National Qualifications

X803/77/11

Statistics
Paper 1

Duration — 1 hour

Total marks — 30

Attempt ALL questions.

You may use a calculator.

To earn full marks you must show your working in your answers.

State the units for your answer where appropriate.

Write your answers clearly in the answer booklet provided. In the answer booklet you must clearly identify the question number you are attempting.

Use **blue** or **black** ink.

Before leaving the examination room you must give your answer booklet to the Invigilator; if you do not, you may lose all the marks for this paper.

You may refer to the Statistics Advanced Higher Statistical Formulae and Tables.

X8037711

Total marks — 30

Attempt ALL questions

1. An extract of a draft report by a researcher is given below.

It is known to contain some flaws and questionable methodology.

Read it and then answer the questions that follow.

Introduction

For several years, Google has been helping generate training data for artificial intelligence (AI) neural networks to learn from. One project is called 'Quick Draw!' where players have 20 seconds to draw a doodle of an object. The time taken to draw each doodle is recorded, along with several other measurements. I found one website that compared the time taken to draw the doodle of a cat with that for a dog, but it did not conduct any analysis of the data and only presented some graphs to compare them. I wanted to explore if there was any statistically significant difference in the time taken to draw a doodle of each animal.

Method

The full data set has over 50 million doodles of 345 different objects, and I did not have the capabilities to process all of this data. I therefore requested two randomly sampled data sets of 150 doodles of cats and 150 doodles of dogs upon which to perform my analysis and from these I extracted the times taken (in seconds) to draw each.

However, within each set of 150 doodles there were a number of doodles that had not been recognised as either a cat or a dog by the AI algorithm. After removal of these, I had 121 valid doodles of cats and 145 valid doodles of dogs. It already seemed clear that there could be a difference when trying to draw a recognisable doodle of a cat or dog!

Data

Here is a back-to-back stem-and-leaf plot of the times taken to draw each animal, followed by computer software output summary statistics of each sampled data set:

	0	5
9	1	6
5778	2	1355577789999
256777	3	001222344555666788889
0011224446667789	4	011223445666666789999
2233444455566889	5	000122223334466677888
000012334444556677888889	6	111222344455566666
0122234555556666899	7	022223345568
0012233445667888	8	1366
11224556666999	9	268
001124456799	10	49
02233367779	11	34
2466	12	6
	13	9
3	14	
2	15	

	min	Q1	median	mean	Q3	max	sd
cats times	0.500	3.700	5.100	5.399	6.500	13.900	2.307
dogs times	1.900	5.500	7.300	7.498	9.500	15.200	2.655

1. (continued) **MARKS**

Analysis & Conclusion

I was not confident that I could assume that both the samples came from normal distributions, so I opted to use the Mann-Whitney Test to determine if the samples had different average drawing times.

After ranking all of the times in order (not shown), I obtained a rank sum of $W_{cats} = 12\,048$.

I then had to use a normal approximation which gave a z-test statistic of $-6{\cdot}57$ which is less than $-2{\cdot}58$ so it is highly significant. This proves that the average time to draw a doodle of a cat was less than that for a dog.

(a) The stem-and-leaf diagram could be improved by stating the number of leaves on each side of the stem. Describe three further improvements that should be made to the diagram to make it acceptable. **2**

(b) An alternative way to display the data would have been to use boxplots.

Determine how many outliers above the median would have to be shown on the boxplot for the drawing times of doodles of cats. **2**

(c) Read lines 15 to 19.

The author makes a claim based upon the number of recognisable doodles of each animal, without any statistical basis.

(i) Name a hypothesis test that could be performed to determine whether there was any difference in the success rate of drawing a recognisable doodle for each type of animal. **1**

(ii) Write down the hypotheses for the test you named in (i). **1**

(d) Read lines 24 to 26.

State the assumptions the author must have made in order to perform the Mann-Whitney test. **2**

(e) Read lines 27 to 30.

Assuming the value of the rank sum is correct, verify the stated value of the z-test statistic of the Mann-Whitney test, showing full calculations. **3**

(f) Read lines 7 to 9 and lines 24 to 31.

State the hypotheses of the test being performed, the associated level of significance used and write an improved conclusion to the test. **3**

(g) Read lines 24 to 26.

