Pearson Edexcel

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

Summer 2023 Exemplar
9ST0-02 A level Statistics
Paper 02: Statistical Inference

Senior Examiner’s feedback on student responses

Contents

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

About this booklet

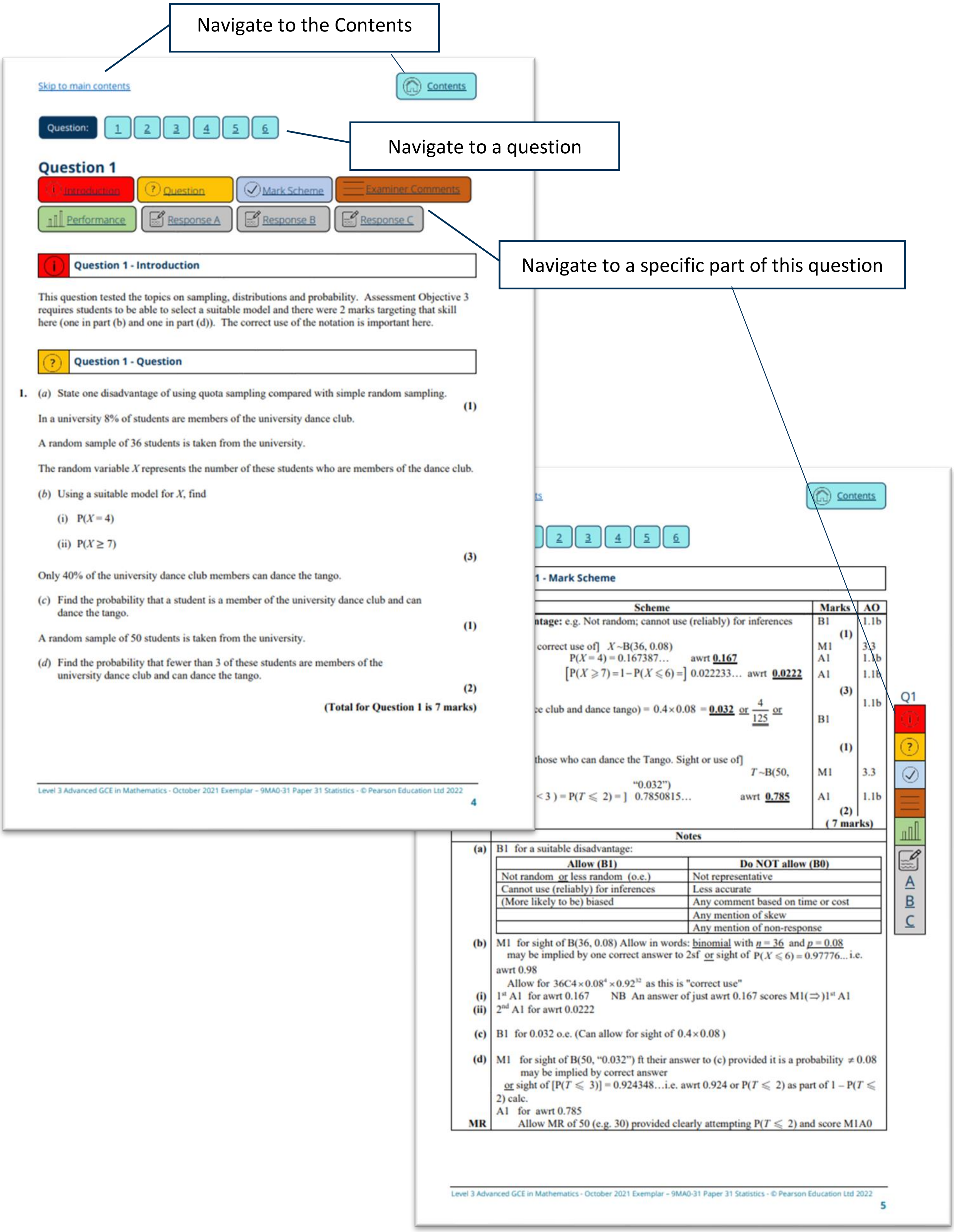
This document has been produced to support mathematics teachers delivering the GCE Statistics specification.

This document looks at questions from the 9ST0-02 A level Statistics Paper 2: Statistical Inference Summer 2023 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. For the mark scheme notes and details of alternative methods please see the full mark scheme for this question paper on our website.

For 2023, the approach all exam boards have taken to grading was to return to pre-pandemic grading, while giving students protection against any impact of disruption. Results in summer 2023 therefore will be far more in line with summer exams that were sat in 2019, but lower than in 2022, when grades awarded were based on a mid-point between 2019 and 2021 outcomes. For more information please read our 'Understanding grade boundaries 2023' document.

* 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.

How to use this booklet



General Examiner Feedback

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

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

It is important for students to be able to comment on reliability and validity to acquire the A03 marks in this specification. Although there were instances in the paper where candidates were successful, for instance identifying the use of Central Limit Theorem in Q03(c), it was apparent that this was a discriminating skill at the top end.

Candidates should continue to be reminded that explanation answers should be given in the context of the question.

Question 1

Question 1 - Question

1 100 professional players qualified to take part in the World Cup for a video game.

In each round, players were awarded points based on their position.

The overall positions in the tournament were determined by the total number of points that the players scored from all the rounds. The winning player gained the highest number of final points.

Keito believes that professional players from countries in Asia are better players than those from other continents.

To test this belief, he obtained a sample of 20 players from those in the World Cup. The final points of the players are shown in **Figure 1**.

Final points of players from countries in Asia	Final points of players from countries not in Asia
32	20
26	17
23	15
19	13
16	12
14	11
4	10
3	6
	5
	2
	1
	0

Figure 1

(a) Carry out a non-parametric test to determine whether there is evidence to support Keito's belief.

(9)

(b) Suggest **one** improvement that could be made to the test in (a).

(1)

Tim suggested using a test on the mean number of points scored by players, from countries in Asia and those not in Asia, using the **t-distribution** to investigate Keito's belief.

(c) Give **one** reason why a non-parametric test was chosen in (a), rather than the test suggested by Tim.

Question 1 – Question (Cont.)

After carrying out his non-parametric test, Keito wrote an article on his blog. An extract from his blog is given in **Figure 2**.

> As my test shows the players from Asian countries in the World Cup did better than those from countries not in Asia, we can conclude that Asian players are better at video games than non-Asian players.
>
> Furthermore it is clear this is due to the living arrangements used by Asian teams to form a team house. The team house culture gives Asian players a huge advantage and is one that the rest of the world has not yet caught up with.

Figure 2

(d) Give **two** criticisms of the comments that Keito has made in his blog.

(2)

(Total for Question 1 is 13 marks)

Question 1 - Mark Scheme

Question	Scheme	Marks	AO	Notes
1(a)	Wilcoxon Rank Sum	M1	2.1a	Used or clearly stated **or** Mann-Whitney (U test) stated
	H_0: $\eta_A = \eta_B$ H_1: $\eta_A > \eta_B$	B1	1.3	H_0**:** Samples from identical populations H_1: Samples from different populations oe

Asia (A)	Not Asia (B)
1 20	4 17
2 19	6 15
3 18	8 13
5 16	10 11
7 14	11 10
9 12	12 9
16 5	13 8
17 4	14 7

Q1

A

B

C

	15 6 18 3 19 2 20 1			
		M1	1.3	Attempt at ranking as one group
		A1	1.3	All ranking correct Or reversed ranks
	$T_A=60$ or $T_B=150$	M1dep	1.3	Effort to obtain T (reverse totals 108,102)
	$U_A = 60 - \frac{1}{2}(8)(9) = 24$ $U_B = 150 - \frac{1}{2}(12)(13) = 72$	A1	1.3	Either U correct
	cv=26 (or 70)	B1	1.3	Either cv correct
	"24" < "26" (or 72>70) Reject H_0	M1	2.1b	Correct conclusion for their ts and their cv
	There is significant evidence at the 5% level that players from Asian countries do better than those from non-Asian countries	E1dep	2.1a	Dep on correct ts & cv
1(b)	**Possible improvements (not exhaustive)** Use a larger sample size Ensure sample is randomly selected Select sample from all professional players not just those taking part in the world cup Use more than one video game Use equal sample sizes Use sample from more than one event	E1	3.1a	Larger than 30 so CLT can be used Any sensible suggestion
1(c)	Numbers of points may not be normally distributed. Positions are discrete rather than continuous.			

Q1
?
A
B
C

	It is not known whether the variances are equal.	E1	3.1a	Any of these
1(d)	Can't necessarily extend to a category that it is a subset of/similar to (all video games) World Cup was the best players from each country – no relevance to a specific player Nothing suggesting it's related to the living situations Keito's statement is too definite	E1, E1	3.1b 3.1b	Any two comments
		Total	**13**	

Q1

A

B

C

Question 1 - Examiner Comments

This question proved to be a good settler for the exam, with many candidates achieving full marks for the Wilcoxon Rank Sum tests in (a). It was pleasing to see several candidates clearly identifying the hypothesis test in their answers, though this was not required in the mark scheme. The use of the wording 'non-parametric test' in the question was identified by many students and enabled them to correctly identify the required test. There were instances of candidates conducting a difference of two mean *t*-test, which could not achieve any marks in (a).

A common misconception across candidates was whether the value of 0 is included in the ranking of the samples. This is the third time that a 0 value has been given in the new specification and it is important that candidates are aware that the **0 is included within the ranking**. Those who did not include it lost accuracy marks.

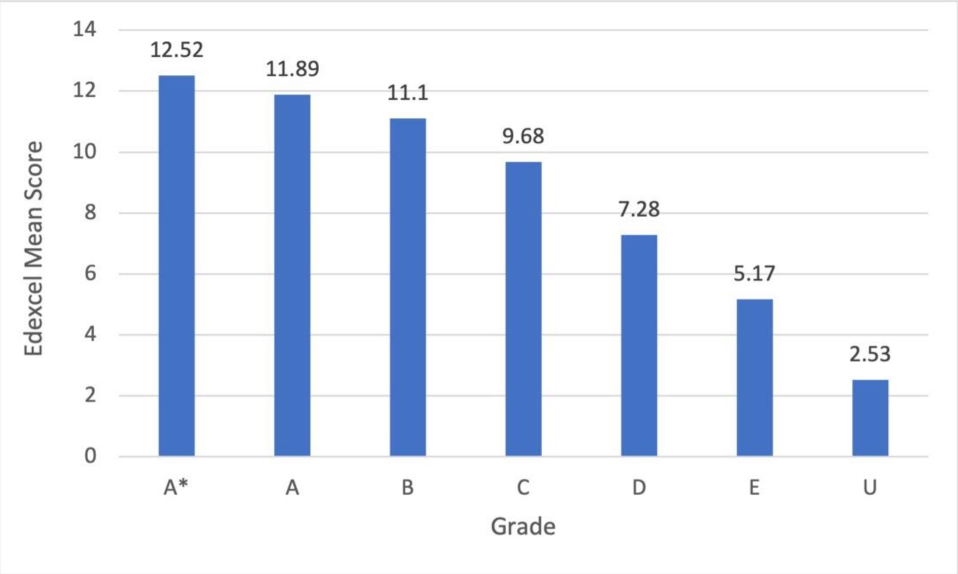
Part (b) was answered very well. Credit was given for having equal sample sizes. However, it should be noted that equal sample sizes would only improve the test if the overall sample size is maintained, or increased. Some candidates suggested an alternative test, it is important that the wording of the question is noted 'an improvement for the test in (a)' therefore answers must be improvements to the Wilcoxon Rank Sum test.

Part (c) was attempted by the majority of candidates with many being able to identify the need for a **normally distributed population**, use of the words it, data or sample were not penalised if used instead of population at this stage in the paper. Credit was also given for candidates who mentioned the need for a **common variance for the populations**.

The final part of this question proved to be the most challenging aspect of the questions. Though it was pleasing to see many candidates picking up at least 1 mark in this part. It is important for candidates to be aware that their responses must be **specific and in context**. With regards to the living arrangements comment there were candidates who said 'this was not valid as the improvement was likely due to practice' this did not gain credit as they themselves have assumed a cause that was not tested in (a). Some candidates criticised the test rather than the comments that had been made.

Question 1 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
8.61	13	66	8.61	12.52	11.89	11.10	9.68	7.28	5.17	2.53



Question 1 - Response A

1 100 professional players qualified to take part in the World Cup for a video game.

In each round, players were awarded points based on their position.

The overall positions in the tournament were determined by the total number of points that the players scored from all the rounds. The winning player gained the highest number of final points.

Keito believes that professional players from countries in Asia are better players than those from other continents.

To test this belief, he obtained a sample of 20 players from those in the World Cup. The final points of the players are shown in **Figure 1**.

Final points of players from countries in Asia	Final points of players from countries not in Asia
32	20
26	17
23	15
19	13
16	12
14	11
4	10
3	6
	5
	2
	1
	0

Figure 1

(a) Carry out a non-parametric test to determine whether there is evidence to support Keito's belief.

(9)

~~Ho: There is no association between Final points and~~ ~~continents~~

~~H1: There is an association between Final points and player~~

$H_0: M_1 = M_2$

$H_1: M_1 > M_2$

$T_1 = 137$ $T_2 = 112$ $T.S = 46$

$n_1 = 8$ $n_2 = 12$ $C.V = 26$

$137 - \frac{8 \times 7}{2} = 109$ $112 - \frac{12 \times 11}{2} = 46$

$46 > 26$ Do not Reject Ho

$U_1 = 109$ $U_2 = 46$

Question 1 continued

There is sufficient evedence to reject the claim that Asian players are better than others.

(b) Suggest **one** improvement that could be made to the test in (a).

(1)

have equal amounts of data

Tim suggested using a test on the mean number of points scored by players, from countries in Asia and those not in Asia, using the ***t*-distribution** to investigate Keito's belief.

(c) Give **one** reason why a non-parametric test was chosen in (a), rather than the test suggested by Tim.

(1)

~~because there weren't equal amounts of data~~

The two factors are not paird data

After carrying out his non-parametric test, Keito wrote an article on his blog. An extract from his blog is given in **Figure 2**.

