **or** 6 > 3 (CR)
	There is no significant evidence… …that the **proportion** is **smaller** in the **UK** than in India.	A1dep	2.1a	**or** …to support Hamish's suspicion. Must be in context and conclusion not definite dep M1A1M1
SC: Use of normal approx. max 3/5 B1M1A0M1A0				

Q2

Hypotheses, model, comparison of ts with (-)1.645 or *p-value*=0.429 with 0.05				
2(b)	**Possible comments (not exhaustive)**			
	Only students at Hamish's university used.			
	Only responses for one night.			Might have been a party/exam time so not independent.
	Small sample [so low power test].			
	Students are self-reporting.			
	They might not know how long they slept.			oe
		E1, E1	3.1a, 3.1a	E1 for each sensible comment (Max E2)
2(c)	**Exact binomial method** H_0: $p = 0.626$ H_1: $p \neq 0.626$	B1	1.3	For both Accept π for p Condone H_0: $p_I = p_{UK}$ H_1: $p_I \neq p_{UK}$ oe well defined in words
	[X = Number of poor sleepers]			
	$X \sim B(105, 0.626)$	M1	1.3	Use of binomial with either $n = 105$ or p=0.626
	$P(X \geq 84)$	M1	1.3	**or** P(X≥ 76) = 0.0226 *or* 0.0227
	$= 0.000\ 09302$	A1	1.3	awrt 9.30×10^{-5} **or** 0.99996 or 0.99997 **or** CR: X≥ 76, (X≤ 55)
	< 0.025 so reject H_0	M1	2.1b	PI Comparison with 0.025 **or** 84 > 76
	There is significant evidence of a **difference** between the **proportion** of poor sleepers in the UK and in India.	E1dep	2.1a	In context, not definite. dep all 4 previous marks.

Q2

A B C

	Normal approximation method 1			
	H_0: $p = 0.626$ H_1: $p \neq 0.626$	(B1)		For both Accept π for p Condone H_0: $p_I = p_{UK}$ H_1: $p_I \neq p_{UK}$ oe well defined in words
	$z = \dfrac{0.80 - 0.626}{\sqrt{\dfrac{0.626 \times 0.374}{105}}}$	(M1)		PI Use of 0.626 **or** use of 84 and 65.73 condone use of 83.5
		(M1)		PI Dividing by their appropriate standard deviation.
	$= 3.68(49)$	(A1)		awfw 3.5~3.7 **or** CR: $\bar{X} \geq 75$ (implies previous M1M1 too)
	"3.68" > 1.96 so reject H_0	(M1)		PI Comparison of 'their' ts with 1.96 **or** p=0.00011 < 0.025 oe **or** $\hat{p} > 0.7185$ required **or** $\bar{x} > 75.45$ required
	There is significant evidence of a **difference** between the **proportion** of poor sleepers in the UK and in India.	(E1dep)		In context, not too definite dep all 4 previous marks.
SC Two proportions test B1M1M0A0M1E0 3/6 max 3rd M1 for their ts comparison with 1.96				
2(d)	$[\bar{x} = 6.48, s = 1.71, n = 105]$			
	$6.48 \pm (1.96) \times \dfrac{1.71}{\sqrt{105}}$	M1	1.3	PI Use of $\sqrt{105}$ or 0.1669
		B1	1.3	PI Using $z = 1.96$ **or** $t_{104} = 1.98(3)$
	CI is (6.15, 6.81)	A1	1.3	awfw 6.14~6.15

Q2

A B C

2(e)	This CI for the UK (6.15, 6.81) is completely within the 95% CI for India which is (6.07, 6.83)…	M1ft	2.1b	For correct comparison for both ends could be seen on a number line ft their CI for the UK as long as consistent
	…so there is no significant evidence of a difference in the mean PSQI scores for students in India and the UK.	E1dep	2.1a	oe In context, not too definite. dep previous M1 and correct CI in (d)
	Total	**18**		

Question 2 - Examiner Comments

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.

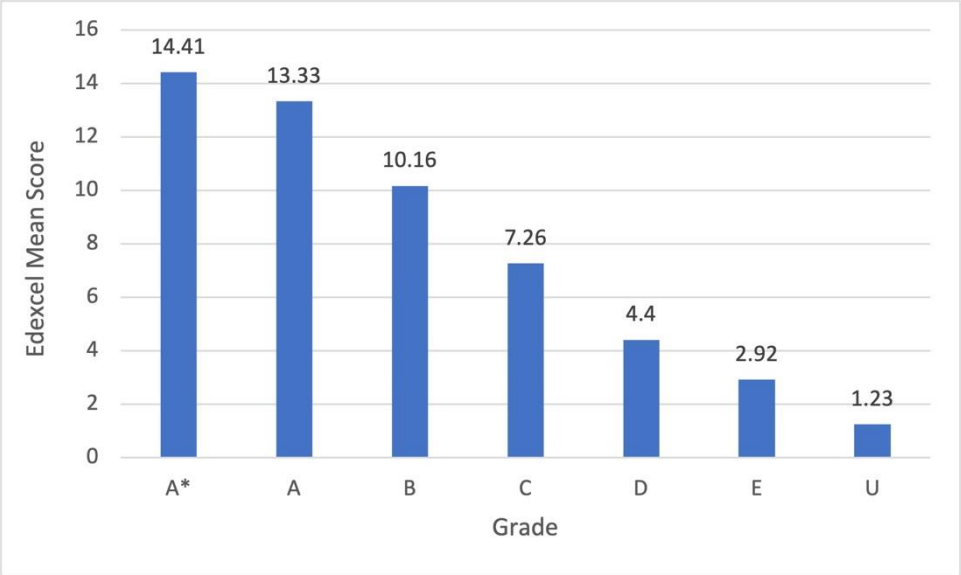
Where students were able to identify 0.626 as the proportion of poor sleepers in India, the majority did correctly go on to use either an exact binomial distribution or a normal approximation to conduct the test. The most common error was to attempt to conduct a hypothesis test for the difference between 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 2 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
7.80	18	43	7.80	14.41	13.33	10.16	7.26	4.40	2.92	1.23



Question 2 - Response A

(5)

$H_0: P_{UK} = P_{India}$
$H_1: P_{UK} < P_{India}$

$\frac{6}{40} \times 100 = 15\%$

mean $= 40 \times 0.15 = 6$
Var $= 40 \times 0.15 \times 0.85 = 5.1$

$UK \sim B(6, 5.1)$

Question 2 continued

(b) Identify **two** limitations of Hamish's study.

(2)

may be external influences effecting how much students can sleep.
there was to be a bigger boost
All students have to be chosen at random
There was to be either a odd or even group
school may start at different times

Hamish also obtains the PSQI scores for a randomly selected sample of 105 UK students.

He finds that 80% of these students can be classified as 'poor sleepers'.

(c) Conduct a test to investigate whether there is evidence for a difference in the proportion of 'poor sleepers' between students in the UK and those in India.

(6)

$H_0: P_{UK} = P_{India}$
$H_1: P_{UK} = P_{India}$

mean $= 105 \times 0.8 = 84$

$UK \sim B(84, 16.8)$

var $= 16.8$

Q2 ? A B C

Question 2 continued

The PSQI scores for these 105 students have mean 6.48 and standard deviation 1.71

(d) Construct a 95% confidence interval for the corresponding population mean PSQI score.

(3)

n large $\Rightarrow Z$

$6.48 \pm 1.96 \times \frac{1.71}{105} = (6.448, 6.512)$

area: 0.975

σ: 1

μ: 0

$= 1.96$

The 95% confidence interval for the mean PSQI score for students in India was found to be (6.07, 6.83).

(e) Using this information and the confidence interval calculated in (d), comment on the mean PSQI scores of the UK and India.

(2)

On average students in the UK have a better PSQI as their CI is more narrow, however there are individual students in India who have a much better PSQI score.

3 /18

Part (a)

This question asked for an exact test to be used to decide whether this sample provides evidence to support Hamish's suspicion. Candidates should be reminded to refer to earlier parts of the given information to identify Hamish's suspicion. 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 at the beginning of the question tells us that the study's reported findings were that 20% of students slept for less than 5 hours a night (in India).

B1: This form for the hypotheses is condoned. $H_0 : \pi = 0.2, H_1 : \pi < 0.2$ (using p or π) was preferred, although $H_0 : \pi_I = \pi_{UK}, H_1 : \pi_I < \pi_{UK}$ was condoned.

M0: This mark was for the use of X~B(40, 0.2). Candidates should also be encouraged to state the model being used. If a value had been stated in their model this method mark could have been awarded, however 5.1 is not a value for p.

A0: No attempt to calculate a probability using the binomial distribution

M0: There is no comparison of 'their p-value' with 0.05

A0: This mark depended upon the previous M1A1M1 so could not be awarded even if a conclusion had been written in context.

Part (b)

E1: This mark was awarded for the comment about their being possible external influences affecting how much students can sleep.

E0: No further relevant comments.

Part (c)

B0: The hypotheses have been stated incorrectly. The use of $H_0 : p_I = p_{UK}, H_1 : p_I \neq p_{UK}$ was condoned, however these have both been written using =.

M0: The correct model has not been stated.

M0: There is no attempt to find the relevant probability or to use the normal approximation to find the test statistic.

A0: Neither test statistic has been calculated (using either the exact binomial distribution or the normal approximation).

M0: No comparison has been shown.

E0: Depended on the previous 4 marks being awarded.

Part (d)

M0: The use of $\sqrt{105}$ rather than just 105 was needed for this mark.

B1: For the use of 1.96

A0: The answer is incorrect.

Part (e)

M0: A comparison of the two confidence intervals either in words or on a number line was needed for this mark.

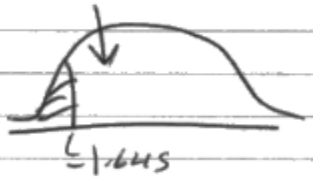
E0: This mark depends upon a comparison of the two confidence intervals.