> As my test shows the players from Asian countries in the World Cup did better than those from countries not in Asia, we can conclude that Asian players are better at video games than non-Asian players.
>
> Furthermore it is clear this is due to the living arrangements used by Asian teams to form a team house. The team house culture gives Asian players a huge advantage and is one that the rest of the world has not yet caught up with.

Figure 2

(d) Give **two** criticisms of the comments that Keito has made in his blog.

(2)

Too quick to conclude, hasent done any ~~[illegible]~~ verifying. Could use more data than one word cup. Has not done anything to support his argument in second paragraph

6 /13

Part (a)

M1: Evidence of Wilcoxon Rank Sum awarded due to the U calculation.
B1: Awarded, though groups should have been clearly labelled on the table as 1 and 2.
M0: No attempt to rank the data.
A0: No attempt to rank the data.
M0: No attempt to sum the ranks.
A0: No correct test statistic value.
B1: Obtained correct critical value from the formula booklet.
M1: Correctly compared their test statistic with their critical value.
E0: Conclusion has context based on the information from the question, but this final mark was dependent on correct test statistic and critical value.

Q1 ? A B C

Part (b)

E1: Awarded. However, it should be noted that this is only an improvement if the overall sample size for the groups either remains the same or increases.

Part (c)

E0: Response is not about either the need for a normal population or a common variance between the populations.

Part (d)

E1: Awarded for the second comment.

E0: The first comment is not specific enough. It is presumed that the candidate may have been trying to bring up the definitiveness of the comment.

Question 1 - Response B

1 100 professional players qualified to take part in the World Cup for a video game.

In each round, players were awarded points based on their position.

The overall positions in the tournament were determined by the total number of points that the players scored from all the rounds. The winning player gained the highest number of final points.

Keito believes that professional players from countries in Asia are better players than those from other continents.

To test this belief, he obtained a sample of 20 players from those in the World Cup. The final points of the players are shown in **Figure 1**.

Final points of players from countries in Asia	Final points of players from countries not in Asia
20 32	17 20
19 26	15 17
18 23	13 15
16 19	11 13
14 16	10 12
12 14	9 11
5 4	8 10
4 3	7 6
↓	6 5
108	3 2
	2 1
	1 0

↓ 102

Figure 1

(a) Carry out a non-parametric test to determine whether there is evidence to support Keito's belief.

(9)

H_1: there is an association between the continent they're from and the skill of the player.

H_0: There is not an association between the continent they're from & the skill of the player.

$\frac{1}{2} \cancel{8} + 8(8+1) - 6 = 66.5$

$\frac{1}{2} + 12(\cancel{8}2+1) - 17 = 139.5$

$\cancel{108 - \frac{20(20+1)}{2}} =$

$60 < 66.5$

Reject Ho as there is clearly an association between Continent & Skill of the player.

(b) Suggest **one** improvement that could be made to the test in (a). **(1)**

~~Improves~~ Increase the Sample Size.

Tim suggested using a test on the mean number of points scored by players, from countries in Asia and those not in Asia, using the ***t*-distribution** to investigate Keito's belief.

(c) Give **one** reason why a non-parametric test was chosen in (a), rather than the test suggested by Tim. **(1)**

The data is not Symetrical & we Can not gaurantee it has Come from a random Sample.

After carrying out his non-parametric test, Keito wrote an article on his blog. An extract from his blog is given in **Figure 2**.

> As my test shows the players from Asian countries in the World Cup did better than those from countries not in Asia, we can conclude that Asian players are better at video games than non-Asian players.
>
> Furthermore it is clear this is due to the living arrangements used by Asian teams to form a team house. The team house culture gives Asian players a huge advantage and is one that the rest of the world has not yet caught up with.

Figure 2

(d) Give **two** criticisms of the comments that Keito has made in his blog. **(2)**

Living arrangements are not mentioned so how can you Conclude this assumption to be true. ~~Valued~~. No numerical justification given

6 /13

Part (a)
M1: Evidence of Wilcoxon Rank Sum awarded due to the ranking of the 2 sets of data.
B0: Incorrect hypotheses.
M1: Attempt for ranking the data as a whole.
A1: Correct ranking of the data as a whole.
M1: Attempt to add the ranks together for each group.

Q1

A
B
C

A0: Incorrect test statistics calculated using the U formula.
B0: Incorrect critical value identified from formula booklet.
M0: Incorrectly compared their test statistic with their critical value.
E0: Conclusion was dependent on the correct test statistic and correct critical value.

Part (b)
E1: Correct improvement.

Part (c)
E0: Although the comment mentions symmetry this was not enough to be awarded the mark.

Part (d)
E1: Awarded for the comment that there is no numerical justification for the living arrangements statement.
E0: No second criticism given by candidate.

Skip to Main Contents

Question: 1 2 3 4 5 6 7

Question 1 - Response C

(a) Carry out a non-parametric test to determine whether there is evidence to support Keito's belief.

(9)

Rank-Sum

$H_0 \rightarrow$ Professional players from countries in Asia are just as good as players from countries not in Asia

$H_1 \rightarrow$ Professional players from countries in Asia are better than professional players from countries not in Asia

5% significance 1 tail

Countries in Asia (A) Points	Ranks	Countries not in Asia (B) Points	Ranks
32	20	20	17
26	19	17	15
23	18	15	13
19	16	13	11
16	14	12	10
14	12	11	9
4	5	10	8
3	4	6	7
		5	6
		2	3
		1	2
		0	1

$T_A = 108 \quad n_A = 8$

$T_B = 102 \quad n_B = 12$

$U_A = 108 - \frac{8(8+1)}{2} \qquad U_B = 102 - \frac{12(12+1)}{2}$

$= 108 - \frac{72}{2} \qquad = 102 - \frac{132}{2}$

$U_A = 72 \qquad U_B = 36 \qquad CV_{12}^{8} = 26$

$36 > 26 \quad \therefore$ Accept H_0

Insufficient evidence at 5% tesignificance level to suggest that professional players from countries in Asia are better than professional players from countries not in Asia

Q1

?

A
B
C

(b) Suggest **one** improvement that could be made to the test in (a).

(1)

Larger sample size

Tim suggested using a test on the mean number of points scored by players, from countries in Asia and those not in Asia, using the ***t*-distribution** to investigate Keito's belief.

(c) Give **one** reason why a non-parametric test was chosen in (a), rather than the test suggested by Tim.

(1)

The data ~~ri~~ is discrete and therefore not normally distributed so t-test is not suitable to use a t-test.

After carrying out his non-parametric test, Keito wrote an article on his blog. An extract from his blog is given in **Figure 2**.

> As my test shows the players from Asian countries in the World Cup did better than those from countries not in Asia, we can conclude that Asian players are better at video games than non-Asian players.
>
> Furthermore it is clear this is due to the living arrangements used by Asian teams to form a team house. The team house culture gives Asian players a huge advantage and is one that the rest of the world has not yet caught up with.

Figure 2

(d) Give **two** criticisms of the comments that Keito has made in his blog.

(2)

There ~~was~~ was not sufficient evidence to prove that ~~e-tay~~ Asian players performed better than non-Asian players

11 /13

Q1

?

A

B

C

Part (a)

M1: Evidence of Wilcoxon Rank Sum awarded due to the ranking of the 2 sets of data, as well as stating 'Rank Sum'

B1: Although not written using symbols this mark was awarded. However, there should have been mention of an average to have an improved answer.

M1: Attempt for ranking the data as a whole.

A1: Correct ranking of the data as a whole.

M1: Attempt to add the ranks together for each group.

A1: Although one test statistic is incorrect, 72 was correct.

B1: Correct critical value identified from formula booklet.

M1: Correctly compared their test statistic with their critical value.

E0: Conclusion has context based on the information from the question, but this final mark was dependent on correct test statistic and critical value.

Part (b)
E1: Correct improvement.

Part (c)
E1: Awarded for normal distribution comment.

Part (d)
E1: Awarded for the comment that there is no numerical justification for the living arrangements statement.
E0: No second criticism given by candidate.

Question 2

Question 2 - Question

2 It is believed that colour preference in children is influenced by many factors.

Pippa is a child psychologist who is interested in investigating whether home location affects colour preference.

Colour preferences were tested in a sample of females aged 4-11 years in four different home locations:

- Shipibo villages in the Peruvian Amazon,
- Kastom villages in the highlands of Tanna Island, Vanuatu in the Pacific Islands,
- BaYaka villages in the northern Republic of Congo,
- Brisbane, a large city in Australia.

The results of this experiment are summarised in **Figure 3**.

Observed Frequencies		Colour Preference						
		Light Pink	Pink	Red	Purple	Blue	Light Blue	Total
Home Location	Shipibo	8	11	9	6	14	7	55
	Kastom	6	9	14	15	11	9	64
	BaYaka	4	8	13	5	14	7	51
	Brisbane	15	32	17	18	11	5	98
	Total	33	60	53	44	50	28	268

[Data Source: https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.13528]

Figure 3

Question 2 – Question (Cont.)

Some of the expected values for the χ^2 test for an association are shown in **Figure 4**.

Expected Frequencies		Colour Preference					
		Light Pink	Pink	Red	Purple	Blue	Light Blue
Home Location	Shipibo		12.31	10.88	9.03	10.26	5.75
	Kastom		14.33	12.66	10.51	11.94	6.69
	BaYaka		11.42	10.09	8.37	9.51	5.33
	Brisbane		21.94	19.38	16.09	18.28	10.24

Figure 4

(a) Complete the table in **Figure 4**.

(2)

Some of the contributions to the χ^2 test statistic are given in **Figure 5.**

		Colour Preference					
		Light Pink	Pink	Red	Purple	Blue	Light Blue
Home Location	Shipibo	0.223	0.140	0.324		1.362	0.274
	Kastom	0.449	1.981	0.143		0.074	0.800
	BaYaka	0.828	1.023	0.842		2.114	0.524
	Brisbane	0.713	4.612	0.292		2.902	2.680

Figure 5

(b) Test whether colour preference of females aged 4-11 years is independent of home location.

(6)

(c) Describe the nature of any association found in (b)

You should include numerical justification.

(2)

It was later discovered that the four samples of children did not have the same proportions of each of the ages from 4 years to 11 years.

(d) Explain why the conclusion in (b) is now likely **not** to be valid and suggest a possible modification to Pippa's experimental design.

(2)

(Total for Question 2 is 12 marks)

Q2 A B C

Question 2 - Mark Scheme

Question	Scheme	Marks	AO	Notes

	Light pink	Pink	Red	Purple	Blue	Light Blue
Shipibo	6.77	12.31	10.88	9.03	10.26	5.75
Kastom	7.88	14.33	12.66	10.51	11.94	6.69
BaYaka	6.28	11.42	10.09	8.37	9.51	5.33
Brisbane	12.07	21.94	19.38	16.09	18.28	10.24

Question	Scheme	Marks	AO	Notes
2(a)	See table above	M1	1.3	Any correct value in bold box to at least 1dp
		A1	1.3	All calculated expected values correct and shown

0.223	0.140	0.324	1.017	1.362	0.274
0.449	1.981	0.143	1.918	0.074	0.800
0.828	1.023	0.842	1.357	2.114	0.524
0.713	4.612	0.292	0.227	2.902	2.680

Question	Scheme	Marks	AO	Notes
2(b)	$\frac{(6-9.03)^2}{9.03}+\cdots+\frac{(18-16.09)^2}{16.0009}$	M1	1.3	PI Method for contributions correct Give mark for one correct value seen Condone small slip
	$= 26.819$	A1	1.3	awfw 25~28
	H_0: Colour preference and home location **are independent** H_1: Colour preference and home location **are not independent**	B1	1.3	oe Both hypotheses allow association
	cv = 24.996	B1	1.3	awrt 25
	"26.824" > "24.996" Reject H_0	M1	2.1b	Stated comparison of their ts and their cv or 0.0302 < 0.05

Question	Scheme	Marks	AO	Notes
	There is sufficient evidence that the colour preference of females aged 4-11 years is not independent of their home location.	E1dep	2.1a	oe Conclusion in context Dep on correct cv and ts
2(c)	Females aged 4-11 years in Brisbane were more likely to like pink than expected	E1	2.1b	
	4.612 is the largest contribution to the test statistic	B1	2.1b	PI by 32 and 21.94
2(d)	The association for colour preference may be based on age rather than home location	E1	3.1b	
	Select ages of females in equal ratios across 4-11 years.	E1	2.1b	
	Total	**12**		

Question 2 - Examiner Comments

Again, the contingency table method has proved to be accessible to many candidates. Though some students struggled with the hypothesis test being broken up across several pages of the paper.

Part (a) was attempted by a significant number of candidates and the majority scored these marks. It would be useful to remind candidates to **double check** calculations on these more accessible questions.

The completion of the hypothesis test in (b) was done well by many candidates using a variety or methods. For those who are using the calculator it is important that they **write down the test statistic contributors**. This is important so that method marks can be awarded if errors have been made in the calculator work, but also so that contributors can be identified for association questions which may follow, as in (c)

As mentioned in last years report the main errors of candidates not gaining full credit were for hypotheses being reversed (this was seen more frequently in candidates using the word independence rather than association) or selection of the incorrect critical value.

Part (c) was answered well by the top end of candidates. For association questions it is required for the candidate to identify **the largest test statistic contributor**, those who scored well on this part identified the test statistic as well as the relevant observed and expected values.

Candidates well practised in exam technique were able to pick up on the key word of '**describe**' in the question. Therefore, those who simply identified Brisbane and Pink did not gain credit, there needed to be an attempt at saying that more females chose pink in Brisbane than would be expected.

Q2

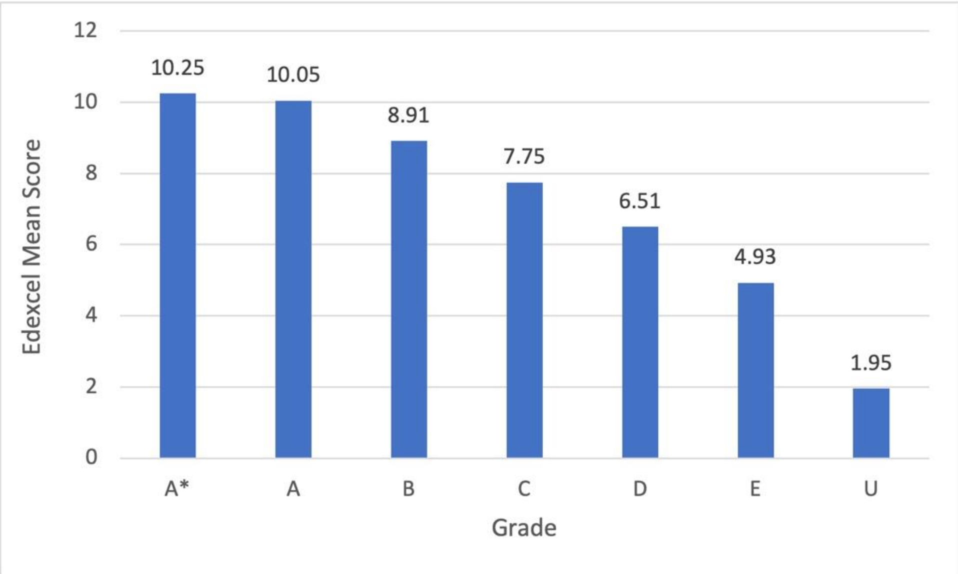
A
B
C

Again, the most challenging part of the question was (d). The first mark required candidates to identify that the association found may now be due to age rather than location, bringing in the concept of gender stereotypes was a common incorrect answer unless the candidate directly explained the effect on the conclusion. The second mark was achieved more frequently, though it was important that candidates specifically mentioned dealing with the ages in their answer rather than other possible variables.

'Introduce a blocking factor of age' was a very common incorrect answer for (d)

Question 2 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
7.22	12	60	7.22	10.25	10.05	8.91	7.75	6.51	4.93	1.95



Question 2 - Response A

2 It is believed that colour preference in children is influenced by many factors.

Pippa is a child psychologist who is interested in investigating whether home location affects colour preference.

Colour preferences were tested in a sample of females aged 4-11 years in four different home locations:

- Shipibo villages in the Peruvian Amazon,
- Kastom villages in the highlands of Tanna Island, Vanuatu in the Pacific Islands,
- BaYaka villages in the northern Republic of Congo,
- Brisbane, a large city in Australia.

The results of this experiment are summarised in **Figure 3**.

Observed Frequencies		Colour Preference						
		Light Pink	**Pink**	**Red**	**Purple**	**Blue**	**Light Blue**	**Total**
Home Location	**Shipibo**	8	11	9	6	14	7	55
	Kastom	6	9	14	15	11	9	64
	BaYaka	4	8	13	5	14	7	51
	Brisbane	15	32	17	18	11	5	98
	Total	33	60	53	44	50	28	268

[Data Source: https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.13528]

Figure 3

Some of the expected values for the χ^2 test for an association are shown in **Figure 4**.

Expected Frequencies		Colour Preference					
		Light Pink	**Pink**	**Red**	**Purple**	**Blue**	**Light Blue**
Home Location	**Shipibo**	7.25	12.31	10.88	9.03	10.26	5.75
	Kastom	7.87	14.33	12.66	10.51	11.94	6.69
	BaYaka	6.28	11.42	10.09	8.37	9.51	5.33
	Brisbane	12.07	21.94	19.38	16.09	18.28	10.24

Figure 4

Q2 ? A B C

Question 2 continued

(a) Complete the table in **Figure 4**.

(2)

$\alpha = 0.05$

Shipibo = 12.31 + 10.58 + 9.03 + 10.26 + 5.75 = 48.23, 55 − 48.23 = 7.25

Kastom = 14.33 + 12.66 + 10.51 + 11.94 + 6.69 = 56.13, 64 − 56.13 = 7.87

BaYaka = 11.42 + 10.09 + 8.37 + 9.51 + 5.33 = 44.72, 51 − 44.72 = 6.28

Brisbane = 21.94 + 19.38 + 16.09 + 18.28 + 10.24 = 85.93, 98 − 85.93 = 12.07

Question 2 continued

Some of the contributions to the χ^2 test statistic are given in **Figure 5**.

		Colour Preference					
		Light Pink	Pink	Red	Purple	Blue	Light Blue
Home Location	Shipibo	0.223	0.140	0.324	1.017	1.362	0.274
	Kastom	0.449	1.981	0.143	1.9182	0.074	0.800
	BaYaka	0.828	1.023	0.842	1.3569	2.114	0.524
	Brisbane	0.713	4.612	0.292	0.2267	2.902	2.680

Figure 5

(b) Test whether colour preference of females aged 4-11 years is independent of home location.

(6)

$\alpha = 0.05$

~~Shtbr~~

Shipibo + Purple $= \dfrac{(6 - 9.03)^2}{9.03}$

$= 1.017$

Kastom + Purple $= \dfrac{(15 - 10.51)^2}{10.51}$

$= 1.9182$

Ba Yaka + Purple $= \dfrac{(5 - 8.37)^2}{8.37}$

$= 1.3569$

Brisbane + Purple $= \dfrac{(15 - 16.09)^2}{16.09}$

$= 0.2267$

(χ^2)

CV = 13.091

$n = 24$

$v = 24 - 1$

$\sum \dfrac{(O-E)^2}{E}$

$= 26.3688$

H_0: not independent

H_1: independent

Question 2 continued

(c) Describe the nature of any association found in (b)

You should include numerical justification.

(2)

$26.3688 > 13.091$ so Reject H_0

significant evidence at a 5% level to suggest that the colour preference of females aged 4-11 years is independent of home location.

It was later discovered that the four samples of children did not have the same proportions of each of the ages from 4 years to 11 years.

(d) Explain why the conclusion in (b) is now likely **not** to be valid and suggest a possible modification to Pippa's experimental design.

(2)

not valid as it is

~~an~~ would be less generalisable to those that weren't as represented in terms of age. I would recommend to PIPPA getting a larger sample of those that are not as proportionate to fill out the difference.

4 /12

Part (a)
M1: Attempt at calculating the expected value.
A0: One the four expected values is incorrect.

Part (b)
M1: Attempt at calculating a test statistic contributor.
A1: Correct test statistic obtained.
B0: Hypotheses were given the wrong way round.
B0: Incorrect critical value obtained from formula booklet (given in (c)).
M1: Correct comparison between their test statistic and their critical value.
E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (c)
E0: No attempt made for (c). The response is the final marks of the hypothesis test (which were awarded previously)
B0: No attempt made for (c). The response is the final marks of the hypothesis test (which were awarded previously)

Part (d)
E0: Candidate did not address the comment on the validity.
E0: Although proportion is mentioned the candidate is not specific about this. 'Larger sample' gains no credit as it does not deal with the issue of age affecting the result.

Question 2 - Response B

2 It is believed that colour preference in children is influenced by many factors.

Pippa is a child psychologist who is interested in investigating whether home location affects colour preference.

Colour preferences were tested in a sample of females aged 4-11 years in four different home locations:

- Shipibo villages in the Peruvian Amazon,
- Kastom villages in the highlands of Tanna Island, Vanuatu in the Pacific Islands,
- BaYaka villages in the northern Republic of Congo,
- Brisbane, a large city in Australia.

The results of this experiment are summarised in **Figure 3**.

Observed Frequencies		Colour Preference						
		Light Pink	Pink	Red	Purple	Blue	Light Blue	Total
Home Location	Shipibo	8	11	9	6	14	7	55
	Kastom	6	9	14	15	11	9	64
	BaYaka	4	8	13	5	14	7	51
	Brisbane	15	32	17	18	11	5	98
	Total	33	60	53	44	50	28	268

[Data Source: https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.13528]

Figure 3

Some of the expected values for the χ^2 test for an association are shown in **Figure 4**.

Expected Frequencies		Colour Preference					
		Light Pink	Pink	Red	Purple	Blue	Light Blue
Home Location	Shipibo	6.77	12.31	10.88	9.03	10.26	5.75
	Kastom	7.88	14.33	12.66	10.51	11.94	6.69
	BaYaka	6.30	11.42	10.09	8.37	9.51	5.33
	Brisbane	12.07	21.94	19.38	16.09	18.28	10.24

Figure 4

Q2 ? A B C

(a) Complete the table in **Figure 4**.

(2)

Some of the contributions to the χ^2 test statistic are given in **Figure 5**.

		Colour Preference					
		Light Pink	**Pink**	**Red**	**Purple**	**Blue**	**Light Blue**
Home Location	**Shipibo**	0.223	0.140	0.324	1.017	1.362	0.274
	Kastom	0.449	1.981	0.143	1.918	0.074	0.800
	BaYaka	0.828	1.023	0.842	1.357	2.114	0.524
	Brisbane	0.713	4.612	0.292	0.227	2.902	2.680

Figure 5

(b) Test whether colour preference of females aged 4-11 years is independent of home location.

(6)

H_0: Colour preference of females aged 4-11 years is not independent of home location.

H_1: Colour preference of females aged 4-11 years is ~~not~~ independent of home location.

1 tailed test $\alpha = 0.05$

$\Sigma e = 26.819$ $n = 24$ $CV = 36.415$

$26.819 < 36.415$ $\therefore$ Reject H_0,

Sufficient evidence at a 5% level to show colour preferance of females aged 4-11 years is ~~[illegible]~~ independent of home location.

Q2

A

B

C

(c) Describe the nature of any association found in (b)

You should include numerical justification.

(2)

The association found in b is very significant as the test statistic is significantly lower than the critical value.

It was later discovered that the four samples of children did not have the same proportions of each of the ages from 4 years to 11 years.

(d) Explain why the conclusion in (b) is now likely **not** to be valid and suggest a possible modification to Pippa's experimental design.

(2)

As there may be a prefferred colour between ages ~~and~~ which would invalidate the results. Pippa could perhaps do a stratified sample instead to get more accurate results of the whole population.

4 /12

Part (a)
M1: Attempt at calculating the expected value.
A1: All four expected values were correct.

Part (b)
M1: Attempt at calculating a test statistic contributor.
A1: Correct test statistic obtained.
B0: Hypotheses were given the wrong way round.
B0: Incorrect critical value obtained from formula booklet.
M0: Incorrect comparison between their test statistic and their critical value.
E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (c)
E0: No attempt to describe where there was large association observed between the groups.
B0: No attempt to use test statistic contributors to identify association.

Part (d)
E0: Although candidate has said that ages may affect colour which would invalidate the results. This was implied through the question. More detail was needed to be awarded this mark.
E0: Candidate has suggested a stratified sample but has not identified the need for age to be a variable considered when selecting it.

Question 2 - Response C

2 It is believed that colour preference in children is influenced by many factors.

Pippa is a child psychologist who is interested in investigating whether home location affects colour preference.

Colour preferences were tested in a sample of females aged 4-11 years in four different home locations:

- Shipibo villages in the Peruvian Amazon,
- Kastom villages in the highlands of Tanna Island, Vanuatu in the Pacific Islands,
- BaYaka villages in the northern Republic of Congo,
- Brisbane, a large city in Australia.

The results of this experiment are summarised in **Figure 3**.

Observed Frequencies		Colour Preference						
		Light Pink	**Pink**	**Red**	**Purple**	**Blue**	**Light Blue**	**Total**
Home Location	**Shipibo**	8	11	9	6	14	7	55
	Kastom	6	9	14	15	11	9	64
	BaYaka	4	8	13	5	14	7	51
	Brisbane	15	32	17	18	11	5	98
	Total	33	60	53	44	50	28	268

[Data Source: https://srcd.onlinelibrary.wiley.com/doi/full/10.1111/cdev.13528]

Figure 3

Some of the expected values for the χ^2 test for an association are shown in **Figure 4**.

Expected Frequencies		Colour Preference					
		Light Pink	**Pink**	**Red**	**Purple**	**Blue**	**Light Blue**
Home Location	**Shipibo**	6.77	12.31	10.88	9.03	10.26	5.75
	Kastom	7.88	14.33	12.66	10.51	11.94	6.69
	BaYaka	6.28	11.42	10.09	8.37	9.51	5.33
	Brisbane	12.07	21.94	19.38	16.09	18.28	10.24

Figure 4

Q2

?

A
B
C

Question 2 continued

Some of the contributions to the χ^2 test statistic are given in **Figure 5.**

		Colour Preference					
		Light Pink	**Pink**	**Red**	**Purple**	**Blue**	**Light Blue**
Home Location	**Shipibo**	0.223	0.140	0.324	1.017	1.362	0.274
	Kastom	0.449	1.981	0.143	1.918	0.074	0.800
	BaYaka	0.828	1.023	0.842	1.357	2.114	0.524
	Brisbane	0.713	4.612	0.292	0.227	2.902	2.680

Figure 5

(b) Test whether colour preference of females aged 4-11 years is independent of home location.

(6)

H0: There is no association between colour preferences of females aged 4-11 and Home location.

H1: There is an association between colour preferences of females aged 4-11 and home location.

$\frac{(6-9.03)^2}{9.03} = 1.017$ $\quad \frac{(15-10.51)^2}{10.51} = 1.918$

$\frac{(5-8.37)^2}{8.37} = 1.357$ $\quad \frac{(18-16.09^2)}{16.09} = 0.227$

$\sum \frac{(O-E)^2}{E} = \chi^2 = 26.82$

5% SL, 2 tail = ~~0.025~~ 0.05, df = 15 $\quad$ CV = ~~27.49~~ ~~24.99~~ 25.00

~~26.82 < 27.49, do not reject H0~~ $\quad$ 26.82 > 25.00, reject H0.

There is significant evidence to suggest that there is an association between colour of females aged 4-11 and home location.

(c) Describe the nature of any association found in (b)

You should include numerical justification.

(2)

The association likely came from people who liked pink because they had the highest contribution total, and there were a lot ~~the~~ more people who liked pink in Brisbane (32) than expected (21.41). There were a lot less people who liked bright blue from Brisbane (3) than expected (10.24).

It was later discovered that the four samples of children did not have the same proportions of each of the ages from 4 years to 11 years.

(d) Explain why the conclusion in (b) is now likely **not** to be valid and suggest a possible modification to Pippa's experimental design.

(2)

Because all of the proportions must be the same to carry out an association test. Pippa could instead only use ~~people~~ children from one specific age.

11 /12

Q2

A
B
C

Part (a)
M1: Attempt at calculating the expected value.
A1: All four expected values calculated correctly.