Question 2 - Response B

H_0: The ~~proportion~~ of Indian ~~Uk~~ students who sleep for less than 5 hrs = the ~~proportion~~ of ~~Indian students~~ Uk students who sleep for less than 5 hrs. (5)

H_1: The ~~proportion~~ of Indian students who sleep for less than 5 hrs < the ~~proportion~~ of Uk students who sleep for less than 5 hrs

$\hat{p} = \frac{6}{40} = 0.15$ (1T, 5%)



$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.15 - 0.2}{\sqrt{\frac{0.2(0.8)}{40}}} = -0.791$$

$k = -1.645$

Accept H_0

Evidence suggests that the proportion of Indian to Uk students sleeping poorly is equal

Question 2 continued

(b) Identify **two** limitations of Hamish's study. (2)

He ~~is~~ only asked from his university and not multiple.

The PSQI score was calculated for Indi.

Hamish also obtains the PSQI scores for a randomly selected sample of 105 UK students.

He finds that 80% of these students can be classified as 'poor sleepers'.

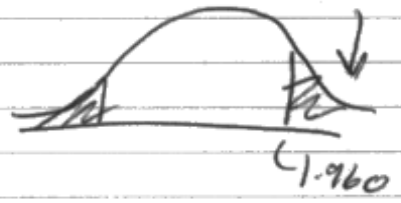
(c) Conduct a test to investigate whether there is evidence for a difference in the proportion of 'poor sleepers' between students in the UK and those in India.

(6)

$H_0 = p_I = p_{UK}$ $H_1: p_I \neq p_{UK}$ (2T, 5%)

p_I = India's proportion

p_{UK} = UK's proportion $k = 1.960$



$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.8 - 0.626}{\sqrt{\frac{0.626(1-0.626)}{105}}} = 3.685$$

Reject H_0

Evidence suggests that there is a difference between the proportion of UK's poor sleepers and India's proportion.

Question 2 continued

The PSQI scores for these 105 students have mean 6.48 and standard deviation 1.71

(d) Construct a 95% confidence interval for the corresponding population mean PSQI score.

(3)

$\mu = \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$

$6.48 \pm 1.645 \times \frac{1.71}{\sqrt{105}}$

$6.21 \longleftrightarrow 6.75$

~~$6.48 \pm 1.645 \times \frac{1.71}{\sqrt{105}}$~~

~~$104.7 \longleftrightarrow 105.3$~~

Q2

A B C

The 95% confidence interval for the mean PSQI score for students in India was found to be (6.07, 6.83).

(e) Using this information and the confidence interval calculated in (d), comment on the mean PSQI scores of the UK and India.

(2)

On average, the Uk has a greater proportion of poor sleepers because their mean is higher. The higher the score (when greater than a value) the worse the sleep. However there is a very small confidence interval of 0.54 India isn't in the confidence interval either, meaning India's value is unreliable and possibly inaccurate.

(Total for Question 2 is 18 marks) Total

11 /18

Part (a)

This question asked for an exact test to be used to decide whether this sample provides evidence to support Hamish's suspicion. Candidates should be reminded to refer to earlier parts of the given information to identify Hamish's suspicion. 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 at the beginning of the question tells us that the study's reported findings were that 20% of students slept for less than 5 hours a night (in India).

B1: The hypotheses are correct; the hypotheses have been well explained in words.

M1: Although this question asked for an exact test to be used, the use of a normal approximation in this question was condoned for this series. This mark has been awarded for the correct use of the normal approximation formula $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ this implies the use of the model X~B(40, 0.2).

A0: No attempt to calculate a probability using the binomial distribution.

M1: For a comparison of their test statistic = −0.791 with their critical value = −1.645

A0: This mark depended upon the previous M1A1M1 so could not be awarded.

Part (b)

E1: 'only asked from his university' is a correct comment.

E0: No further relevant comments.

Part (c)

B1: We condone the use of $H_0 : p_I = p_{UK}, H_1 : p_I \neq p_{UK}$ so this mark has been awarded.

M1: Normal approximation method used; this mark awarded for the use of 0.626 in the formula $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

M1: For dividing by their appropriate standard deviation.

A1: 3.685 is correct.

M1: Comparison has been shown leading to reject H_0

E1: Correct conclusion has been stated in context and the previous 4 marks were awarded.

Part (d)

M1: For the use of $\sqrt{105}$

B0: 1.645 is incorrect.

A0: The answer is incorrect.

Part (e)

M0: A comparison of the two confidence intervals either in words or on a number line was needed for this mark.

E0: This mark depends upon a comparison of the two confidence intervals.

Question 2 - Response C

(3)

· Random sample

H0: $p = 0.2$
H1: $p < 0.2$

$p(x \leq 6) = 0.28589137$

T.S = 0.286 (3 sf)
C.V = 0.05

T.S > C.V. Accept H0.
At a 5% sig level there is insufficient evidence to suggest students who sleep less than 5 hours a night is smaller in the UK than India

Question 2 continued

(b) Identify **two** limitations of Hamish's study.

(2)

· Relatively small sample to represent the entire UK

· Only students where used so again not representative of all of the UK.

Hamish also obtains the PSQI scores for a randomly selected sample of 105 UK students.

He finds that 80% of these students can be classified as 'poor sleepers'.

(c) Conduct a test to investigate whether there is evidence for a difference in the proportion of 'poor sleepers' between students in the UK and those in India.

(6)

$105 \times 0.8 = 84$

$H_0: p = 0.626$

$H_1: p \neq 0.626$

$p(X \geq 84) = 0.00009302$

$1.5 = 0.00009302$ (3 s.f.)

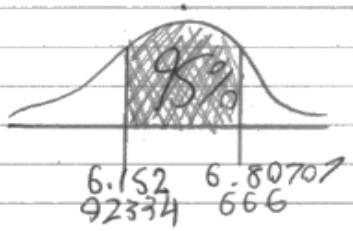
$C.V = 0.025$

$1.5 < C.V$. Reject H_0. Reject H_0. At a 5% significance level, there is sufficient evidence to suggest a difference between the proportion of poor sleepers in the UK and India.

The PSQI scores for these 105 students have mean 6.48 and standard deviation 1.71

(d) Construct a 95% confidence interval for the corresponding population mean PSQI score.

(3)



95% mean [6.15292334, 6.80707666]

The 95% confidence interval for the mean PSQI score for students in India was found to be (6.07, 6.83). (6.15, 6.81)

(e) Using this information and the confidence interval calculated in (d), comment on the mean PSQI scores of the UK and India.

(2)

The mean PSQI scores in India are wider spread than in the UK.

Q2

A B C

15 /18

Part (a)

This question asked for an exact test to be used to decide whether this sample provides evidence to support Hamish's suspicion. Candidates should be reminded to refer to earlier parts of the given information to identify Hamish's suspicion. 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 at the beginning of the question tells us that the study's reported findings were that 20% of students slept for less than 5 hours a night (in India).

B1: Correct hypotheses.

M1: This mark was for the use of $X \sim B(40, 0.2)$. The correct value of 0.2859 implies the use of the correct model.

A1: Correct test statistic.

M1: The test statistic and critical value are stated followed by 'their test statistic > critical value' which is a correct comparison.

A1: The previous M1A1M1 were awarded, and a correct conclusion has been stated in context, in this series we have condoned the absence of the word proportion in the conclusion.

Part (b)

E1: Correct statement about the sample being relatively small.

E0: There was not enough information given in this answer, reference to only using students at Hamish's university for this mark was required rather than just referring to students.

Part (c)

B1: The hypotheses have been stated correctly. The use of p or π in the hypotheses was condoned.

M1: The exact binomial method has been used, an answer of 0.00009302 implies the use of the model $X \sim B(105, 0.626)$

M1: The exact binomial method has been used, an answer of 0.00009302 implies $P(X \geq 84)$ has been found

A1: 0.00009302 is correct.

M1: Comparison of their test statistic with their critical value.

E1: Correct conclusion in context and previous 4 marks awarded.

Part (d)

M1: Correct answer implies use of $\sqrt{105}$

B1: Correct answer implies use of 1.96

A1: The answer is correct.

Part (e)

M0: A comparison of the two confidence intervals either in words or on a number line was needed for this mark.

E0: This mark depended upon a comparison of the two confidence intervals.

Question 3

Question | Mark Scheme | Examiner Comments | Performance | Response A | Response B | Response C

Question 3 - Question

3 A study was conducted into whether social media data could be used to understand tourists' preferences for nature-based experiences.

A large number of photographs taken by tourists at Kruger National Park over a given period, and then posted on Instagram and Flickr, was analysed.

The researchers divided the photographs into categories according to subject matter. They then used two-sample *z*-tests to compare proportions of pictures posted on

Instagram with those posted on Flickr, for each category.
The *p*-values from some of these tests are given in **Figure 1**.

All differences were calculated as (proportion on Instagram) – (proportion on Flickr).

Category	*p*-value
People active	$p < 0.0001$
People posing	$p < 0.01$
Birds	$p < 0.01$
Arthropods (insects, spiders, etc.)	$p < 0.01$
Reptiles	$p < 0.01$

Figure 1
[Data source: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12343]

(a) Interpret, in context, the *p*-value in **Figure 1** for 'People active'.

(2)

The researchers also calculated Cohen's *d* values for the differences in proportions, which are interpreted in the same way as Cohen's *d* values for differences in means.

All differences were calculated as (proportion on Instagram) – (proportion on Flickr).

Note that a **negative** value of Cohen's *d* indicates that the proportion on Instagram is **smaller** than the proportion on Flickr.

These values, for the categories used in **Figure 1**, are summarised in **Figure 2**.

Category	Cohen's *d*
People active	0.82
People posing	1.49
Birds	–1.00
Arthropods (insects, spiders, etc.)	–1.50
Reptiles	–0.44

Figure 2

(b) Interpret, in context, the *d*-values in **Figure 2** for

(i) people posing,

(ii) reptiles.