Part (b)
M1: Attempt at calculating a test statistic contributor.
A1: Correct test statistic obtained.
B1: Correct hypotheses.
B1: Correct critical value obtained from formula booklet. Would be preferred as an unrounded figure.
M1: Correct comparison between their test statistic and their critical value.
E1: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (c)
E1: The candidate has described the association by stating there were a lot **more** people who liked pink in Brisbane than expected.
B1: The candidate identified the expected and observed values for the largest test statistic contributor. Would be improved by quoting the largest contributor as well.

Part (d)
E0: Candidate did not address the comment on the validity. It is not a requirement to have the proportions the same for a contingency test. However, the proportions being the same would be required to know that the association was due to location, and not age.
E1: A reasonable suggestion given as to how to adapt the design to eliminate age as a variable.

Question 3

Question 3 - Question

3 On the 25th June 2019 the BBC wrote an article comparing key statistics for the highest leagues, Premier and Super, of men's and women's football in England.

Within this article it was stated that there were more goals scored per match in the women's games than in the men's games.

[Source: https://www.bbc.co.uk/news/uk-48742850]

Paul is interested in investigating whether the claim made in 2019 is still true. He found that the mean number of goals scored per match for the 2021/22 season in the men's Premier League was 2.82

He then gathered data on numbers of goals scored for a random sample of the women's Super League matches for the 2021/22 season. The results are shown below.

1 2 3 5 7 4 0 4 5 3 1 0

[Source: https://www.soccerstats.com/table.asp?league=england17&tid=d]

(a) Making any necessary assumptions, perform a hypothesis test to investigate whether, for the 2021/22 season, there were more goals scored per match, on average, in the women's Super League than in the men's Premier League.

(6)

A separate claim, made in the same article, is that the women break the rules in football **less** often than men.

A yellow card is given to a player as a warning about involvement in a rule breaking incident.

During the men's Premier League's 2021/22 season, the mean number of yellow cards per player was found to be 3.37

Paul gathered data for the number of yellow cards given to each of a random sample of 100 female players playing in the women's Super League for the same season. Using his data, Paul found that the mean number of yellow cards per player was 1.93 and the **variance** was 2.76

(b) Perform a test to investigate whether women in the Super League, on average, are given **fewer** yellow cards than men in the Premier League.

(5)

(c) Explain why there was no need to make any assumptions with regards to the population distribution when carrying out the test in (b)

(2)

Question 3 – Question (Cont.)

(d) Write a short article suitable for the general public comparing the highest leagues of men's and women's football in England.

You should ensure that your answer includes

- references to your conclusions in (a) and (b)
- relevant comments with regards to the original sample data.

(4)

(Total for Question 3 is 17 marks)

Question 3 - Mark Scheme

Question	Scheme	Marks	AO	Notes
3(a)	$H_0: \mu = 2.82$ $H_1: \mu > 2.82$	B1	1.3	Both hypotheses
	$\bar{x} = 2.917$ $s = 2.193$	B1	1.2	PI Both correct $\bar{x}$ awrt 2.92 s awrt 2.2
	$\text{ts} = \dfrac{\text{"2.917"} - 2.82}{\frac{\text{"2.193"}}{\sqrt{12}}}$	M1	1.3	PI Method for ts
	= 0.153	A1ft	1.3	PI awfw 0.15~0.16 **or** p-value = 0.441
	"0.153" < 1.796, Do not reject H_0	M1	2.1b	Comparison of their ts and correct cv **or** 0.441 > 0.05 **or** use of population mean
	There is insufficient evidence to suggest that the average number of goals scored per match is greater for Women's Super League than men's Premier League	E1dep	2.1a	Conclusion in context dep correct cv and ts

Question	Scheme	Marks	AO	Notes
Alternative 1				
3(a)	$H_0: \mu = 2.82$ $H_1: \mu > 2.82$	(B1)		Both hypotheses
	$\bar{x} = 2.917$ $s = 2.193$	(B1)		PI Both correct $\bar{x}$ awrt 2.92 s awrt 2.2
	cv $= 2.82 + 1.796 \times \frac{\text{"2.193"}}{\sqrt{12}}$	(M1)		PI Method for cv
	3.957	(A1ft)		PI awrt 3.96 **or** p-value = 0.441 ft only for incorrect rounding of sample mean and sd not t
	"2.917" < 3.957, Do not reject H_0	(M1)		Comparison of their ts and correct cv **or** 0.441 > 0.05 **or** use of population mean
	There is insufficient evidence to suggest that the average number of goals scored per match is greater for women's Super League than men's Premier League	(E1dep)		Conclusion in context dep correct cv and ts
Alternative 2				
3(a)	$H_0: \eta_W = 2.82$ $H_1: \eta_W > 2.82$	(B1)		oe H_0: (Population) median for women = 2.82 H_1: (Population) median for women > 2.82

Diff	-1.82	-0.82	+0.18	+2.18	+4.18	+1.18	-2.82	+1.18	+2.18	+0.18	-1.82	-2.82
Ranks	6.5	3	1.5	8.5	12	4.5	10.5	4.5	8.5	1.5	6.5	10.5

	Differences (see table above)	(M1)		Attempt at calculating differences

Q3
?
A
B
C

	Ranks (see table above)	(B1)		Attempt at ranking the differences
	$T_+=41$ or $T_-=37$	(A1)		Either T correct
	"37">17 Do not reject H_0	(M1)		Comparison their smaller T with 17 oe 61
	There is insufficient evidence to suggest that the average number of goals scored per match is greater for Women's Super League than Men's Premier League	(E1dep)		Dep on correct ts & cv
3(b)	$H_0: \mu = 3.37$ $H_1: \mu < 3.37$	B1	1.3	Both hypotheses
	ts = $\frac{1.93-3.37}{\sqrt{\frac{2.76}{100}}}$	M1	1.3	PI Method for ts
	= -8.67	A1	1.3	PI **or** p-value ≈ 0
	"-8.67" < -1.6449, Reject H_0	M1	2.1b	Comparison of their ts and correct cv **or** 0 < 0.05 **or** use of population mean
	There is sufficient evidence to suggest that the average number of yellow cards per player is fewer for women's Super League than men's Premier League	E1dep	2.1a	Conclusion in context dep correct cv and ts
Alternative				
3(b)	$H_0: \mu = 3.37$ $H_1: \mu < 3.37$	(B1)		Both hypotheses
	Standard error = $\sqrt{\frac{2.76}{100}}$	(M1)		PI
	cv = 3.10	(A1)		PI **or** p-value ≈ 0
	"1.93" < 3.10, Reject H_0	(M1ft)		Comparison of their ts and correct cv

	There is sufficient evidence to suggest that the average number of yellow cards per player is fewer for Women's Super League than men's Premier League	(E1dep)		**or** $0 < 0.05$ **or** use of population mean Conclusion in context dep correct cv and ts
3(c)	Since sample is large and random therefore CLT applies	E1	3.1a	ref to $n > 30$ or CLT
	Therefore, population distribution is not required to be known (normal)	E1dep	3.1a	
3(d)	There is no evidence to suggest that women score more goals per game than men (based on the Super and Premier League)	E1ft	2.1b	Comment with regards to (a)
	However, in 7 out of the 12 womens' matches the number of goals per match was higher than the men's average.	E1	2.1b	Any relevant comment with regards to original sample data provided. e.g. The sample is only for one year with justification
	There is evidence to suggest that women do receive fewer yellow cards than men (based on the Super and Premier League)	E1ft	2.1b	Comment with regards (b)
		E1	2.1b	Style appropriate to general public. No difficult statistical vocabulary. No oversimplification.
	Total	**17**		

Question 3 - Examiner Comments

This question was particularly challenging, though it was pleasing to see that the vast majority of candidates had attempted all parts of the question.

The paper did not specifically state a particular test in (a). Therefore, this was left for interpretation by candidates. However, it is important that candidates are encouraged to select the **most powerful** test with **fewest assumptions** where appropriate, this was reflected in the marking. Those who attempted either a t-test or Wilcoxon signed rank test could achieve all 6 marks. Those who attempted a z-test achieved a maximum of 4 marks (2 assumptions made) and those who attempted the sign test achieved a maximum of 3 marks (no assumptions but least powerful).

For both (a) and (b) it is important that candidates using hypotheses like $H_0: \mu_A = \mu_B$ that the subscripts are clearly identified in their method. Hypotheses like $H_0: \mu_F = \mu_M$ gained full credit as it was identifiable by the context.

A significant number of candidates correctly identified that the large sample would lead to a z test in (b) and this proved to be completed more successfully than (a).

Both z and t tests are tests on an average (specifically the mean). Therefore, candidates should be referring to the mean in their conclusions.

It was pleasing to see many fully correct responses to (c). Those who only identified the use of central limit theorem but not how this was relevant to the assumptions achieved 1 mark.

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

A common misconception was that 'Do not reject H_0' meant that there were more goals in the premier league than in the super league.

The audience mark was penalised for use of: random sample, mean, variance, standard deviation, 5% significance or naming a specific test.

For the second mark candidates could either comment on the values seen in the sample or indicate that this result is for 2022 and has therefore changed since 2019.

Q3

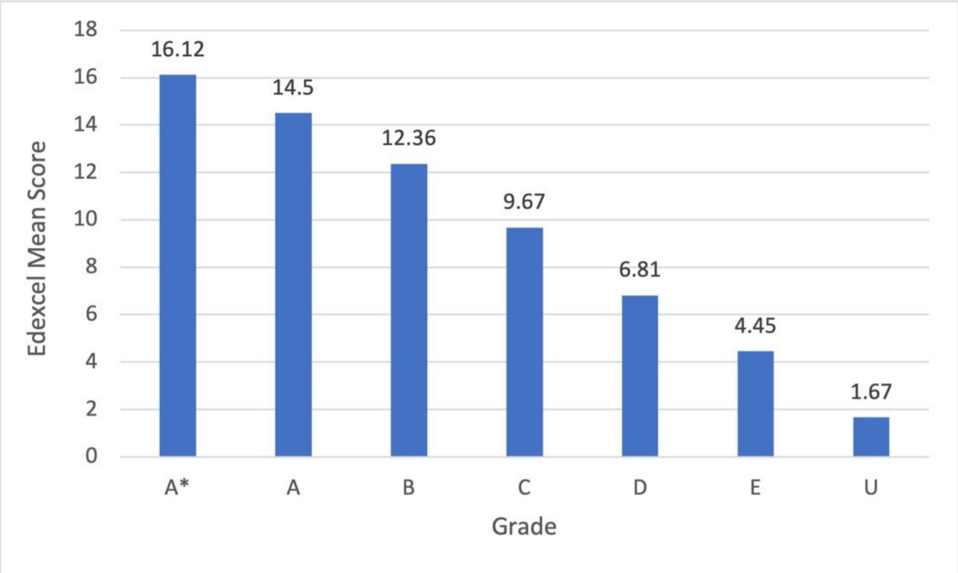
A

B

C

Question 3 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
8.90	17	52	8.90	16.12	14.50	12.36	9.67	6.81	4.45	1.67



Question 3 - Response A

(a) Making any necessary assumptions, perform a hypothesis test to investigate whether, for the 2021/22 season, there were more goals scored per match, on average, in the women's Super League than in the men's Premier League.

(6)

~~$H_0: \mu_s = \mu_p$ $H_1: \mu_s > \mu_p$~~ μ_s = super ~~μ_p = premier~~

~~$\mu = 2.82$ for Premier $\bar{x}_s = 2.92$~~

$H_0: \mu_s = 2.82$ $H_1: \mu_s > 2.82$ SL = 5%.

$\bar{x} = 2.92$ $S = 2.19$ $2.92 + 1.6449 \times \frac{2.19}{\sqrt{12}} =$

$\sim Z(0, 1^2)$ $= 3.96$

0.95

1.6449

$3.96 > 1.6449$ reject H_0

There is significant evidence to suggest that there were more goals scored per match, on average, in the women's Super League than the men's premier league.

(b) Perform a test to investigate whether women in the Super League, on average, are given **fewer** yellow cards than men in the Premier League.

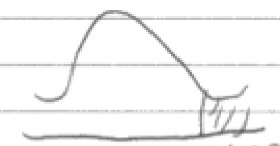
(5)

~~$H_0: \mu_s = \mu_p$ $H_1: \mu_s < \mu_p$~~

$H_0: \mu_s = 3.37$ $H_1: \mu_s < 3.37$

$\bar{x} = 1.93$ $\sqrt{2.76} = \sigma$ 5% SL $\sim Z(0, 1^2)$

$1.93 + (1.6449 \times \frac{\sqrt{2.76}}{\sqrt{12}} =$



1.6449

$= 2.72$ $2.72 > 1.6449$ reject H_0

There is significant evidence to suggest that women in the Super league, on average, are given fewer yellow cards than men in the premier league.

Q3

A

B

C

(c) Explain why there was no need to make any assumptions with regards to the population distribution when carrying out the test in (b)

(2)

because the sample size is less than 30 and s can be used as variance and $\bar{x}$ can be used as the mean.

(d) Write a short article suitable for the general public comparing the highest leagues of men's and women's football in England.

You should ensure that your answer includes

- references to your conclusions in (a) and (b)
- relevant comments with regards to the original sample data.

(4)

The ~~super~~ women's super league and mens premier league both see lots of goals going in but on average the women's league gets more. The mean for the super league is 2.92 but only 2.82. So when you watch either of these leagues, the most likely outcome you will see is 3 goals. The women also commit less yellow card offenses with an average of 1.93 compared to almost double for the men with average yellow cards of 3.86 per match. So if you want to watch a match with the most goals then watch the Super league and if you want to see more yellow cards then there is significant evidence to suggest that you should ~~watch~~ watch the men's premier league.

Q3

A B C

7 /17

Part (a)
B1: Correct hypotheses given.
B1: Both sample mean and sample standard deviation found to 3 significant figures.
M1: Candidate attempted to use the critical value formula correctly.
A0: Not awarded as candidate used a z value rather than the correct t value. Also 2.92 should be 2.82 in this method.
M0: Did not compare their test statistic to the correct critical value.
E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (b)
B1: Correct hypotheses given.
M1: Candidate has recognised that 2.76 is a **variance** and has dealt with this correctly. Although an incorrect value has been used for the sample size the candidate has shown evidence of attempting to find the standard error.
A0: Incorrect critical value found.