(3)

(c) Provide an overall summary for a reader with limited statistical knowledge of the main findings from **Figure 1** and **Figure 2**.

(3)

(Total for Question 3 is 8 marks)

Question 3 - Mark Scheme

Qu	Scheme	Marks	AO	Notes
3(a)	$p < 0.0001$ suggests there is (strong) evidence of a **difference** between the of pictures of **people** being active on **Instagram** and on **Flickr**	E1	2.1b	Evidence of difference.
		E1	2.1a	Completely correct in context.
3(b)	$d = 1.49$ suggests a **large effect** between difference in proportion of pictures of people posing on Instagram and that on Flickr (proportion of pictures of people posing is greater on Instagram.)	E1	1.3	Large effect or big difference
	d = −0.44 suggests a **medium** or **small effect** between difference in proportion of pictures of reptiles on Instagram and that on Flickr (proportion of pictures of reptiles on Flickr is greater).	E1	1.3	Medium/small effect or small difference Do not allow for very small
		E1	2.1a	Context correct for at least one.
3(c)	**Possible comments (not exhaustive)**			

Qu	Scheme	Marks	AO	Notes
	Instagram and Flickr have different proportions of pictures posted in these categories.			
	All the differences are 'large', except for in the Reptile category.			Reptiles posted least
	There is a higher proportion of human-based photos on Instagram.			**or** Clear difference between the platforms for animal and human based photographs.
	There is a higher proportion of animal-based photos on Flickr.			
	There is a higher proportion of arthropod photos on Flickr.			
		E1, E1	2.1b, 2.1b	E1 for each sensible comment (max E2)
		E1dep	1.3	Use of non-technical language for non-specialist audience dependent upon a sensible comment.
	Total	**8**		

Question 3 - Examiner Comments

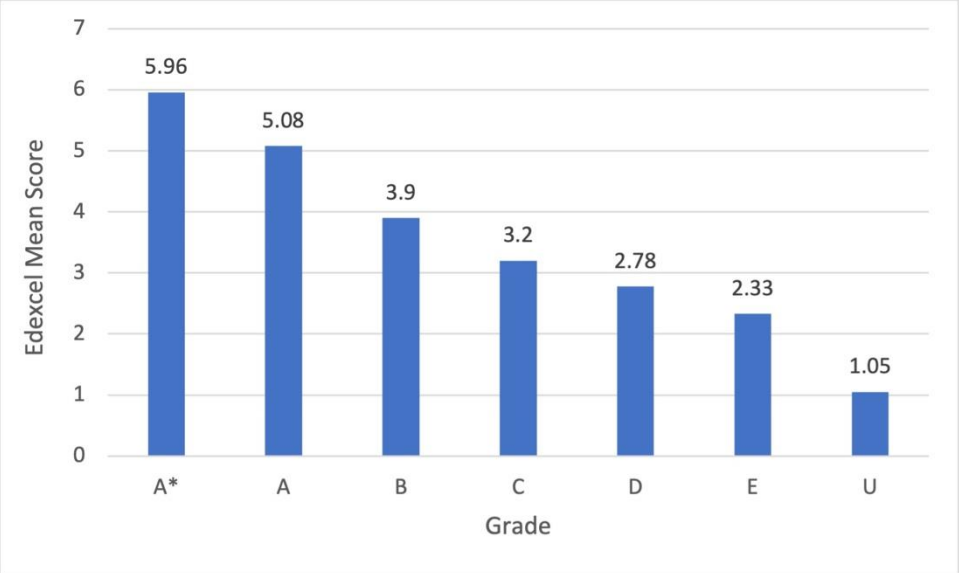
In (a) it was rare to see full marks and many candidates did not seem to understand the relevance of the given *p*-values. Common errors included interpretations about the given *p*-values indicating a very small probability of posting pictures of people being active, rather than identifying that a very 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.

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 3 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
3.48	8	44	3.48	5.96	5.08	3.90	3.20	2.78	2.33	1.05



Question 3 - Response A

Category	p-value
People active	$p < 0.0001$
People posing	$p < 0.01$
Birds	$p < 0.01$
Arthropods (insects, spiders, etc.)	$p < 0.01$
Reptiles	$p < 0.01$

Figure 1

[Data source: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12343]

(a) Interpret, in context, the *p*-value in **Figure 1** for 'People active'

(2)

There's less than a 0.01% chance that people post pictures of them being active on Instagram or Flickr.

These values, for the categories used in **Figure 1**, are summarised in **Figure 2**.

Category	Cohen's *d*
People active	0.82
People posing	1.49
Birds	–1.00
Arthropods (insects, spiders, etc.)	–1.50
Reptiles	–0.44

$\ominus = lg < F$

Figure 2

(b) Interpret, in context, the *d*-values in **Figure 2** for

(i) people posing,

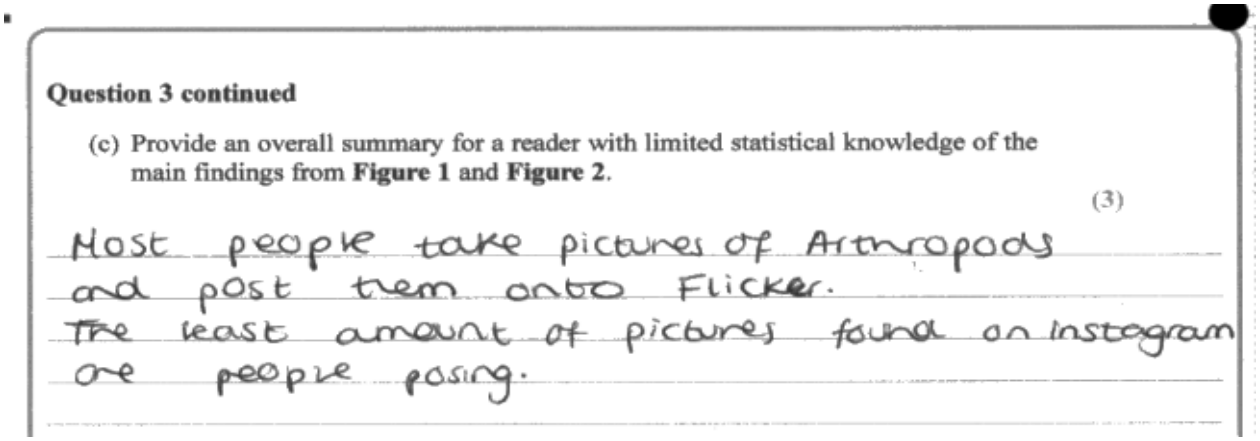
There are more people posing on Instagram than Flickr.

(ii) reptiles.

There are more people, but not many all together, & posting pictures of reptiles on Flickr than Instagram.

Q3

A B C



Question 3 continued

(c) Provide an overall summary for a reader with limited statistical knowledge of the main findings from **Figure 1** and **Figure 2**.

(3)

Most people take pictures of Arthropods and post them onto Flicker. The least amount of pictures found on Instagram are people posing.

1 /8

Part (a)

E0: The comment is incorrect. A small *p*-value suggests the alternative hypothesis will be rejected; we require a comment about there being evidence of a difference for this mark to be awarded.

E0: A completely correct answer in context was required for this mark.

Part (b)

E0: A comment about the effect size for people posing $d = 1.49$ was required for this mark.

E0: A comment about the effect size for reptiles $d = -0.44$ was required for this mark.

E1: This mark was awarded for the correct context, more pictures posted of reptiles on Flickr than on Instagram was enough for this mark.

Part (c)

E0: Most people take pictures of arthropods is not clear enough for this mark, some reference to there being a higher proportion of arthropod photos on Flickr was needed to award a mark.

E0: The second comment is not correct; a comment about there being a higher proportion of human-based photos on Instagram was needed for this mark.

E0: This mark is for the use of non-technical language but is dependent upon at least one sensible comment as stated in the mark scheme.

Question 3 - Response B

Category	*p*-value
People active	$p < 0.0001$
People posing	$p < 0.01$
Birds	$p < 0.01$
Arthropods (insects, spiders, etc.)	$p < 0.01$
Reptiles	$p < 0.01$

Figure 1

[Data source: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12343]

(a) Interpret, in context, the *p*-value in **Figure 1** for 'People active'.

(2)

the proportion of photos taken in kruger National Park by tourists of people being active that were then posted on instagram is very similar to the proportion of these photos of active people that were posted on flickr

Category	Cohen's *d*
People active	0.82
People posing	1.49
Birds	−1.00
Arthropods (insects, spiders, etc.)	−1.50
Reptiles	−0.44

Figure 2

(b) Interpret, in context, the *d*-values in **Figure 2** for

(i) people posing,

when cohens d values for differences are analysed, flickr had a smaller proportion of people posing compared to instagram. And when divided by S it gave a larger value.

Q3

A B C

(ii) reptiles.

when comes d values for differences are analysed, flickr has the higher proportion of reptiles posted compared to instagram.

(c) Provide an overall summary for a reader with limited statistical knowledge of the main findings from **Figure 1** and **Figure 2**.

(3)

figure 1 finds that the amount of tourist photos taken at kruger park that are posted on instagram and flickr are ~~extremely~~ slightly different for photos of people posing, birds, arthropods, and reptiles. And for photos of people active, the amount of photos over that period of time are posted equal amounts on the two sites. // figure 2 finds that for reptiles, Arthopods, and birds instagram has less of those photos compared to flickr, but for people posing and people active, instagram has more.

(Total for Question 3 is 8 marks) Total

4 /8

Q3

A B C

Part (a)

E0: The comment is incorrect. A small *p*-value suggests the alternative hypothesis will be rejected; a comment about there being evidence of a difference was required for this mark to be awarded.

E0: A completely correct answer in context was required for this mark

Part (b)

E0: A comment about the effect size for people posing $d = 1.49$ was required for this mark.

E0: A comment about the effect size for reptiles $d = -0.44$ was required for this mark.

E1: This mark was awarded for the correct context; Flickr had a smaller proportion of people posing compared to Instagram was enough for this mark.