M0: Did not compare their test statistic to the correct critical value.
E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (c)
E0: Incorrectly identified the sample as less than 30. Therefore, no comment made about central limit theorem.
E0: No comment made about there being no assumption for the population to be normally distributed.

Part (d)
E1: Although candidate came to the incorrect conclusion in (a) they are awarded follow through for the comment in (d).
E0: Candidate did not address the original sample data and how this may affect the results.
E1: Again although hypothesis test was not completed correctly in (b) they are awarded follow through for their comment in (d).
E0: Use of the word ‘mean’.

Question 3 - Response B

(a) Making any necessary assumptions, perform a hypothesis test to investigate whether, for the 2021/22 season, there were more goals scored per match, on average, in the women's Super League than in the men's Premier League.

(6)

$H_0: \mu = 2.82$

~~$H_0: \mu_A - \mu_b = 0$~~ $H_1: \mu > 2.82$ (1 tail, 5%)

~~$H_1: \mu_A - \mu_b > 0$~~ (1 tail, 5%)

mens

$\bar{x} = 2.82$

Womans

$\bar{x} = 2.91$

$S_x = 2.193$

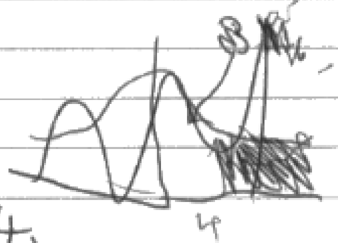
$n_b = 12$

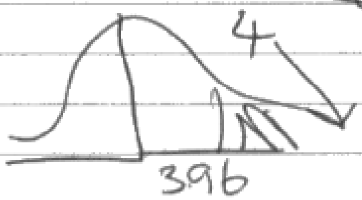
$(+) = (7)$

$(-) = (4)$

$cv = 3.96$

$ts = t =$





∴ Reject H_0 ∴ evidence to suggest more goals where scored in the womans league than in ~~[illegible]~~ mens team.

Q3
?
A
B
C

Question 3 continued

A separate claim, made in the same article, is that the women break the rules in football **less** often than men.

A yellow card is given to a player as a warning about involvement in a rule breaking incident.

During the men's Premier League's 2021/22 season, the mean number of yellow cards per player was found to be 3.37

Paul gathered data for the number of yellow cards given to each of a random sample of 100 female players playing in the women's Super League for the same season. Using his data, Paul found that the mean number of yellow cards per player was 1.93 and the variance was 2.76

(b) Perform a test to investigate whether women in the Super League, on average, are given **fewer** yellow cards than men in the Premier League.

(5)

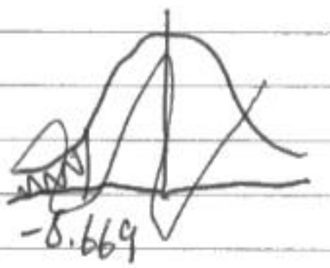
$H_0 = \mu \leq 3.37$

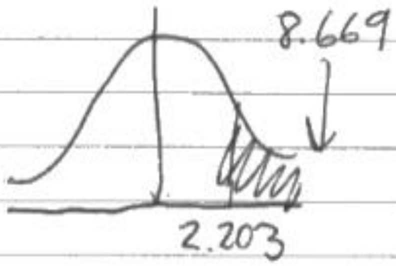
$H_0: \mu_A = \mu_b$ $\quad H_1 = \mu < 3.37$

$H_1: \mu_A > \mu_b$ (1 tail, 5%)

$cv = 2.203$

$ts = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \cancel{3.37} \dfrac{1.93 - 3.37}{1.661/\sqrt{100}} = -8.669$





∴ Reject H_0 there is evidence to suggest women in Super league receive fewer yellow cards than men in Premier League.

(c) Explain why there was no need to make any assumptions with regards to the population distribution when carrying out the test in (b)

(2)

Because the data used is the same as part A and the assumptions have already been made.

Question 3 continued

(d) Write a short article suitable for the general public comparing the highest leagues of men's and women's football in England.

You should ensure that your answer includes

- references to your conclusions in (a) and (b)
- relevant comments with regards to the original sample data.

(4)

For the 2021 season women were more likely to score more goals in the superleague than men in the premier league as the mens average goals scored per game (2.82) was lower than the womens (2.917). However due to not being able to score decimals of a goal it could be argued that both genders would be rounded to 3 and both score the same.

Women are ~~to~~ also less likely to get yellow carded, however the standard deviation is 1.661 so could argue that in some games women could see more yellow cards as we don't the variance of yellow cards in mens games.

10 /17

Part (a)
B1: Correct hypotheses given.
B1: Both sample mean and sample standard deviation found to at least 3 significant figures.
M1: No working shown but implied by correct answer.
A1: Correct critical value found for Alternative 1 method.
M0: Did not compare their test statistic (2.92) to the correct critical value.
E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (b)
B1: Correct hypotheses given.
M1: Candidate attempted to use the standardisation formula to find the test statistic.
A1: Correct test statistic found.
M1: Correctly compared their test statistic to the correct critical value.

E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well. Not awarded as no mention of **average**/**mean**.

Part (c)
E0: Comment was not about sample size.
E0: No comment made about there being no assumption for the population to be normally distributed.

Part (d)
E1: Although candidate came to the incorrect conclusion in (a) they are awarded follow through for the comment in (d).
E0: Candidate did not address the original sample data and how this may affect the results.
E1: Correct comment based on the conclusion from (b).
E0: Use of the phrase 'standard deviation'.

Question 3 - Response C

(a) Making any necessary assumptions, perform a hypothesis test to investigate whether, for the 2021/22 season, there were more goals scored per match, on average, in the women's Super League than in the men's Premier League.

(6)

The population is normally distributed & the ~~8o~~ values are selected ~~for~~ for the sample randomly from the population. The values should also be independent.

$H_0: \mu = 2.82$ $H_1: \mu > 2.82.$

$TS = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

$n = 12$, $\mu = 2.82$

$s = 2.193$

$\bar{x} = 2.917.$

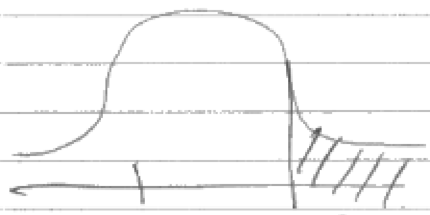
$CV = INVT$

$df = 11$ 0.05

$= 1.796.$

$TS = \dfrac{2.917 - 2.82}{2.193/\sqrt{12}}$

$= 0.1532$



0.1532 1.796

Since $0.1532 < 1.796$ do not reject H_0. There is no significant evidence (at the 5% level) to suggest for the 21/22 season, there were more goals scored per match, on average, in the womens Super League than the mens Premier League.

Q3

A

B

C

Question 3 continued

A separate claim, made in the same article, is that the women break the rules in football **less** often than men.

A yellow card is given to a player as a warning about involvement in a rule breaking incident.

During the men's Premier League's 2021/22 season, the mean number of yellow cards per player was found to be 3.37 $\mu = 3.37$

Paul gathered data for the number of yellow cards given to each of a random sample of 100 female players playing in the women's Super League for the same season. Using his data, Paul found that the mean number of yellow cards per player was 1.93 and the **variance** was 2.76 $n = 100$, 1.93, $\sqrt{2.76} = 1.661$.

(b) Perform a test to investigate whether women in the Super League, on average, are given **fewer** yellow cards than men in the Premier League.   Z test.

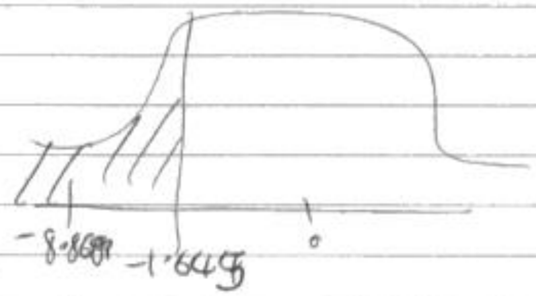
(5)

$H_0: \mu = 3.37$ $H_1: \mu < 3.37$ 5% SL.

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{1.93-3.37}{1.661/\sqrt{100}}$$

$$= -8.669$$

INVN = −1.6448536



Since −8.669 < −1.645 reject H0. There is significant evidence at the 5% SL to suggest women in the Super League, on average, are given fewer yellow cards than men.

(c) Explain why there was no need to make any assumptions with regards to the population distribution when carrying out the test in (b)

(2)

Because $n > 30$, or n was large, so we could assume it was a Z test.

(d) Write a short article suitable for the general public comparing the highest leagues of men's and women's football in England.

You should ensure that your answer includes

- references to your conclusions in (a) and (b)
- relevant comments with regards to the original sample data.

(4)

We can conclude that in the 21/22 Season, mens premier league there were more goals scored per match, on average than in the womens super league. The ~~mean for~~ average for the men per match was higher than the womens super league.
However, it can also be deduced that men are given more yellow cards than women, meaning there are more times when men break the rules and so women are better behaved when it comes to adhering to the rules.
In conclusion, we can see that although there are less goals in the WSL, they receive less yellow cards.

14 /17

Part (a)
B1: Correct hypotheses given.
B1: Both sample mean and sample standard deviation found to 3 significant figures.
M1: Candidate attempted to use the standardisation formula to find the test statistic.
A1: Correct test statistic found.
M1: Correctly compared their test statistic to the correct critical value.
E1: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (b)
B1: Correct hypotheses given.
M1: Candidate attempted to use the standardisation formula to find the test statistic.
A1: Correct test statistic found.
M1: Correctly compared their test statistic to the correct critical value.
E1: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value as well.

Part (c)
E1: Correctly identified the sample as more than 30. Central limit theorem did not need to be recognised for this mark.
E0: No comment made about there being no assumption for the population to be normally distributed.

Part (d)
E0: A common misconception across a lot of candidates was that 'Do not Reject H_0' meant that the men's average was greater.
E0: Candidate did not address the original sample data and how this may affect the results.
E1: Correct comment based on the conclusion from (b).
E1: Candidate ensured non-statistical language was used throughout their written response.

Q3

?

A

B

C

Question 4

Question 4 - Question

4 Alder is a first year psychology student at a university. He decides to use a non-parametric test on some data he has gathered.

He reads some statistics books and decides that either a sign test or the Wilcoxon signed-rank test could be appropriate.

Compare these two tests for Alder.

Your comparison should include comments on

- their similarities,
- their differences,
- the assumptions necessary, if any, for each test, for that test to be valid.

(5)

(Total for Question 4 is 5 marks)

Question 4 - Mark Scheme

Question	Scheme	Marks	AO	Notes
4	**Similarities** Both can be carried out on a single sample or a paired sample Both are tests of the null hypothesis that the population median takes some value (or two population medians are equal in paired case) Both require a random sample Ignore differences of 0			Condone both test for a median (or difference in medians)

Question 4 - Mark Scheme (Cont.)

	Differences			
	Sign test can be used for non-numerical or summarised data, Wilcoxon signed-rank cannot			oe
	Wilcoxon signed-rank test takes into account size of differences, rather than just signs			Condone more powerful
	The critical value for the Wilcoxon signed-rank test is given in a table, but the critical value for a sign test is found using the binomial distribution			
	Wilcoxon signed-rank can be used to test for the mean			
		B1, B1, B1	1.3 1.3 1.3	Any 3 of the above reasons
	Assumptions			
	Sign test – no assumptions regarding distribution needed	B1	1.3	
	Wilcoxon Signed-Rank test – Need symmetry of distribution for single sample (or symmetry of differences for paired)	B1	1.3	
	Total	**5**		

Q4

?

A
B
C

Question 4 - Examiner Comments

A new style of question that saw those candidates with a more refined exam technique scoring well. In questions where specific bullet points are given, in this case 3, the responses given must reference **all 3 bullet points**.

Those who scored well broke their answer up into: Similarities, Differences and Assumptions, this meant that comments were **clearly identified** as one of these 3 categories.

Some candidates wrote in long prose with a general description of the test. These did not gain credit as it is the candidate who is required to clearly identify similarities and differences, not the examiner.

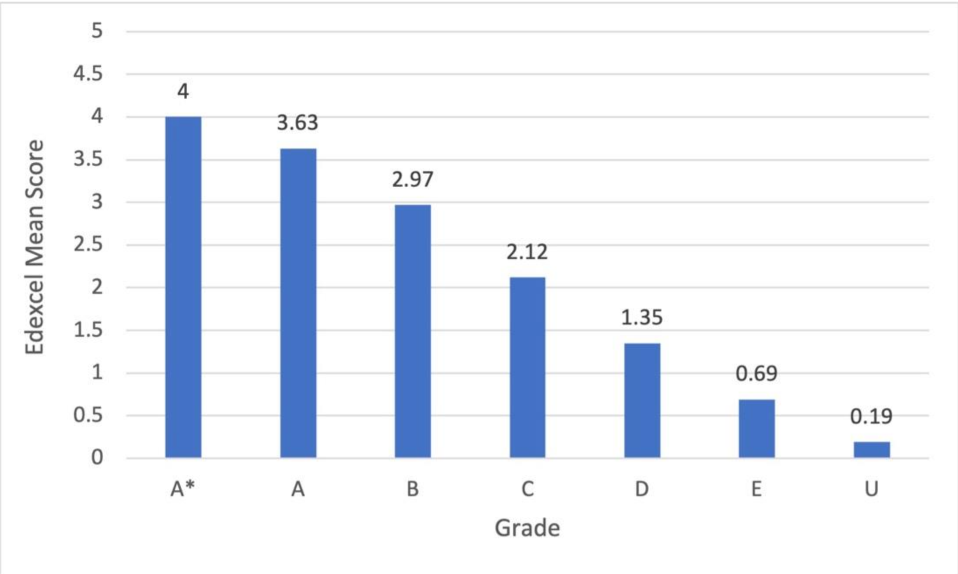
Similarities and differences that could have been true for other hypothesis tests as well were not given credit eg 'if the test statistic is in the critical region the conclusion is to Reject H_0'

It was important for candidates to **specifically link** the hypothesis test to its assumption eg the sign test has no distributional assumptions and the Wilcoxon signed rank test with symmetry for the final marks. Some candidates did not clearly state that the sign tests had no distributional assumptions.