Part (c)

E1: Comments about Figure 1 at the start can be ignored. Figure 2 finds that for reptiles, arthropods and birds Instagram has less of those photos compared to Flickr was enough to award this mark.

E1: The final comment is enough for the second mark. '… for people posing and people active, Instagram has more'.

E1: This mark is for the use of non-technical language throughout part (c) and as there is also at least one sensible comment this mark can be awarded.

Question 3 - Response C

Category	*p*-value
People active	$p < 0.0001$
People posing	$p < 0.01$
Birds	$p < 0.01$
Arthropods (insects, spiders, etc.)	$p < 0.01$
Reptiles	$p < 0.01$

Figure 1

[Data source: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12343]

(a) Interpret, in context, the *p*-value in **Figure 1** for 'People active'.

(2)

a) This means the probability of an active person posting a picture is less than 0.01% of 0.01 of significant this means there is a difference in the proportion posting on instagram and Flickr of people being active.

Category	Cohen's *d*
People active	0.82
People posing	1.49
Birds	–1.00
Arthropods (insects, spiders, etc.)	–1.50
Reptiles	–0.44

Figure 2

(b) Interpret, in context, the *d*-values in **Figure 2** for

(i) people posing,

very large effect size between the proportion on instagram and flickr meaning a lot more people post on instagram than flickr

Q3

A B C

(ii) reptiles.

a medium effect size but in this case flickr has more posts of this than instagram.

(c) Provide an overall summary for a reader with limited statistical knowledge of the main findings from **Figure 1** and **Figure 2**.

(3)

In figure 1 it shows ~~that~~ every category has a significant result ~~~~ meaning of difference between the two however in figure 2 it shows a large effect size for ~~most~~ all categories other than reptiles which is a surprise due to the significant result there should be a large effect size this means further investigation in to reptile pictures.

(Total for Question 3 is 8 marks) Total

Q3

7 /8

Part (a)

E1: The comment is correct there is reference to a difference between Instagram and Flickr.

E1: Completely correct answer in context.

Part (b)

E1: (Very) Large effect size is enough for this mark.

E1: Mention of medium effect is enough for this mark.

E1: This mark was awarded for the correct context.

Part (c)

E1: A difference between the two is equivalent to the first statement in the mark scheme.

E1: '... all categories other than reptiles on Flickr' is equivalent to the second statement in the mark scheme.

E0: This mark is for the use of non-technical language, mentions significant and effect size so this mark cannot be awarded.

Question 4

Question | Mark Scheme | Examiner Comments | Performance | Response A | Response B | Response C

Question 4 - Question

4 A farmer wanted to maximise the spring oat yield, tonnes per hectare, that she could get from a field.

The farmer was able to control two factors

- variety of oat seed planted: Aspen, Canyon, Delfin or Merlin
- concentration of fertiliser used: 10%, 20% or 30%

She wanted to investigate whether either of these factors might have an impact on the yield.

For her experiment, she split a field into four strips and randomly allocated one of the four different types of oat seeds for planting in each strip.

She then subdivided each strip into three smaller plots and randomly allocated one of the three different fertiliser concentrations to each plot during the growing season.

At harvest time, she measured the oat yield for each plot. **Figure 3** gives the plot yields in tonnes per hectare.

	Plot fertiliser concentration			
Oat seed variety	**10%**	**20%**	**30%**	**Total**
Aspen	5.5	5.8	5.7	**17**
Canyon	5.6	5.6	5.5	**16.7**
Delfin	5.5	5.8	5.6	**16.9**
Merlin	5.8	6.1	6.3	**18.2**
Total	**22.4**	**23.3**	**23.1**	**68.8**

Figure 3

Note that $\sum\sum x_{ij}^2 = 395.14$

(a) Making any necessary assumptions, perform a two-factor ANOVA to investigate for a difference between varieties **and** for a difference between fertiliser concentration.

(11)

(b) State **two** assumptions necessary to make the analysis in (*a*) valid.

(2)

(c) With numerical justification, what advice would you give to the farmer to maximise her yield with reference to fertiliser concentration and seed variety.

(2)

(Total for Question 4 is 15 marks)

Question 4 - Mark Scheme

Qu	Scheme	Marks	AO	Notes
4(a)	H_0: No difference in means between varieties H_1: At least two means differ (between varieties) H_0: No difference in means between fertiliser concentration H_1: At least two means differ (between fertiliser concentration)	B1	1.3	oe At least one correct pair Accept equivalent in symbols, provided two sets of hypotheses (or hypotheses combined) e.g. H_0: $\mu_{i,j} = \mu$ for all i,j H_1: $\mu_{i,j} \neq \mu$ for some i,j H_0: $\mu_A = \mu_C = \mu_D = \mu_M$ H_1: At least two means differ H_0: $\mu_{10} = \mu_{20} = \mu_{30}$ H_1: At least two means differ May also be seen combined, e.g. H_0: No difference in means between varieties **and** no difference in means between fertiliser concentration H_1: At least two means differ between varieties **or** at least two means differ between fertiliser concentration
	Total SS $= 395.14 - \frac{68.8^2}{12}$ $= 395.14 - 394.45$ $= 0.6867$	M1	1.3	PI by correct table SS Total awrt 0.687
	Variety SS $= \frac{17^2+16.7^2+16.9^2+18.2^2}{3} - 394.45$	M1	1.3	PI (At least) one of these seen

Qu	Scheme	Marks	AO	Notes
	= 0.46			**Condone** small slip
	Fertiliser SS = $\frac{22.4^2+23.3^2+23.1^2}{4} - 394.45$ = 0.1117			awrt 0.11
	Error SS= $0.6867 - 0.46 - 0.1117$ = 0.115	M1ft	1.3	PI ft their SS values not negative
	<table><tr><th></th><th>ss</th><th>df</th><th>ms</th></tr><tr><td>**Variety**</td><td>0.460</td><td>3</td><td>0.153</td></tr><tr><td>**Fertiliser**</td><td>0.112</td><td>2</td><td>0.056</td></tr><tr><td>**Error**</td><td>0.115</td><td>6</td><td>0.019</td></tr><tr><td>**Total**</td><td>0.687</td><td>11</td><td></td></tr></table>			Correct table awrt (ignoring total) Scores M1M1M1B1M1
		B1	1.3	PI df correct (3,2,6)
		M1ft	1.3	PI MS=SS/df for variety or fertiliser
	To compare variety, $F = \frac{their\ MSv}{their\ MSe} = 7.984$ To compare fertiliser concentration, $F = \frac{their\ MSf}{their\ MSe} = 2.913$	M1	1.3	PI At least one of (variety or fertiliser) their MS/Error MS not if negative
		A1	1.3	awfw F=7.8~8.6 (or p = 0.0161) **or** awfw F=2.8~3.2 (or p = 0.1306)
	Critical value $F_6^3(0.05) = 4.757$ Critical value $F_6^2(0.05) = 5.143$	B1	1.3	For either awrt 4.76 **or** awrt 5.14
	"8"> "4.757" so reject H_0 for varieties. "2.913" < "5.143" so do not reject H_0 for fertiliser concentration.	M1	2.1b	PI Correct comparisons 'their ts' with either correct cv Either needed. **or** comparing p-values with 0.05
	Thus there is significant evidence of a **difference** between mean oat yields for the **varieties** of oat seed … …but there is **no** significant evidence of a **difference** between the mean oat yields for the different **fertiliser** concentrations.	E1dep	2.1a	For **both** conclusions in context, not too definite. dep previous 3 marks

Qu	Scheme	Marks	AO	Notes
4(b)	(Oat yield is approximately) normally distributed (for each variety/fertiliser concentration combination).			
	(Oat yield has approximately) equal variances (for each variety/fertiliser concentration combination).			
	No interaction between oat variety and concentration of fertiliser.			
		E1, E1	3.1a, 3.1a	E1 for each assumption Max E1 if no context
4(c)	<table><tr><th></th><th>Total</th><th>Mean</th></tr><tr><td>Aspen</td><td>17</td><td>5.67</td></tr><tr><td>Canyon</td><td>16.7</td><td>5.57</td></tr><tr><td>Delfin</td><td>16.9</td><td>5.63</td></tr><tr><td>Merlin</td><td>18.2</td><td>6.07</td></tr></table>			
	It appears Merlin would be the best variety to maximise yield because it has the greatest mean yield, 6.07 (or total 18.2)	E1	2.1b	Merlin suggested
	There is no significant difference in concentration of fertiliser so no advice on which to use.	E1	2.1b	Correct fertiliser comment
	Total	**15**		

Question 4 - Examiner Comments

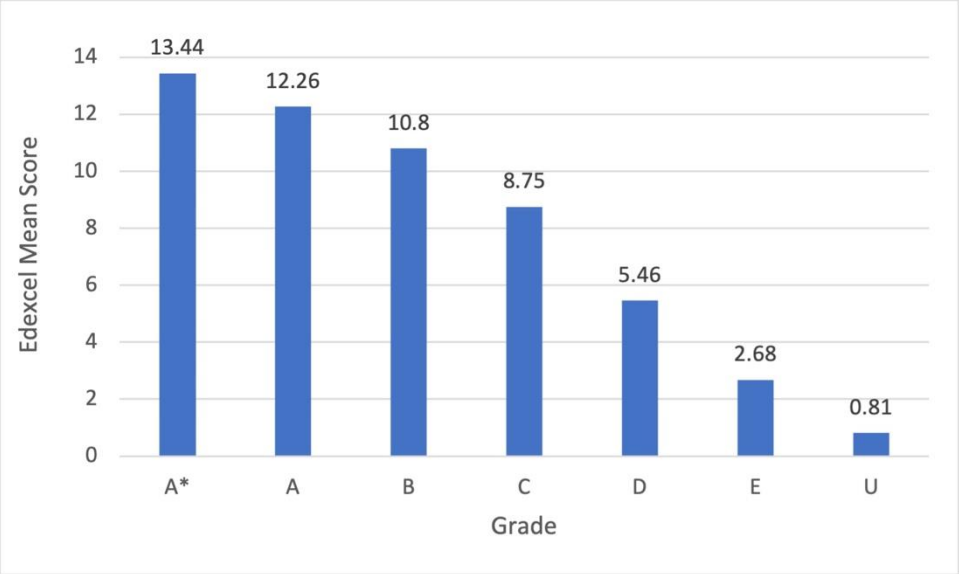
Part (a) of this question was well answered in general with the majority of students earning at least 6 of the available 11 marks. Candidates had clearly been well prepared for this style of question and were able to carry out a two-factor ANOVA test. Those students who made use of the statistical functions 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 4 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
8.32	15	55	8.32	13.44	12.26	10.80	8.75	5.46	2.68	0.81