A common incorrect response was to say 'the Wilcoxon signed rank tests is more accurate than the sign test' this did not gain credit as all hypothesis tests are accurate went conducted correctly. The term the specification would require is **powerful** (this was previously mentioned in the exemplar materials from 2206 9ST0 03)

Question 4 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
1.97	5	39	1.97	4.00	3.63	2.97	2.12	1.35	0.69	0.19



Question 4 - Response A

A Sign test Measures the Median.

One Similarity is that the data does not have to be normally distributed for it to work

• For a WSRT, we assign ranks to the data, and use Σ Positive Rank values or Σ Negative rank values as a test Statistic

• For a Sign test, which is more Simple, we just compare the data to the given mean, and count the amount of positive or negative values, the lower of the two should be used in the TS using Binomial model

• Wilcoxon Sign rank test and a Sign test are Similar, they both measure whether the data is less than, equal to or greater than the mean

• We must assume that the data is independent

we must assume that the data is

0 /5

The candidate received no credit for any of the comments. This is explained below:

- 1st comment – no credit, as it is not a similarity/difference/assumption.
- 2nd comment – no credit was given for comments about being non-parametric or not requiring a normal.
- 3rd comment – no credit, a description of the Wilcoxon Signed-Rank test.
- 4th comment – no credit, a description of the Sign test.
- 5th comment – no credit, candidate used mean instead of median.
- 6th comment – no credit.

Q4

A

B

C

Question 4 - Response B

(5)

Similarities
- Both tests can be used ~~for~~ either paired or non paired data.
- Both test for discrete data sets.

Differences
- Sign tests can be used for opinions or non numerical data whereas signed rank needs numerical data.

Assumptions
- For the Wilcoxon signed rank test, the data needs to be symmetrical and for paired tests, the differences need to be symmetrical.

3 /5

E1: Stated that the tests can be used for both single and paired data.
E1: Stated that sign test can be used for non-numerical data.
E0: No third similarity/difference given.
E0: Did not state any points about the sign test's assumptions.
E1: Clearly stated that the Wilcoxon Signed-Rank test requires a symmetric distribution, or that the differences are symmetric.

Extra Comments:
Very good exam technique as the candidate has made it clear whether comments are similarities, differences, or assumptions.

Question 4 - Response C

Both tests look at the difference ~~bet~~ either between two data sets, or a median value and another data set. Also, they both compare medians. ~~They look~~ Furthermore, both tests use the difference by looking at the signs of the difference.

The main difference between Wilcoxon signed-rank and the sign test, is that Wilcoxon ~~[illegible]~~ takes the data, and then looks at the sign. Instead, the sign test only looks at the sign of each difference and counts the amount of values for each sign. Because of this the Wilcoxon signed-rank is a more powerful test.

You should always try to do the Wilcoxon signed-rank, but for the Wilcoxon signed-rank test to be valid, the differences of the distribution has to be symmetrical. On the other hand the sign test doesn't have any assumptions. So it is useful if the assumptions for the Wilcoxon signed-rank aren't met.

4 /5

E1: Stated that both tests can test for the median.
E1: Identified that the Wilcoxon Signed-Rank test takes into account the magnitude of the differences. Correctly used the word powerful.
E0: No third similarity/difference given.
E1: Clearly stated that the sign test requires no distributional assumptions.
E1: Clearly stated that the Wilcoxon Signed-Rank test requires a symmetric distribution.

Q4

?

A
B
C

Question 5

Question 5 - Question

5 In March 2011, there was an accident at the Fukushima Nuclear Power Plant in Japan that caused a massive release of radioactive materials into the environment.

An investigation was carried out to see whether the release of radioactive materials affected the *Zizeeria maha*, a common butterfly in Japan.

The **first** generation of butterflies from the Fukushima area was observed in May 2011.

The offspring of the first generation, called the **second** generation, were then observed.

Also, the offspring of the second generation, called the **third** generation, were observed.

For each of these generations, the wing size, cm, for a random sample of male butterflies was measured.

The recorded measurements are shown in **Figure 6**.

	First Generation	Second Generation	Third Generation
	1.43	1.30	1.27
	1.44	1.33	1.31
	1.44	1.38	1.38
	1.43	1.34	1.31
	1.42	1.37	1.35
	1.40	1.33	1.32
	1.45	1.40	1.41
Total	10.01	9.45	9.35

[Data Source https://www.nature.com/articles/srep00570]

Figure 6

The data produced the following summary statistic.

$$\sum x_i^2 = 39.5831$$

Question 5 – Question (Cont.)

(a) Carry out a hypothesis test to investigate whether there is any difference in the average wing size for male butterflies between the three generations.

NOTE: When using Table 7 in the Statistical formulae and tables booklet to obtain the critical value, if the exact critical value is not available you should use the closest critical values either side of the degrees of freedom required.

(10)

It was later discovered in this investigation that, for each generation, a male butterfly was observed in each of seven different locations surrounding the Fukushima area.

(b) Explain how your analysis in (a) could be adapted using this further information and state the name of the new experimental design.

(2)

(c) Comment on how the experimental error will be affected by using the design stated in (b).

(1)

For each of the seven locations, the distance, in km, from the Fukushima Nuclear Power Plant was recorded alongside the average abnormality rate, %, for the male butterflies in the **first** generation.

One of the researchers, Joji, believed that the abnormality rate would be lower for locations further away from the Fukushima Nuclear Power Plant.

The data for the seven locations is shown in **Figure 7**.

Location	A	B	C	D	E	F	G
Distance from Power Plant	20	30	57	58	77	81	128
Abnormality Rate	6.8	6.6	4.1	3.8	1.5	3.8	0.7

Figure 7

(d) Calculate Pearson's product moment correlation coefficient between the distance from the Power Plant and the abnormality rate.

(1)

(e) Making any necessary assumptions, conduct a hypothesis test to investigate Joji's belief.

(4)

(Total for Question 5 is 18 marks)

Q5

A

B

C

Question 5 - Mark Scheme

Question	Scheme	Marks	AO	Notes
5(a)	$H_0: \mu_1 = \mu_2 = \mu_3$ H_1: At least two of the means differ	B1	1.3	oe Or $\mu_i = \mu$ for $i = 1$ to 3
	T = 28.81 $SS_T = 39.5831 - \frac{28.81^2}{21}$ $= 0.05852$	M1	1.3	SS_T method PI
	$SS_B = \frac{10.01^2}{7} + \frac{9.45^2}{7} + \frac{9.35^2}{7} - \frac{28.81^2}{21}$ $= 0.03615$	M1	1.3	SS_B method PI
		M1dep	1.3	SS_E method No negative SS values. Dep one previous M1
	(see table below)	B1	1.3	PI df 2 and 18
		M1	1.3	PI MS divide SS by df ft on df values
	$F = \frac{0.018076}{0.001243} = 14.5$	M1	1.3	PI F method ft on df values **or** p–value = 0.00017
	Consideration of cvs cv $F_{15}^2 = 3.682$ and $F_{20}^2 = 3.493$	B1	1.3	Testing both/either cvs or $F_{18}^2 = 3.55$

	SS	df	MS
Between generations	0.03615	2	0.018076
Error	0.02237	18	0.001243
Total	0.05852	20	

Question	Scheme	Mark	AO	Notes
	"14.5" > "3.682" and/or "3.493" Reject H_0	M1	2.1b	Comparison of their ts and their cv or "14.5" > 3.55 or 0.000174 < 0.05 Correct conclusion for their ts and their cv
	There is significant evidence to suggest that **mean** wing size for male butterflies for at least 2 of the generations differ	E1dep	2.1a	Correct conclusion in context Dep on test fully correct
5(b)	Use a 2 factor ANOVA using location as a blocking factor	E1	3.1a	
	Randomised Block Design	E1	2.1b	
5(c)	The experimental error will **reduce**	E1	3.1a	oe
5(d)	r = -0.924	B1	1.2	
5(e)	H_0: No association H_1: Negative association	B1	1.3	oe Both Accept H_0: $\rho = 0$ H_1: $\rho < 0$
	cv = -0.6694 1-tail	B1	1.3	Ignore sign for cv
	"-0.924" < "-0.6694" Reject H_0	M1	2.1b	Comparison of their ts and their cv, consistent sign
	There is significant evidence to suggest that there is a negative association between the distance, in km, from the Fukushima Nuclear Power Plant and the average abnormality rate, %, for the male butterflies	A1	2.1a	Conclusion in context dep ts, cv correct oe
	Total	**18**		

Question 5 - Examiner Comments

The largest question on the exam paper. However, it was clear candidates had been well prepared for ANOVA (this was also true in 2206). Many candidates managed to achieve full marks in (a) or only the final explanation mark. It is important to note that ANOVA is a test on the population mean and therefore it was **expected that the mean was referred to** in both the hypotheses and conclusion.

Where candidates had stated a negative for any of their SS values, they were then unable to get any further method marks.

Candidates should also be encouraged to use the formulas that are given in the formula booklet, there were cases where the student had misremembered the formulas and therefore lost marks.

Students who do this on their calculator should include the total row in their table.

In this series we were generous by allowing candidates to use either test statistic. If candidates are using the tables they must state and compare against **both critical values** (or an interpolated value/calculated value).

Several candidates read ahead and then incorrectly took the data to be a 2-factor table, in these cases they could still gain credit but it is important to note that in an exam candidates may need to read back (as in Q05(e)) in an exam question they will **never be required to read forward** to answer a question.

Part (b) was answered well with many candidates identifying that location could be used as a blocking factor and correctly identifying the design as a Randomised Block Design. In some cases, candidates named another variable as a possible blocking factor, this was not given credit as the question clearly stated 'using this further information' so, as in Q02(d), location was required to be mentioned specifically.

An easy mark was obtained by the vast majority of candidates in (c). It was not required they explain why experimental error had been reduced, though it was pleasing to see many correct attempts at reasons.

Parts (d) and (e) were completed by the vast majority of candidates correctly which was good to see so late on in the paper. Full credit was only given to those testing for a negative correlation as this is the relationship that Joji had described in (d)

Q5

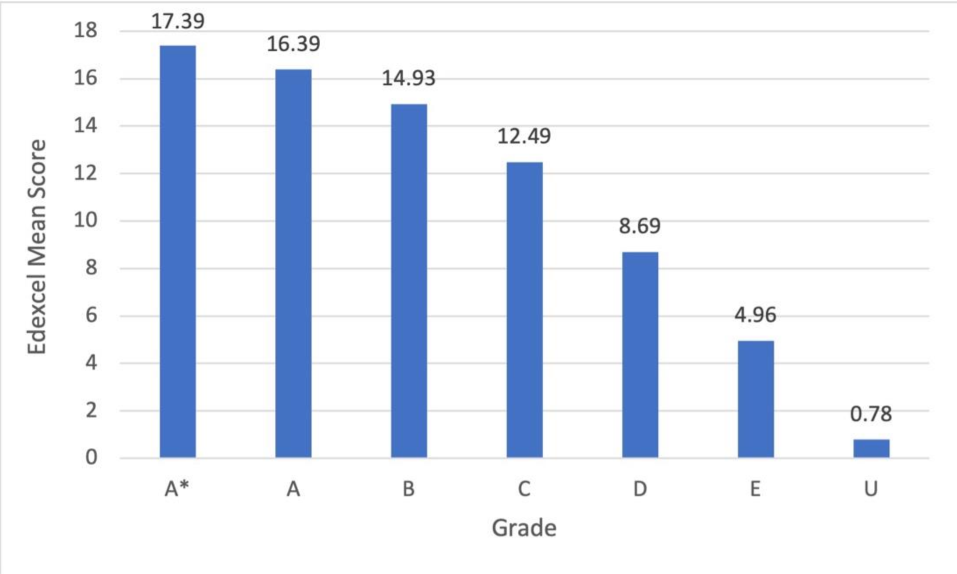
A

B

C

Question 5 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
10.78	18	60	10.78	17.39	16.39	14.93	12.49	8.69	4.96	0.78

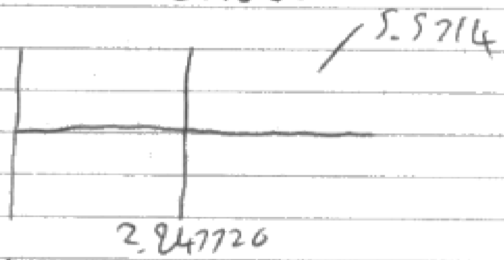


Question 5 - Response A

$H_0: \mu_E = \mu_S = \mu_T$
H_1: At least two means differ

	SS	d.f	MS	F
Factor A	13 ÷	14 =	0.9285	$\frac{0.9285}{0.1666} = 5.5714$
Standard Error	1 ÷	6 =	0.1666	

$F^{6}_{14} = 2.847726$



Reject H_0, evidence of a difference in wing size between generations

It was later discovered in this investigation that, for each generation, a male butterfly was observed in each of seven different locations surrounding the Fukushima area.

(b) Explain how your analysis in (a) could be adapted using this further information and state the name of the new experimental design.

(2)

Block by location

Randomised block design

(c) Comment on how the experimental error will be affected by using the design stated in (b).

(1)

Error will be reduced as it eliminates differences caused by location

(d) Calculate Pearson's product moment correlation coefficient between the distance from the Power Plant and the abnormality rate.

(1)

$r = -0.924$

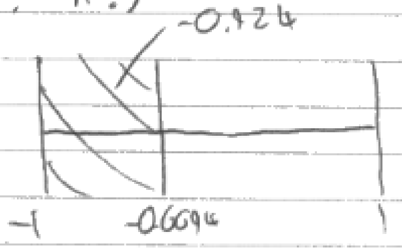
(e) Making any necessary assumptions, conduct a hypothesis test to investigate Joji's belief.

(4)

$H_0: \rho = 0$

$H_1: \rho < 0$ 1 tail 5%, n=7

CV=-0.6694



Reject H_0, evidence to suggest lower abnormality rate the further from the nuclear power plant you go

12 /18

Part (a)
B1: Correct hypotheses, as seen in the mark scheme.
M0: Value for SS_T incorrectly calculated. Since there is no working no credit can be given.
M0: Value for SS_B not calculated. Since there is no working no credit can be given.
M0: Value for SS_E cannot be credited as we cannot see where the value has come from.
B0: Degrees of freedom stated are incorrect.
M1: Mean squares values have been found using their values for SS and degrees of freedom.
M1: Test statistic found by dividing their MS_B by MS_E
B0: Incorrect critical value was found, due to incorrect degrees of freedom.
M1: Correct comparison between their test statistic and their critical value.
E0: Conclusion has context based on the information from the question. However, it does not include the word **mean**. This final mark was dependent on correct test statistic and critical value.