Question 4 - Response A

Question 4 continued

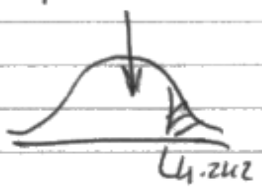
(a) Making any necessary assumptions, perform a two-factor ANOVA to investigate for a difference between varieties **and** for a difference between fertiliser concentration. (11)

$H_0: \mu_A = \mu_C = \mu_O = \mu_M$ $H_1: \mu_A \neq \mu_C \neq \mu_O \neq \mu_M$ (2T, 5%)

$H_0: \mu_1 = \mu_2 = \mu_3$ $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

	SS	V	MS	F
Seed	$\frac{17^2}{3^2} + \frac{16.2^2}{3^2} + \frac{16.4^2}{3^2} + \frac{18.2^2}{3^2} - \frac{68.8^2}{12^2} = 0.460$	$4-1 = 3$	$\frac{0.46}{3} = 0.153$	$\frac{0.153}{0.128} = 1.195$
fertiliser	$\frac{22.4^2}{4^2} + \frac{23.3^2}{4^2} + \frac{23.1^2}{4^2} - \frac{68.8^2}{12^2} = 0.112$	$3-1 = 2$	$\frac{0.112}{2} = 0.056$	$\frac{0.056}{0.128} = 0.4375$
error	$0.687 - 0.112 - 0.460 = 0.115$	$12-3 = 9$	$\frac{0.115}{9} = 0.128$	
total	$395.14 - \frac{68.8^2}{12^2} = 0.687$	14		

$F_{seed} = 1.195$ $k = 4.242$



$F_{fertiliser} = 0.4375$ $k = 4.857$

Q4

A B C

Question 4 continued

Accept Ho for both

Evidence suggests that there is no diggerence for both fertiliser concentration and type of seed, for how much a farmer can yeild.

(b) State **two** assumptions necessary to make the analysis in (a) valid.

(2)

-The variences had to be (very) similar for both seeds and fertiliser.

- Had to be approximately normally distributed

(c) With numerical justification, what advice would you give to the farmer to maximise her yield with reference to fertiliser concentration and seed variety.

(2)

To maximise her yeild she should manage the type of seed she uses because the F_{seed} is greater than $F_{fertiliser}$

1.145 > 0.4375. This shows that the seed is more ~~right~~ influential on the yeild.

Q4

A
B
C

7 /15

Part (a)

B0: Incorrect hypotheses we need to see at least two means differ for H_1

M1: SS_T is correct, note this mark could also have been implied by the correct values in the table.

M1: SS_R (and SS_C) are correct, note this mark could also have been implied by the correct values in the table.

M1: SS_E correct, note this mark could have been implied by the correct values in the table.

B0: Incorrect degrees of freedom.

M1: Values incorrect and no method shown to enable follow through marks to be awarded.
Note if a method for calculating the mark scheme values had been shown this would have enabled follow through marks to have been awarded here even if the two mark scheme values had been incorrect.

M1: Although the *F* values (test statistic) are incorrect, this mark can still be awarded as the method has been shown using their incorrect MS_E value.

A0: Neither *F* value are within the required range.

B0: Neither critical value is correct.

M0: A comparison of their test statistic with a correct critical value was needed. Neither critical value is correct so this mark cannot be awarded.

E0: The previous three marks have not been awarded so this mark cannot be gained.

Part (b)

E1: Normally distributed is one correct assumption.

E1: Seeds and fertiliser mentioned, this is enough to enable the mark for context to be awarded.

Part (c)

E0: Merlin not suggested.

E0: As there was no significant difference in concentration of fertiliser in part (a) there is no advice on which fertiliser to use.

Question 4 - Response B

Question 4 continued

(a) Making any necessary assumptions, perform a two-factor ANOVA to investigate for a difference between varieties **and** for a difference between fertiliser concentration.

(11)

	SS	DoF	MSS	F ratio
Between rows	0.46	4-1 = 3	0.1533	0.1533 ÷ 0.01922 = 7.976
Between columns	0.1117	3-1 = 2	0.0559	0.0559 ÷ 0.01922 = 2.908
Total	0.687			
Error		6	0.01922	

$SS_R = \frac{17^2}{3} + \frac{16.7^2}{3} + \frac{16.9^2}{3} + \frac{18.2^2}{3} - \frac{68.8^3}{12}$ 0.46

$\Sigma x^2 = 395.14$

$SS_C = \frac{22.4^2}{4} + \frac{23.3^2}{4} + \frac{23.1^2}{4} - \frac{68.8^2}{12} = 0.1117$

$SS_T = 395.14 - \frac{68.8^2}{12} =$ 0.687

$SS_E =$ 0.687 − 0.1117 − 0.46 = 0.1153

$MSS_R = 0.46 \div 3 = 0.1533$

$MSS_C = 0.1117 \div 2 = 0.0559$

$MSS_E = 0.1153 \div 6 = 0.01922$

Q4

A B C

(b) State **two** assumptions necessary to make the analysis in (a) valid.

(2)

Assuming the samples have ~~are equal~~ a common variance and the factors do not effect each other

(c) With numerical justification, what advice would you give to the farmer to maximise her yield with reference to fertiliser concentration and seed variety.

(2)

Advise the 20%. fertiliser concentration because ~~overall~~ it gave the better results in every type of seed. The Merlin seed variety was most successful with a total of 18·2

9 /15

Part (a)

B0: No hypotheses have been stated.

M1: SS_T correct, note this mark can be implied by a correct table.

M1: SS_R (or SS_C) correct, note this mark can be implied by a correct table.

M1: SS_E correct, note this mark can be implied by a correct table.

B1: Correct degrees of freedom.

M1: Values correct in the table.

M1: Test statistic values, *F*-ratio, correct; note this mark can be awarded for just one of the test statistic values correct.

A1: Test statistic = 7.976 is within the required range, test statistic = 2.908 is also within the required range.

B0: No critical values have been stated.

M0: There is no comparison shown.

E0: No conclusions stated.

Part (b)

E1: Common variance is enough to suggest the variances are equal.

E0: No context has been given for either assumption so a maximum of E1 can be awarded.

Part (c)

E1: Merlin suggested.

E0: As there was no significant difference in concentration of fertiliser in part (a) there is no advice on which fertiliser to use.

Question 4 - Response C

Question 4 continued

(a) Making any necessary assumptions, perform a two-factor ANOVA to investigate for a difference between varieties **and** for a difference between fertiliser concentration.

(11)

assumptions: the population distributions are ~~dont~~ Normal, the variances are ~~the same~~ equal/similar, the samples ~~are~~ are independant and random

$H_0: M_A = M_C = M_O = M_M$

H_1: atleast two of the population means diff. for the oat seed varieties

	SS	dof	MS
B_O	0.46	3	0.153
B_F	0.11167	2	0.055835
E	0.11533	6	0.0192
T	0.68667	11	

$\Sigma x_i = 68.8$

$\Sigma x_i^2 = 395.14$

$n = 12$

$ts = \frac{0.153}{0.0192} = 7.97$

$F^{3\rightarrow}_{6\downarrow} = 4.757$

$SS_T = 395.14 - \frac{68.8^2}{12} = 0.686667$

$SS_{B_{OF}} = \frac{22.4^2 + 23.3^2 + 23.1^2}{4} - \frac{68.8^2}{12} = 0.11167$

$SS_{B_O} = \frac{17^2 + 16.7^2 + 16.9^2 + 18.2^2}{3} - \frac{68.8^2}{12} = 0.46$

$SS_E = 0.687 - 0.11167 - 0.46 = 0.11533$

$7.97 > 4.757$ ∴ reject H_0

there is significant evidence to suggest that atleast two of the population means diff for the oat types

$H_0: M_{1O} = M_{2O} = M_{3O}$

H_1: atleast two of the population means diff for the fertilizer concentration

Q4

A B C

$F^{2,6}_{0.05} = 5.143$

$\frac{0.055835}{0.0192} = 2.89$

$2.89 < 5.143$

$\therefore$ accept H_0

there is insufficient evidence to suggest atleast two of the population means diff for the concentration of fertiliser.

(b) State **two** assumptions necessary to make the analysis in (a) valid.

(2)

- the population distributions must have equal variances
- the population distributions must be normally distributed.

(c) With numerical justification, what advice would you give to the farmer to maximise her yield with reference to fertiliser concentration and seed variety.

(2)

Merlin as the ~~expected~~ total yield is the highest (18.2) and 20% because the total yield is the highest (23.3 tonnes per Hectare)

13 /15

Q4

Part (a)

B1: Correct hypotheses; this mark can be given for hypotheses written in words or symbols.

M1: SS_T correct, note this mark can be implied by a correct table.

M1: SS_R (or SS_C) correct, note this mark can be implied by a correct table.

M1: SS_E correct, note this mark can be implied by a correct table.

B1: Correct degrees of freedom.

M1: Values correct in the table.

M1: Test statistic values correct, note this mark can be awarded for just one of the test statistic values correct.

A1: Test statistic = 7.97 is within the required range, test statistic = 2.89 is also within the required range.

B1: This mark can be awarded for either critical value = 4.757 or critical value = 5.143

M1: Test statistic = 7.97 compared with 4.757 which is a correct critical value. This mark can be awarded for either of this comparison or for 2.89 compared with 5.143

E1: Both conclusions are correctly stated in context and the previous three marks were awarded.

Part (b)

E1: Either equal variances or normally distributed would gain this mark.

E0: No context has been given for either assumption so a maximum of E1 can be awarded.

Part (c)

E1: Merlin suggested.

E0: As there was no significant difference in concentration of fertiliser in part (a) there is no advice on which fertiliser to use.

Question 5

Question | Mark Scheme | Examiner Comments | Performance | Response A | Response B | Response C

Question 5 - Question

5 Based on an earlier study, Sinead decided to investigate the extent to which people are influenced by other people's laughter.

Participants were asked to rate the 'funniness' of a set of 10 jokes.

The 10 jokes were videoed being told by a professional comedian. Random samples of participants were then shown the video and asked to rate the funniness of each joke on a scale of 1 (unfunny) to 7 (extremely funny).

A first sample of participants watched the video with no laughter track added.

A second sample watched it with a backing track of fake machine-generated laughter.

A third sample watched it with a backing track of real laughter generated by an audience.

The funniness ratings for each joke were then averaged for each of the three samples. The mean ratings obtained are given in **Figure 4**.

		No laughter	Fake laughter	Real laughter
Joke	1	2.3	1.9	1.0
	2	3.2	2.9	4.2
	3	2.1	4.2	3.8
	4	3.4	1.9	3.8
	5	2.5	3.5	3.8
	6	2.2	3.0	3.7
	7	1.3	1.9	2.5
	8	1.3	2.4	2.9
	9	1.4	3.5	4.3
	10	2.5	4.0	5.0

Figure 4

Sinead's belief is that jokes are perceived to be more funny if accompanied by the sound of laughter.

The test values calculated using the data in **Figure 4** are

$t = 1.93$ for 'Fake laughter' compared to 'No laughter' and

$t = 3.51$ for 'Real laughter' compared to 'No laughter'

(a) State the hypotheses being tested and decide whether these data support Sinead's belief.

(4)

(b) Conduct a *t*-test using data from **Figure 4** to investigate whether there is evidence of any difference between 'Fake laughter' and 'Real laughter' in terms of the perceived funniness of jokes.

(8)

(Total for Question 5 is 12 marks)

Question 5 - Mark Scheme

Qu	Scheme	Marks	AO	Notes
5(a)	N = No laughter F = Fake laughter R = Real laughter			
	H_0: $\mu_F - \mu_N = 0$ H_1: $\mu_F - \mu_N > 0$	B1	1.3	oe At least one pair correct
	H_0: $\mu_R - \mu_N = 0$ H_1: $\mu_R - \mu_N > 0$			H_0:$\mu_d = 0$ H_1: $\mu_d > 0$
	[5% one-tailed with $\nu = 9$] cv = 1.83(3)	B1	1.3	cao **or** p-values: awrt 0.043 **and** awrt 0.0033
	$1.93 > 1.83$ or $3.51 > 1.83$ so reject both H_0	M1	2.1b	PI Correct comparison of at least one test value with their critical t-value. Signs consistent. **or** Correct comparison of at least one p-value with 0.05
	The data does support Sinead's belief.	E1dep	2.1a	oe referring to funniness of jokes with laughter. Not too strong and in context.

Qu	Scheme	Marks	AO	Notes
				dep B1M1
5(b)	H_0: $\mu_R - \mu_F = 0$ H_1: $\mu_R - \mu_F \neq 0$	B1	1.3	oe **Accept** μ_d
	Differences: -0.9, 1.3, -0.4, 1.9, 0.3, 0.7, 0.6, 0.5, 0.8, 1.0	B1	1.2	PI Attempt at differences **or** negative
	mean $= (-)0.58$ $s = 0.798(3)$	B1	1.2	PI Both correct **Condone** $s = 0.757(4)$
	ts: $t = \frac{(-)0.58}{\frac{0.798}{\sqrt{10}}}$	M1ft	1.3	PI Numerator may be –ve **or** $\frac{\bar{d}}{\frac{0.798}{\sqrt{10}}}$
	$= 2.297$	A1	1.3	awrt $(\pm)2.30$ **or** $\frac{\bar{d}}{\frac{0.798}{\sqrt{10}}} = 2.262$
	[two-tailed, $\nu = 9$]			
	cv $= 2.262$	B1	1.3	**or** cv $= -2.26$ **or** p-value $= 0.0236$ **or** p-value $= 0.0472$ **or** $\bar{d} = \pm 0.571$
	"2.297" > "2.262" so reject H_0	M1	2.1b	PI Comparison their ts with the correct cv. **or** $0.0236 < 0.025$ **or** $0.0472 < 0.05$ **or** $0.58 > 0.571$

Q5

Qu	Scheme	Marks	AO	Notes
	There is significant evidence of a **difference** between '**Fake** laughter' and '**Real** laughter' in terms of the perceived funniness of jokes. (It appears jokes are thought to be funnier with 'Real laughter' than with 'Fake laughter').	E1dep	2.1a	Correct conclusion in context. Dep previous 4 marks.
SC: Use of two independent samples B1B0B0M1A0B1M1A0 Max4/8 M1 ts=awrt(-)1.3 B1 cv=$\pm$2.101 M1 comparing their 1.3 with 2.10				
	Total	**12**		

Question 5 - Examiner Comments

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.

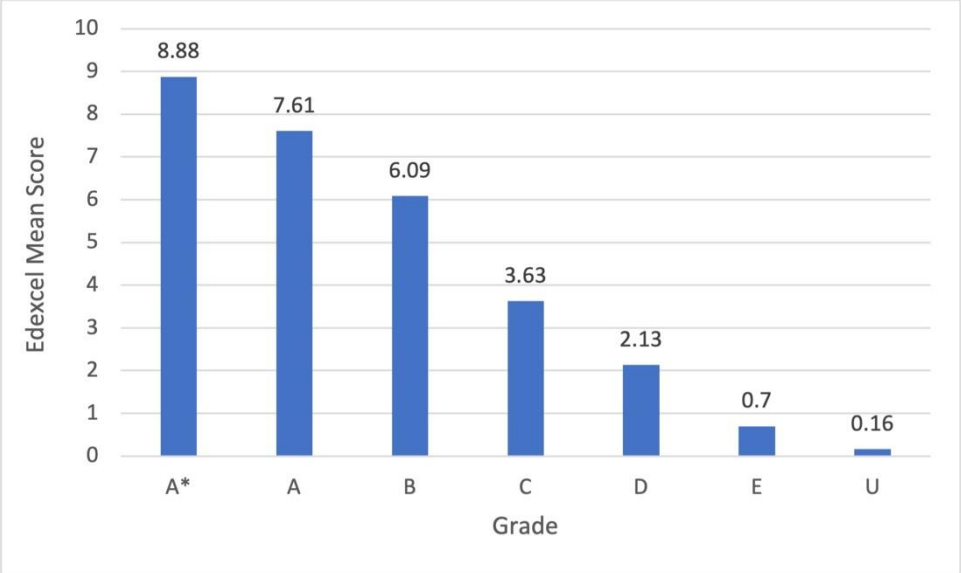
In part (a) the jokes were all numbered 1 to 10, clearly indicating that the results for 'no laughter' and 'fake laughter', and for' no laughter' and 'real laughter' were paired. A common error was to compare the two given test values (t = 1.93 and t = 3.51) together rather than finding the critical 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 5 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
4.22	12	35	4.22	8.88	7.61	6.09	3.63	2.13	0.70	0.16



Q5
A
B
C

Question 5 - Response A

Question 5 continued

(a) State the hypotheses being tested and decide whether these data support Sinead's belief.

(4)

H_0: No laughter ≠

$H_0: \mu_1 = \mu_2 \neq \mu_3$

$H_1: \mu_1 < \mu_2 \neq \mu_3$

The test stats imply a difference.

(b) Conduct a *t*-test using data from **Figure 4** to investigate whether there is evidence of any difference between 'Fake laughter' and 'Real laughter' in terms of the perceived funniness of jokes.

A t-test at the 5% using $\sqrt{=9}$

(8)

$H_0: d = 0$

$H_1: d \neq 0$

Let x be fake laughter funniness and y be that of real laughter

$\sqrt{=9}$

differences: 0.9, −1.3, 0.4, −1.9, −0.3, −0.7, −0.6

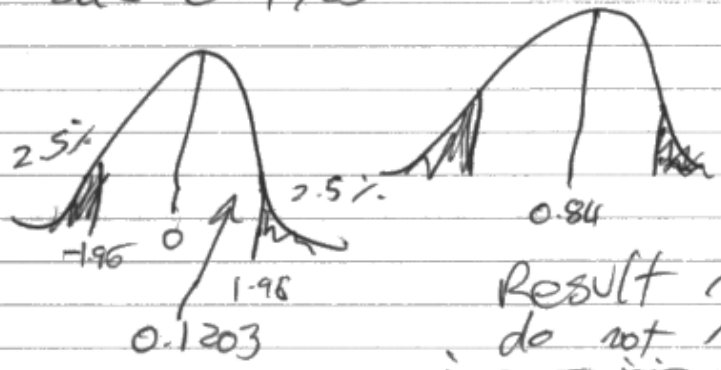
x − y −0.5, −0.8, −1

$\bar{d} = 0.84$

$Sd = 0.4765$

Test stat = 2.228

Crit = 0.1203



Result is significant

do not reject H_0

insufficient evidence of a difference between funniness of fake real laughter.

Q5

A B C

1 /12

Part (a)

B0: Incorrect hypotheses, due to the "^μ_3". A '1' and '2' should be clearly defined.

B0: No attempt to find the required critical value using the *t*-distribution.

M0: No comparison made.

E0: This mark depended upon B1 M1 above (finding the correct critical value).

Part (b)

B0: Incorrect hypotheses using d rather than μ_d

B1: Attempt to find differences; these are all correct.