Part (b)
E1: Identified that location could be used as a blocking factor.
E1: Correctly stated the new experimental design in full.

Part (c)
E1: Correct response.

Part (d)
B1: Correct value for PMCC calculated.

Part (e)
B1: Hypotheses are correct using ρ
B1: Correct critical value found from the formula booklet.
M1: Correct comparison of their test statistic with their critical value (signs must be consistent at this point)
A1: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value.

Question 5 - Response B

$Ho: u_i = u$ for all $i = 1,2,3$
$H_1: u_i \neq u$ for some $i = 1,2,3$

$SS_T = 39.5831 - \frac{28.81^2}{21} = 0.058538$

$SSB = \left(\frac{10.01^2}{7} + \frac{9.45^2}{7} + \frac{9.35^2}{7}\right) - \frac{28.81^2}{21} = 0.0361524$

$SSw = 0.0223714$

Source of var:

W: $0.0223714 \div 18 = 0.001243$ $F = 14.56$ $Cv = 3.55$
B: $0.0361524 \div 2 = 0.0181$

$14.56 > 3.55$ so Reject Ho. There is significant evidence to suggest that there is a difference in the average wing size for male butterflies between three generations

It was later discovered in this investigation that, for each generation, a male butterfly was observed in each of seven different locations surrounding the Fukushima area.

(b) Explain how your analysis in (a) could be adapted using this further information and state the name of the new experimental design.

(2)

Two factor Annova

(c) Comment on how the experimental error will be affected by using the design stated in (b).

(1)

It will be reduced

(d) Calculate Pearson's product moment correlation coefficient between the distance from the Power Plant and the abnormality rate.

(1)

$P = -0.935$

(e) Making any necessary assumptions, conduct a hypothesis test to investigate Joji's belief.

(4)

~~H0: ρ = −0.935~~
~~H1: ρ < −0.935~~

H0: ρ = 0   ρ = −0.935   C.V. = −0.6694
H1: ρ < 0

−0.935 < −0.6694   So Reject H0: There is significant evidence to suggest that the abnormality rate would be lower for breeding further away from the Fukushima Nuclear Power plant.

13 /18

Part (a)
B1: Correct hypotheses, a different version from those seen in scheme of work but is acceptable.
M1: Value for SS_T correctly calculated and working shown.
M1: Value for SS_B correctly calculated and working shown.
M1: Value for SS_W has been calculated using the two previous values correctly.
B1: Degrees of freedom stated correctly for each row in the ANOVA table.
M1: Mean squares values have been found using their values for SS and degrees of freedom.
M1: Test statistic found by dividing their MS_B by MS_E
B1: A correct critical value was found, in this case from the calculator.
M1: Correct comparison between their test statistic and their critical value.
E0: Conclusion has context based on the information from the question. However, the word **mean** was not mentioned. This final mark was dependent on correct test statistic and critical value.

Part (b)
E0: Identified that two factor ANOVA could be used but not that location was a blocking factor.
E0: Did not state the experimental design.

Part (c)
E1: Correct response.

Part (d)
B0: Incorrect value for PMCC calculated.

Part (e)
B1: Hypotheses are correct using ρ
B1: Correct critical value found from the formula booklet.
M1: Correct comparison of their test statistic with their critical value (signs must be consistent at this point).
A0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value.

Question 5 - Response C

$H_0: \mu_F = \mu_S = \mu_T$
H_1: at least two of the means differ

	SS	df	MS	F
Between Generation	0.0361	2	0.0180	14.544
Error	0.0223	18	0.0012	
total	0.0584	20		

ndf = 2 ddf = 18
CV = 3.555 < F

$\therefore$ reject H_0, sufficient evidence at 5% level to conclude that there is a difference between at least two generations in wing size.

It was later discovered in this investigation that, for each generation, a male butterfly was observed in each of seven different locations surrounding the Fukushima area.

(b) Explain how your analysis in (a) could be adapted using this further information and state the name of the new experimental design.

(2)

- ~~Add new blocking factor for Sex of butterfly~~
- Add new blocking factor for location or distance from location to exploded reactor for each butterfly.
- Perform two-factor ANOVA to check for difference in means in both variables.

(c) Comment on how the experimental error will be affected by using the design stated in (b).

(1)

- ~~[illegible]~~ Experimental error will be increased as more variables are added.

Q5 ? A B C

(d) Calculate Pearson's product moment correlation coefficient between the distance from the Power Plant and the abnormality rate.

(1)

$r = -0.9242$

(e) Making any necessary assumptions, conduct a hypothesis test to investigate Joji's belief.

(4)

H_0: no correlation between distance and abnormality rate, $r = 0$

H_1: $r < 0$

$CV = -0.6694 > -0.9242$

∴ reject H_0, sufficient evidence at 5% level to conclude that the abnormality rate is lower ~~the~~ further away from the plant.

15 /18

Part (a)

B1: Correct hypotheses, as seen in the mark scheme.

M1: Value for SS_T correctly calculated. If this value had been miscalculated mark would not have been awarded as there is no working. **Candidates are reminded that working out should be shown**.

M1: Value for SS_B correctly calculated, again would be improved by including working.

M1: Value for SS_E has been calculated using the two previous values correctly.

B1: Degrees of freedom stated correctly for each row in the ANOVA table.

M1: Mean squares values have been found using their values for SS and degrees of freedom.

M1: Test statistic found by dividing their MS_B by MS_E

B1: A correct critical value was found, in this case from the calculator.

M1: Correct comparison between their test statistic and their critical value.

E1: Conclusion has context based on the information from the question as well as using the work **mean**. This final mark was dependent on correct test statistic and critical value.

Part (b)

E1: Identified that location could be used as a blocking factor.

E0: Did not state the experimental design, the candidate stated an inference instead which was a common error on this question.

Part (c)

E0: Incorrect response.

Q5

?

A

B

C

Part (d)
B1: Correct value for PMCC calculated.

Part (e)
B0: Null hypothesis could be considered correct (ignoring the r=0). However, the alternative hypothesis should either be in words or using ρ
B1: Correct critical value found from the formula booklet.
M1: Correct comparison of their test statistic with their critical value (signs must be consistent at this point).
A1: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value.

Extra comments:
Those candidates using a calculator to fill in their ANOVA table should be encouraged to show working, a single mistype in this case would have cost them several marks.

Q5

?

A
B
C

Question 6

Question 6 - Question

6 A study investigated the effects of smartphone displays with blue light on the sleepiness of adults playing smartphone games.

[Source: Effects of smartphone use with and without blue light at night in healthy adults: A randomized, double-blind, cross-over, placebo controlled comparison - ScienceDirect]

Two groups of adults took part in the trial.

- One group played games **with** blue light.
- One group played games **without** blue light.

Sleepiness levels were measured for each adult after playing the games.

A *t*-test was carried out to investigate whether the adults playing games **with** blue light had a **decreased** average sleepiness level compared to those playing games **without** blue light.

The test gave a *p*-value of 0.04 and a Cohen's *d* value of 0.49

Use both the *p*-value and the value of Cohen's *d* for this test to make conclusions about the difference between average sleepiness levels for adults playing games with blue light and those playing without blue light.

You may assume that sleepiness levels for all adults are normally distributed.

(4)

(Total for Question 6 is 4 marks)

Question 6 - Mark Scheme

Question	Scheme	Marks	AO	Notes
6	***p*-value** ($0.04<0.05$) evidence of significant difference (so reject H_0).	B1	1.3	Knowledge of *p*-value
	….there is evidence to suggest that adults playing games with blue light have decreased average sleepiness levels compared to those playing games without blue light.	E1	2.1a	Explained fully in context

Question 6 - Mark Scheme (Cont.)

Question	Scheme	Marks	AO	Notes
	Cohen's *d* Medium effect …..	B1	1.3	Knowledge of Cohen's *d*
	….so suggests that there is a medium size magnitude in the difference between the average sleepiness levels for the two groups.	E1	2.1a	Explained fully in context
	Total	**4**		

Question 6 - Examiner Comments

At this point in the paper some candidates did not make an attempt at this question. Although Cohen's *d* is one of the later topics in the A-level content, there are some easily accessible marks for learnt terminology and methods.

Candidates who were the most successful on this question **broke their answers up** by dealing with the interpretation of the hypothesis test first and then interpreting the value of Cohen's *d* separately.

When interpreting the *p*-value, many students incorrectly interpreted this as the size of the difference between the averages. Rather than identifying that a small *p*-value meant that such an extreme observed outcome would be unlikely under the null hypothesis (quoted from 2206 examiner feedback Q03). For the first mark students were only required to show (or imply) Reject H_0.

When interpreting the value of Cohen's *d*, many students used incorrect terminology. The specification states the following:

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

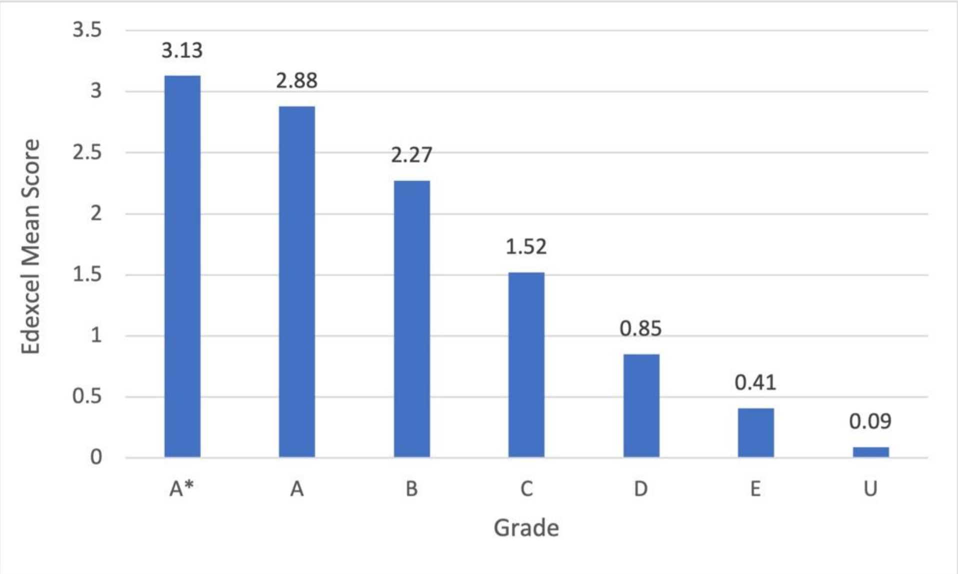
However, marks were allowed for small or medium as 0.49 was close to the boundaries.

Candidates did not gain full credit if:

- they omitted the word average/mean in the comments (penalised once)
- they did not write two separate comments (one for the interpretation of the *p*-value and one for Cohen's *d*)

Question 6 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
1.43	4	36%	1.43	3.13	2.88	2.27	1.52	0.85	0.41	0.09



Question 6 - Response A

$h_0 : \mu_b = \mu_{b'}$

$h_1 \; \mu_b < \mu_{b'}$

$p = 0.04$ do not accept h_0

$CV = 0.05$ $CV > P$ ∴ ~~Rej~~

There is sufficient evidence to suggest that the adults playing games with blue light had a decreased ~~slep~~ Sleepiness level compared to those playing games without blue light.

dVt Cohens d → $0.4 < d < 0.6$

∴ medium affect size

2 /4

B1: Correctly compared the *p*-value to the significance level and concluded to Reject H_0
E0: Correctly interpreted Reject H_0 within the context of the question but did not use the word average/mean.
B1: Correctly described the effect size as medium/small.
E0: Did not attempt to interpret the effect size.

Q6

?

A
B
C

Question 6 - Response B

0.04 < 0.05 so there is sufficient evidence to suggest that the adults playing games with blue light had a significantly decreased average sleepiness level compared to those playing games without blue light.
However, the cohen's d value of 0.49 suggests only a moderate effect size. This means that the effect of playing games with blue light on average sleepiness level compared with playing games without blue light is moderate.

3 /4

B1: Correctly compared the *p*-value to the significance level and concluded to Reject H_0
E1: Correctly interpreted Reject H_0 within the context of the question using the word average/mean.
B0: Incorrectly described the effect size as moderate.
E1: Correctly interpreted the effect size within the context of the question using the word average/mean.

Q6
?
A
B
C

Question 6 - Response C

(4) Q06E2

0.04 < 0.05 so is sig. Evidence of difference between average sleepiness levels for adults playing games with or without blue light.

0.49 shows a medium effect size, so there is a relative difference between average sleepiness levels for adults playing games with or without blue light.

$H_0: p = 0$ The t-test shows that playing
$H_1: p < 0$ games with blue light does ~~increase~~ decrease average sleepiness levels, due to the p-value being small.

4 /4

B1: Correctly compared the *p*-value to the significance level and concluded to Reject H_0
E1: Correctly interpreted Reject H_0 within the context of the question using the word average/mean.
B1: Correctly described the effect size as medium/small.
E1: Correctly interpreted the effect size within the context of the question using the word average/mean.

Question 7

Question 7 - Question

7 Justin investigated how advertising food and drinks to children affects their food consumption.

A group of 20 children was shown an episode of a particular TV show on two consecutive Wednesdays.

One Wednesday the children were shown an episode with adverts for **toys** during the break.

The other Wednesday the children were shown an episode with adverts for **food** during the break.

After they had seen each episode, the children were given a tray of food. The total amount of food each child ate was measured.

(a) Explain which data Justin should use if he wanted to carry out a paired t-test as part of his investigation.

(2)

Kareema says a paired t-test is **only** valid if both samples are selected from normally distributed populations.

(b) Explain why it is valid to use a paired t-test when both samples are from normally distributed populations.

(1)

(c) Explain whether or not it is necessary for **both** samples to be selected from normally distributed populations for a paired t-test to be valid.

(2)

Justin wanted to make sure that the only factor causing the difference in food eaten was the type of advert the children were shown.