B0: Incorrect values for $\bar{d}$ and s_d

M0: There is no method shown for their test statistic. As 2.228 is not correct the method cannot be implied from this answer.

A0: The test statistic is not an answer which rounds to ±2.30

B0: The critical value is not correct.

M0: This mark is for a comparison of their test statistic with the correct critical value for their test. Their critical value is not correct so the mark could not be awarded.

E0: There is a conclusion in context, but this mark can only be awarded if the previous 4 marks had also been awarded.

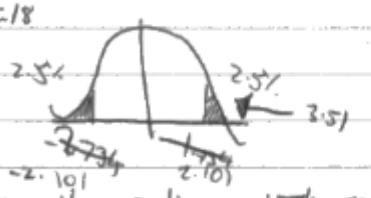
Question 5 - Response B

(a) State the hypotheses being tested and decide whether these data support Sinead's belief.

(4)

H_0: all of the population distributions are identical.
H_1: At least two or more means are different from eachother

I believe these data do support Sinead's belief as there was a t-value of 3.51 for 'real laughter' compared to joke laughter, which would be significant evidence as: $\nu = 18$



Meaning the result is significant.
Reject H_0.
There is significant evidence to suggest that the audience ~~feels~~ perceives jokes as funnier with a background of real laughter compared to no laughter.

(b) Conduct a *t*-test using data from **Figure 4** to investigate whether there is evidence of any difference between 'Fake laughter' and 'Real laughter' in terms of the perceived funniness of jokes.

(8)

Let X be the sunniness rating of each joke when participants are accompanied by joke laughter and Y be that for real laughter.

H_0: $\mu_x - \mu_y = 0$ Use a two-tailed paired t-test at the 5% level using $\nu = 18$

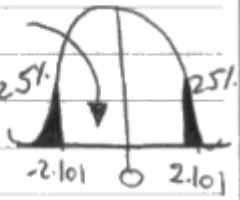
H_1: $\mu_x - \mu_y \neq 0$

$\bar{x} = 2.92$ $\bar{y} = 3.5$

$S_x^2 = 0.764$ $S_y^2 = 1.256$

$$s_p^2 = \frac{(10-1)\,0.764 + (10-1)\,1.256}{10+10-2} = 1.01$$

$$\frac{(2.92-3.5)-(0)}{\sqrt{1.01\left(\frac{1}{10}+\frac{1}{10}\right)}} = -1.29$$



Critical Region:

The result is not significant.
Do not reject H_0.
There is insufficient evidence to suggest a difference, on average, between the ~~funniness~~ sunniness rating of jokes when accompanied by 'real laughter' and 'fake laughter'.

5 /12

Part (a)

B0: Incorrect hypotheses.

B0: Incorrect critical value using the t-distribution with degrees of freedom (df) = 18 rather than df = 9

M1: Comparison shown comparing 3.51 with their critical value.

E0: This mark depended upon B1 M1 above (finding the correct critical value).

Part (b)

This is an example of how marks have been awarded using the special case (SC) for carrying out an incorrect method. This candidate has carried out a t-test for two independent samples with unknown population variance rather than a paired *t*-test.

B1: Correct hypotheses using $\mathrm{H}_1 : \mu_X = \mu_Y$

B0: No attempt to find differences.

B0: No attempt to calculate $\bar{d}$ and s_d

M1: This is not the paired *t*-test but the special case (SC) is applied. The test statistic of −1.29 is correct for their method.

A0: The test statistic is not an answer which rounds to ±2.30

B1: As we are applying the special case the critical value of ±2.101 is correct for their method.

M1: This mark is for a comparison of their test statistic with the correct critical value for their test. Comparison of -1.29 with -2.101 gains this mark.

E0: This mark is not awarded as the previous 4 marks need to have been awarded.

Question 5 - Response C

		No laughter	Fake laughter	Real laughter	
Joke	1	2.3	1.9	1.0	0.9
	2	3.2	2.9	4.2	-1.3
	3	2.1	4.2	3.8	0.4
	4	3.4	1.9	3.8	-1.9
	5	2.5	3.5	3.8	-0.3
	6	2.2	3.0	3.7	-0.7
	7	1.3	1.9	2.5	-0.6
	8	1.3	2.4	2.9	-0.5
	9	1.4	3.5	4.3	-0.8
	10	2.5	4.0	5.0	-1

Figure 4

Sinead's belief is that jokes are perceived to be more funny if accompanied by the sound of laughter.

The test values calculated using the data in **Figure 4** are

$t = 1.93$ for 'Fake laughter' compared to 'No laughter' and

$t = 3.51$ for 'Real laughter' compared to 'No laughter'

Question 5 continued

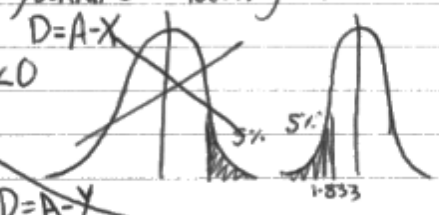
(a) State the hypotheses being tested and decide whether these data support Sinead's belief.

(4)

Let X be funniness rating with fake laughter
Let Y be funniness rating with real laughter
Let A be funniness rating with no laughter.

H_0: $\mu_D = 0$ D=A-X
H_1: $\mu_D < 0$
and
H_0: $\mu_D = 0$
H_1: $\mu_D < 0$ D=A-Y

$\nu = 9$



1.93>1.833 Result not significant- does not support her belief.

3.51>1.833 Result not significant- does not support her belief.

(b) Conduct a *t*-test using data from **Figure 4** to investigate whether there is evidence of any difference between 'Fake laughter' and 'Real laughter' in terms of the perceived funniness of jokes.

(8)

~~H0: μD=0~~

H0: μD=0 D=X-Y

H1: μD≠0

x̄ = 1.92 sx² = 0.764 d̄ = −0.58

ȳ = 3.5 sy² = 1.256 s = 0.7983

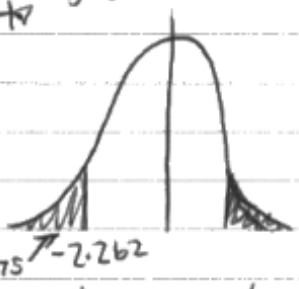
n = 10

2 tailed t test, 5% significance level, ν=9

~~sp² = (10−1)×1.256 + (10−1)×0.764 = 1.01~~

~~pooled~~

~~(3.5−1.92)−0~~ $\frac{-0.58-0}{\frac{0.7983}{\sqrt{10}}} = -2.2975$



Result significant, reject H0 - sufficient evidence to suggest a difference in the mean funniness rating between fake and real laughter

11 /12

Part (a)

B1: Correct hypotheses. Note the clearly labelled letters X, Y and A. This is an example of an equivalent pair of hypotheses since D=A-Y their $H_1 : \mu_D < 0$ is correct.

B1: 1.833 is the correct critical value.

M1: Clear comparison of 1.93 and 3.51 from the question with the critical value of 1.833

E0: Incorrect conclusion following their comparison, their result is actually significant.

Part (b)

B1: Correct hypotheses using μ_D

B1: Differences calculated next to Figure 4, these are equivalent to the differences given in the mark scheme, these differences will lead to a negative test statistic.

B1: Both the mean and *s* values are correct.

M1: Test statistic calculated correctly.

A1: Test statistic = −2.2975 is a correct form of the test statistic.

B1: Critical value = −2.262 is correct for their method.

M1: Comparison of their test statistic with the correct tail of the critical value.

E1: Correct conclusion in context**,** this mark can be awarded as the previous 4 marks had been awarded.

Question 6

Question | Mark Scheme | Examiner Comments | Performance | Response A | Response B | Response C

Question 6 - Question

6 A large gardening company conducted a survey on customer satisfaction by e-mailing a random sample of 400 of its customers, of whom 62 responded.

The company wanted to see if there was a relationship between whether it was the customer's first experience and the likelihood that the customer would use the company again.

The results of this survey are summarised in **Figure 5**.

		First experience		
		Yes	**No**	**Total**
Likelihood of using company again	**Definitely yes**	9	12	**21**
	Probably yes	18	2	**20**
	No	17	4	**21**
	Total	**44**	**18**	**62**

Figure 5

(a) Investigate whether there is evidence of an association between whether it was the customer's first experience and the likelihood that the customer would use the company again.

(8)

(b) Describe, with numerical justification, the nature of any association identified in your conclusion to the test in (a).

(2)

(c) Give **two** possible sources of bias in this investigation.