He was concerned that the hunger level of the children may have been different on the two Wednesdays.

Therefore, he also asked the 20 children to rate their hunger on a scale of 1, not hungry, to 5, very hungry, before they watched the episode.

The mean difference of the hunger rates for the children between the two Wednesdays was 0.03 and the standard deviation of these differences was 0.89

[Source: https://livrepository.liverpool.ac.uk/3027509/1/201153268_Oct2018.pdf.pdf]

Justin decided not to use hunger as a factor in his experiment due to this data.

Question 7 – Question (Cont.)

(d) Carry out a hypothesis test to explain Justin's decision.

(5)

It is also possible to conduct a paired test with a sign test or the Wilcoxon signed-rank test. These tests are non-parametric.

(e) Explain what is meant by a non-parametric test.

(1)

(Total for Question 7 is 11 marks)

Question 7 - Mark Scheme

Question	Scheme	Marks	AO	Notes
7(a)	The differences between the total amounts of food consumed by **each child** on each of the two Wednesdays	E1 E1	1.3 1.3	Clear indication that differences must be obtained Explained fully in context indication **both** samples are required
7(b)	The difference of two normal distributions is a normal distribution (with the means subtracted and variances added)	E1	3.1a	
7(c)	It is not necessary that **both samples** are selected from normally distributed populations provided that the population of **differences** are normally distributed	E1 E1	3.1a 3.1a	Must have an attempt at a correct justification
7(d)	$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$ $ts = \frac{0.03-0}{\frac{0.89}{\sqrt{20}}} = 0.151$ (3sf) $cv = \pm 2.093$	B1 B1 B1	1.3 1.3 1.3	Both Hypotheses **or** p-value = 0.882

Q7 A B C

	" -2.093" < "0.151" < "2.093" Do not reject H_0.	M1	2.1b	Comparison their ts with their cv **or** 0.882 > 0.025
	No significant evidence to suggest any differences in average hunger levels so it was reasonable for Justin to ignore it	E1dep	2.1a	Conclusion in context dep ts, cv correct
Alternative				
7(d)	H_0: $\mu_d = 0$ H_1: $\mu_d \neq 0$	(B1)		Both Hypotheses
	cv = $0 \pm 2.093 \times \frac{0.89}{\sqrt{20}}$	(B1)		PI Method for cv
	$= \pm 0.4165$	(B1)		PI **or** p-value = 0.882
	" -0.417" < "0.03" < "0.417" Do not reject H_0.	(M1)		Comparison their ts with their cv **or** 0.882 > 0.025
	No significant evidence to suggest any differences in average hunger levels so it was reasonable for Justin to ignore it	(E1dep)		Conclusion in context dep ts, cv correct
7(e)	A test which does not rely on the sample being selected from a specific probability distribution.	E1	1.3	
	Total	**11**		

Question 7 - Examiner Comments

This was the most challenging question of the paper; it was encouraging to see many candidates making a good attempt at all parts.

In (a) the first mark was awarded for evidence that the student understood **differences** would need to be found, **context** was required for the second mark as well as identifying that both samples should be used.

Parts (b) and (c) were discriminating for the high achieving students. Similarly, to Q04, those candidates who had invested time into **learning the assumptions** for the hypothesis tests were rewarded.

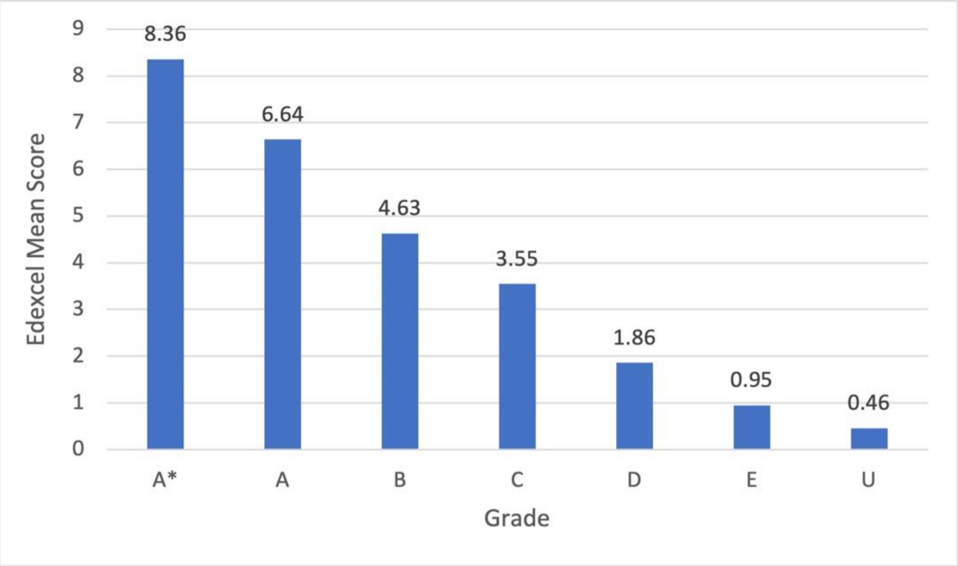
The paired t test was attempted well by candidates, though students should be encouraged to keep going to the end of the paper and **attempt all questions** set as there were some accessible marks.

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

In the final mark of the exam **incorrect or vague terminology was penalised** eg the use of sample/it/data instead of population distribution. It is important that candidates understand that assumptions are about the distribution of the population (or differences for paired tests) and this terminology is expected in their responses as well.

Question 7 - Performance

Mean score	Max score	Mean %	Edexcel averages: mean scored by candidates achieving grade:							
			ALL	A*	A	B	C	D	E	U
3.17	11	29%	3.17	8.36	6.64	4.63	3.55	1.86	0.95	0.46



Question 7 - Response A

(a) Explain which data Justin should use if he wanted to carry out a paired t-test as part of his investigation.

(2)

Justin Should analyse ~~both~~ the amount of food consumed by the Children on both wednesdays where the adverts for toys ~~and~~ and adverts for food were Shown.

Kareema says a paired t-test is **only** valid if both samples are selected from normally distributed populations.

(b) Explain why it is valid to use a paired t-test when both samples are from normally distributed populations.

(1)

~~t-tests do not~~ t-tests are unaffected by the Population distribution the Samples are Selected from.

(c) Explain whether or not it is necessary for **both** samples to be selected from normally distributed populations for a paired t-test to be valid.

(2)

t-tests do not require a normal distribution to be performed, instead choosing to replace Z with a t-value.

Q7

?

A

B

C

(d) Carry out a hypothesis test to explain Justin's decision.

(5)

σ is unknown and n is small $(20 < 30)$ so t-dist must be used.

$H_0: \mu_x = 0$ 5% Significance x = hunger rates for the 20 children between
$H_1: \mu_x \neq 0$ two-tailed the two wednesdays $X \sim N\left(0.03, \frac{0.89^2}{20}\right)$

Assuming H_0 is true, $\bar{X} \sim N\left(~~0.03~~\ 0, \frac{0.89^2}{20}\right)$

$C.V = 0 \pm t\left(\frac{0.89}{\sqrt{20}}\right) = 0 \pm 2.086\left(\frac{0.89}{\sqrt{20}}\right) = (-0.4151, 0.4151)$
$t = 2.086$
$v = 20 - 1 = 19$

Since $0.03 > -0.4151$ and $0.03 < 0.4151$, there is insufficient evidence to reject H_0 at a 5% significance level, suggesting that on average, the hunger rate between the two wednesdays was 0, meaning that for Justin's investigation, hunger was not a significant factor that affected the amount of food each child ate.

(e) Explain what is meant by a non-parametric test.

(1)

A hypothesis test that does not use/assume any distribution when it is performed.

5 /11

Part (a)
E0: Did not identify the need for differences to be found for the data.
E1: Correctly identified that the data from both weeks would be required and used context throughout answer.

Part (b)
E0: Did not identify that differences would be normally distributed from a linear combination.

Part (c)
E0: Did not correctly identify that the test would still be valid.
E0: Did not understand that the assumption is on the differences for paired *t* test and not the original population distribution.

Part (d)
B1: Hypotheses are correct as although they did not use μ_d. They used μ_x and defined its meaning.
B1: Candidate attempted to use the critical value formula correctly.
B0: Not awarded as candidate used the incorrect *t* value.

M1: Correctly interpreted their test statistic and critical value.
E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value.

Part (e)
E1: Although a little vague it felt that the candidate was getting to what was needed for this question.

Question 7 - Response B

(a) Explain which data Justin should use if he wanted to carry out a paired t-test as part of his investigation.

(2)

· he should use both sets of data ~~to~~ to test for a diference in population mean ~~betwee~~ between the amount of food ate
· he should do a paired t ~~t~~ test

Kareema says a paired t-test is **only** valid if both samples are selected from normally distributed populations.

(b) Explain why it is valid to use a paired t-test when both samples are from normally distributed populations.

(1)

because the diferences will then also be normaly distributed ~~you~~ ~~can~~ so you can use the mean diference to complete the test

(c) Explain whether or not it is necessary for **both** samples to be selected from normally distributed populations for a paired t-test to be valid.

(2)

~~No beca~~ not necessary because if one sample is from a uniform dis distribution and the other is normal, then the mean diference will still follow ~~a~~ a normal distribution

Q7

? ✓ A B C

(d) Carry out a hypothesis test to explain Justin's decision.

(5)

$H_0: \mu = 0$

~~$H_1:$~~

$H_1:$ ~~$\mu > 0$~~ $\mu \neq 0$

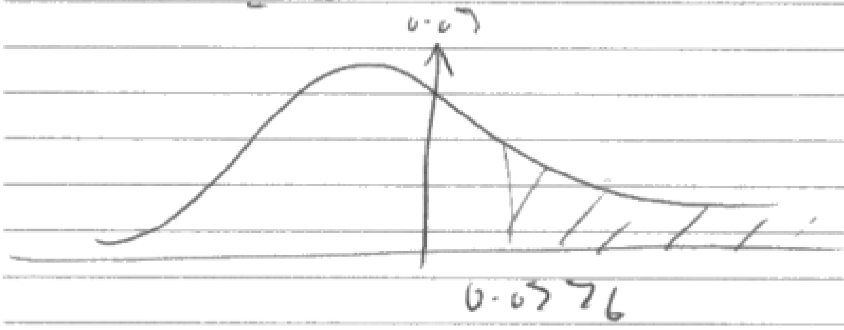
where μ is the mean population diference in hunger levels ~~bw~~

$\alpha = 0.05$

$\bar{X} \sim N\left(0, \frac{0.89^2}{20}\right)$

$t.s = 0.07$

$c.v. = 0.0776$



$t.s < c.v$ $\therefore$ ~~[illegible]~~ reject H_0

, sufficient evidence do ~~suggest~~ ~~that~~ ~~the~~ support Justin's decision

(e) Explain what is meant by a non-parametric test.

(1)

a test on data that comes from a population which ~~does~~ doesn't follow a distribution

6 /11

Part (a)

E1: Correctly identified that differences would need to be calculated.

E1: Correctly identified that the data from both weeks would be required and used context throughout answer.

Part (b)

E1: Correctly understood that a linear combination of two normal distributions would be normal.

Q7

?

A

B

C

Part (c)

E1: Had an attempt at explaining why it was not required.

E1: Correctly understood that the assumption is on the differences for paired t test and not the original population distribution.

Part (d)

B0: Hypotheses are not correct as did not use μ_d

B0: No attempt to use critical value formula (z test approach attempted).

B0: Incorrect critical value found.

M0: Incorrectly interpreted their test statistic and critical value.

E0: Conclusion has context based on the information from the question. This final mark was dependent on correct test statistic and critical value.

Part (e)

E1: Clear explanation which identifies the population assumption.

Question 7 - Response C

(a) Explain which data Justin should use if he wanted to carry out a paired t-test as part of his investigation.

(2)

Justin should use ~~the data~~ both data sets if he wants to carry out a paired t-test by subtracting the values ~~tested~~ for the ~~data~~ amount of food eaten on both wednesdays and then running a 1-sample t-test

Kareema says a paired t-test is **only** valid if both samples are selected from normally distributed populations.

(b) Explain why it is valid to use a paired t-test when both samples are from normally distributed populations.

(1)

~~The sample size is small~~ It is assumed that the samples ~~must~~ share a common Variance

(c) Explain whether or not it is necessary for **both** samples to be selected from normally distributed populations for a paired t-test to be valid.

(2)

~~The differences in the samples~~

For a ~~paired~~ t-test to be valid the samples must be random and independent, ~~be normal~~ be normal distributed, and share a common variance, ~~and~~ The central limit theorem cannot be applied as the sample size is not greater than 30. n=20. It is necessary.

(d) Carry out a hypothesis test to explain Justin's decision.

(5)

$H_0: \mu_d = 0$
$H_1: \mu_d \neq 0$
$\alpha = 0.05 \div 2$
$= 0.025$

$X \sim N(0.03, 0.89)$

$\bar{x} = 0.03$

$s_x = 0.89$

~~0.1507>2~~

$n = 20$

$0.1507 < 2.093$

$t = 0.1507$

$p = 0.8818$

Do not reject H_0, insufficient evidence to suggest a difference in the hunger level of the children on the two wednesdays.

CV =

$df = 19$

$t = 2.093$

(e) Explain what is meant by a non-parametric test.

(1)

~~A test that~~

A hypothesis test which is not rely on a distribution e.g normal, binomial, chi, poisson,

(Total for Question 7 is 11 marks)

7 /11

Part (a)

E1: Correctly identified that differences would need to be calculated, although they didn't state the word specifically.

E1: Correctly identified that the data from both weeks would be required and used context throughout answer.

Part (b)

E0: Did not identify that differences would be normally distributed from a linear combination.

Q7

?

A
B
C

Part (c)
E0: Did not correctly identify that the test would still be valid.
E0: Did not understand that the assumption is on the differences for paired t test and not the original population distribution.

Part (d)
B1: Hypotheses are correct and used μ_d
B1: Correctly obtained a standardised test statistic.
B1: Correct critical value found.
M1: Correctly interpreted their test statistic and critical value.
E0: Conclusion has context based on the information from the question, did not mention average. This final mark was dependent on correct test statistic and critical value.

Part (e)
E1: Although not as clear as Response B, the candidate included some examples of possible probability distributions so was awarded the mark.