(2)

(Total for Question 6 is 12 marks)

Question 6 - Mark Scheme

<table>
<tr><td rowspan="8">6(a)</td><td>H_0: no association (between whether it was their first experience and the likelihood they would use the company again)
H_1: an association (between whether it was their first experience and the likelihood they would use the company again)</td><td>B1</td><td>1.3</td><td>oe
Both correct</td></tr>
<tr><td colspan="4">Expected frequencies
<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="2">First experience</th><th></th></tr>
<tr><th>Yes</th><th>No</th><th>Totals</th></tr>
<tr><td rowspan="4">Use again likelihood</td><td>Definitely yes</td><td>14.9</td><td>6.1</td><td>21</td></tr>
<tr><td>Probably yes</td><td>14.2</td><td>5.8</td><td>20</td></tr>
<tr><td>No</td><td>14.9</td><td>6.1</td><td>21</td></tr>
<tr><td>Totals</td><td>44</td><td>18</td><td>62</td></tr>
</table></td></tr>
<tr><td></td><td>M1</td><td>1.3</td><td>PI
At least one expected value correct to 1 dp</td></tr>
<tr><td>Contribution to χ^2 :
<table>
<tr><td>2.34</td><td>5.72 (or 5.71)</td></tr>
<tr><td>1.02</td><td>2.50 (or 2.49)</td></tr>
<tr><td>0.30</td><td>0.72</td></tr>
</table></td><td>M1ft</td><td>1.3</td><td>PI
Attempt at $\frac{(O-E)^2}{E}$
(at least one correct contribution seen to 1dp)</td></tr>
<tr><td>ts: $\chi^2 = \frac{(9-14.90)^2}{14.90} + \cdots + \frac{(4-6.10)^2}{6.10}$</td><td>M1ft</td><td>1.3</td><td>Intention to sum
PI</td></tr>
<tr><td>ts = 12.6</td><td>A1</td><td>1.3</td><td>awfw 12.4~12.8</td></tr>
<tr><td>[χ_2^2 at 5% level]
cv = 5.99</td><td>B1</td><td>1.3</td><td>or p-value
awrt 0.0018 ~ 0.0019
Condone 7.38</td></tr>
<tr><td>“12.6” > 5.99
so reject H_0</td><td>M1</td><td>2.1b</td><td>PI
compare their ts with 5.99
or 0.00185 < 0.05</td></tr>
<tr><td>There is significant evidence of an association between whether it was their first experience and whether they would use the company again.</td><td>E1dep</td><td>2.1a</td><td>In context
dep ts and cv correct</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</table>

6(b)	The joint category "definitely use the company again" and "not first experience" make the biggest contribution to χ^2 at 5.72 (or 5.71)	E1	2.1a	Condone mention of other contributions, but must identify joint category 5.72 **or** quote numerical justification 12 and 6.1
	This suggests that, far more customers than expected, who have used the company before, would definitely use them again.	E1	2.1a	oe Full explanation in context Comment to include contextualised reference to the direction of the difference between the obs and exp frequencies
6(c)	**Possible sources of bias** **(Not exhaustive)**			
	Only 62 out of 400 (ie only 15.5%) responded to the email. This very large non-response rate could introduce bias.			Comment on low response rate.
	Customers who chose not to give their email address are excluded.			
	Not everyone checks their email.			Comment on use of email. Survey only done by email.
	The categories used for 'Use Again'? are likely to introduce bias as there are two 'positive' responses but only one 'negative' response available.			Comment on two positive choices.
	The customers that responded to the email may be the customers with the strongest opinions.			Idea of customers self-selecting condone volunteer sample
		E1, E1	3.1a, 3.1a	
				E1 for each sensible comment (Max E2)
	Total	**12**		

Q6

A
B
C

Question 6 - Examiner Comments

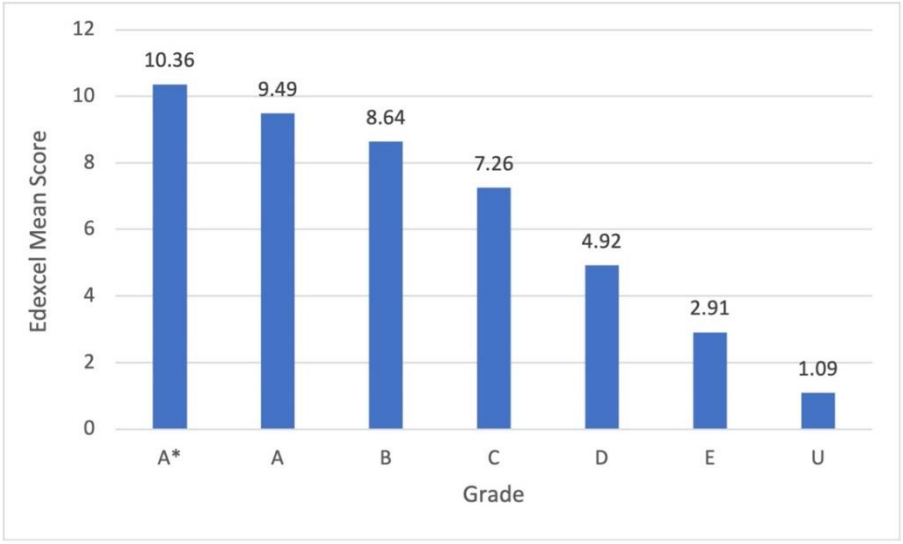
This question was a standard contingency table question. Over 50% of candidates were able to score the full marks available in part (a) demonstrating that they were able to recognise and analyse a contingency table with ease. Many candidates appeared to be familiar with the use of the statistical functions 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.

Question 6 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
6.89	12	57	6.89	10.36	9.49	8.64	7.26	4.92	2.91	1.09



Question 6 - Response A

		First experience		
		Yes	No	Total
Likelihood of using company again	Definitely yes	9 ✓	12	21
	Probably yes	18	2	20
	No	17	4	21
	Total	44	18	62

Figure 5

(a) Investigate whether there is evidence of an association between whether it was the customer's first experience and the likelihood that the customer would use the company again.

(8)

	obs	E	
First (Def yes)	9	14.903	2.338
First (Prob)	18	14.194	1.021
F (NO)	17	14.903	
N (DY)	12	6.097	
N (PY)	2	5.806	
N (N)	4	6.097	

(c) Give **two** possible sources of bias in this investigation.

(2)

* Sent to their Customers,
* Only 62 responded, So that population may not relate to represent the whole population

3 /12

Part (a)
B0: No hypotheses stated.
M1: At least one expected frequency correct.
M1: At least one correct contribution shown.
M0: No attempt to sum the contributions.
A0: Test statistic not calculated.
B0: Critical value not stated.
M0: No comparison shown.
E0: No conclusion drawn.

Part (b)
E0: Not attempted.
E0: Not attempted.

Part (c)
E1: Comment on low response rate leading to bias.
E0: No further relevant comment.

Question 6 - Response B

Figure 5

5%.

(a) Investigate whether there is evidence of an association between whether it was the customer's first experience and the likelihood that the customer would use the company again.

(8)

~~H0: ... all i = 1,2,3~~
~~H1: ... some i = 1,2,3 alternative~~

Ho: An association exists between the customers first experience and the likelihood that they'd use the company again.
H1: There is no association between the customers first experience and the likelihood they'd use the company again.

Expected values.

	Y	N
Det	14.903 22581	6.096774194
Pro	14.19354839	5.806451613
N	14.903 22581	6.096774194

No pooling required.

$\chi^2 = 12.58633877$
df = 2

CV = 1-0.05 = 0.95 → df = 2
CV = 5.991 (usf) ~~TS = 12.58~~
TS = 12.59 (4sf).

Q6

Question 6 continued

TS > CV so Accept Ho at the 5% significance level. There is insufficient evidence to suggest that ~~there exists~~ no association exists between the customers first experience and the likelihood that a customer will buy from the company again.

(c) Give **two** possible sources of bias in this investigation.

(2)

- Some customers may not have/use emails so they weren't able to respond to the survey sent out.
- some customers may have given a fake email.
- older customers are less likely to use/have emails so the responses to the surveys may be heavily dominated by a younger population.
- only 42 responded which may not be representative of the whole sample.

(Total for Question 6 is 12 marks) Total

8 /12

Part (a)

B0: Incorrect hypotheses, they have been written the wrong way round.

M1: At least one expected frequency is correct.

M1: The correct test statistic implies this mark**;** it is advisable to write down the contributions or method in case of an error when using the calculator functions.

M1: The correct test statistic implies this mark.

A1: Test statistic = 12.59 is correct.

B1: Critical value =5.991 is correct.

M1: Comparison of their test statistic and 5.991

E0: Incorrect conclusion.

Part (b)

E0: Not attempted.

E0: Not attempted.

Part (c)

E1: Comment on the use of emails leading to bias.

E1: Comment on low response rate leading to bias.

Question 6 - Response C

Obs
Exp
error

		First experience		
	Obs	Yes	No	Total
Likelihood of using company again	Definitely yes	9	12	21
	Probably yes	18	2	20
	No	17	4	21
	Total	44	18	62

row total X col total / grand total

Figure 5

(a) Investigate whether there is evidence of an association between whether it was the customer's first experience and the likelihood that the customer would use the company again.

(8)

Exp	Yes	NO	total
D yes	14.903	6.097	21
P yes	14.194	5.806	20
No	14.903	6.097	21
Total	44	18	62

Error	Yes	NO	total
D yes	2.338	5.715	8.053
P yes	1.021	2.495	3.516
No	0.2451	0.7212	1.0163
total	3.6541	8.9312	12.5853

$\frac{(O-E)^2}{E}$

$\chi^2 = 125853$

if its

Ho: There is no association between the customers first experience and likelyhood of using the company again

H1: There is an association between the customers first experience and likelyhood of using the company again.

Q6

A
B
C

Question 6 continued

Use a χ^2 test for association at the 5% level using $\nu = 2$

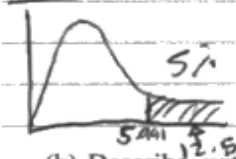
$T_S = 12.5853$

Result is significant

Reject H_0

There is significant evidence to suggest an association between if its the customers first experience and likelyhood of using the company again

CR



(b) Describe, with numerical justification, the nature of any association identified in your conclusion to the test in (a).

(2)

- There were alot more definetly yes' with customes that had been before, with the observed amount at 12 when it was only expected at 6.097.
- There here also alot less probably yes from previous customes, with observed at 2 Compared to expected which was 5.806.

(c) Give **two** possible sources of bias in this investigation.

(2)

- bias towards the people that answered as only 62 out of 1600 responded.
- bias towards those who use email as not everyone does.

11 /12

Q6

A B C

Part (a)

B1: Hypotheses both correct.
M1: All expected frequencies correct.
M1: All contributions correct.
M1: Intention to sum implied by the correct test statistic.
A1: Test statistic = 12.5853 lies within the given range.
B1: Critical value = 5.991 can be seen on the diagram.
M1: Comparison of their test statistic with 5.991
E1: Previous 4 marks awarded and correct conclusion in full context.

Part (b)

E1: The values 12 and 6.097 have been quoted.
E0: A fuller explanation in context was required for this mark e.g. 'There were a lot more customers who have used the company before than expected who would definitely use the company again.'

Part (c)

E1: Comment on low response rate leading to bias.
E1: Comment on the use of emails leading to bias.