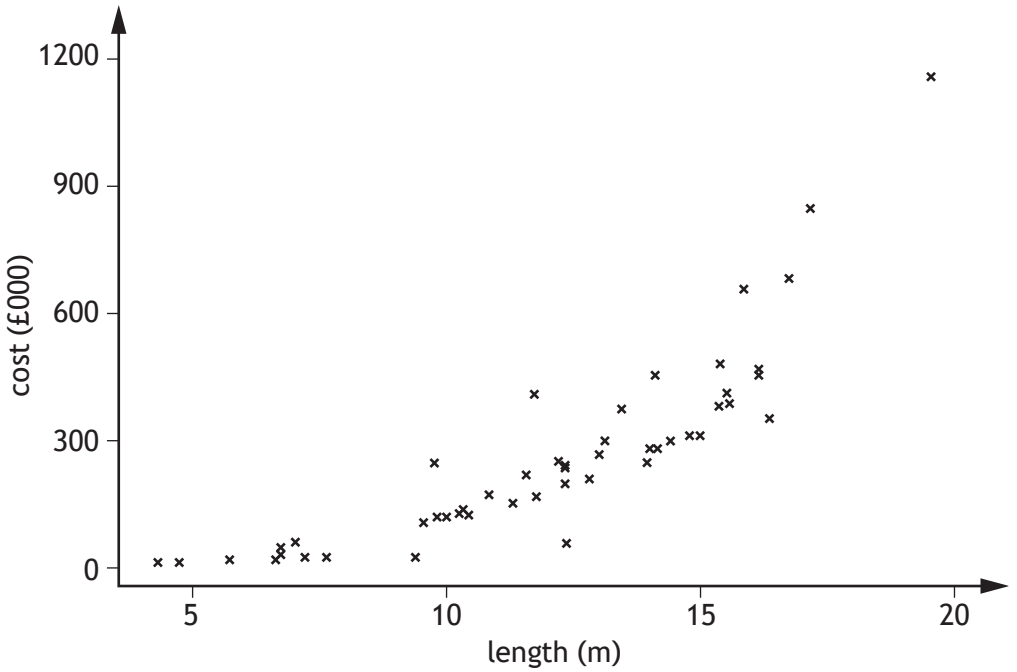
A reviewer of the report considers that the samples provide sufficient evidence that their parent distributions are roughly symmetrical and that it could be assumed that the distributions of times taken to draw cats and dogs are each normally distributed. In light of this they suggest that a t-test for a difference in population means could have been used.

State a further assumption required to perform this t-test and what information from the computer software output summary statistics table would be used to judge if it was valid. **2**

MARKS

2. An important factor influencing the cost of a sailing yacht is its length. To investigate this conjecture, a random sample of yacht lengths (in metres) and cost (pounds) was obtained. The sample data was selected from listings of second hand yachts sourced via an internet marketplace. A scatterplot of this data is shown in **Figure 1**.

Figure 1



(a) Comment on the relationship between the cost and length of a yacht. **2**

Two different transformations of the cost data were used to construct two linear regression models:

Model A used a square root transformation, regressing `sqrt(cost)` against `length`.

Model B used a base 10 logarithmic transformation, regressing `log10(cost)` against `length`.

The summary outputs for each model are given below.

Model A Output

```
model:  sqrt(cost) = -204.693 + 55.437 length
sample correlation coefficient, r = 0.8971518
t = 14.786, df = 53, p-value < 0.0001
alternative hypothesis: true correlation is not equal to 0
```

Model B Output

```
model:  log10(cost) = 3.70340 + 0.12582 length
sample correlation coefficient, r = 0.8948589
t = 14.595, df = 53, p-value < 0.0001
alternative hypothesis: true correlation is not equal to 0
```

(b) On the basis of the hypothesis tests in these outputs, comment upon the appropriateness of the transformation in each model. **2**

(c) Calculate the coefficient of determination for Model A and explain what its value means in this context. **2**

2. (continued)

Further checking of the Models A and B involved generating residual plots, shown in **Figures 2A** and **2B**.

Figure 2A

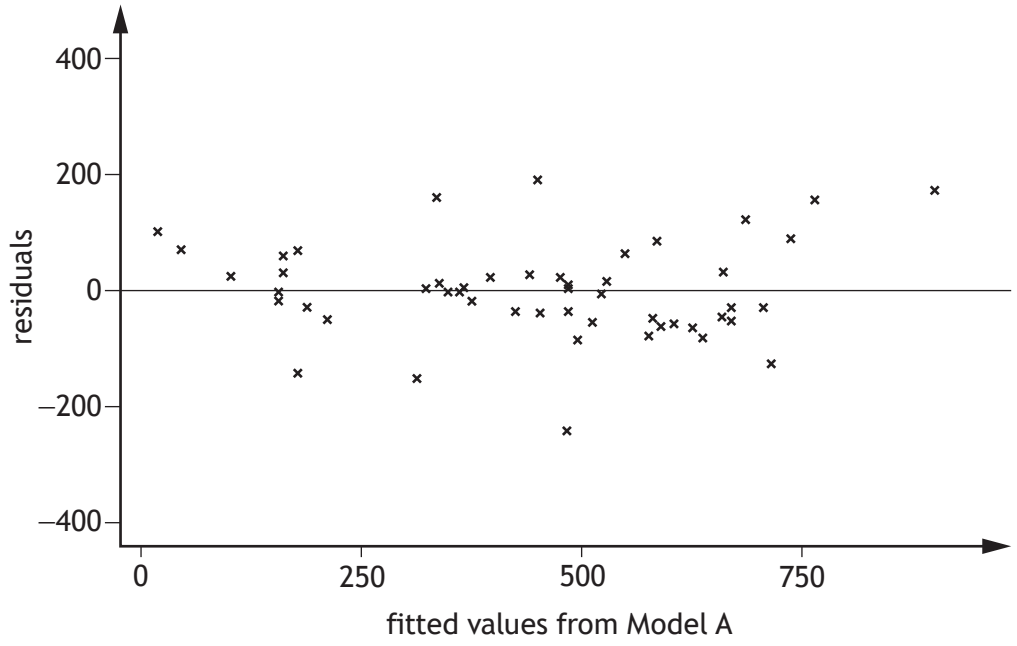
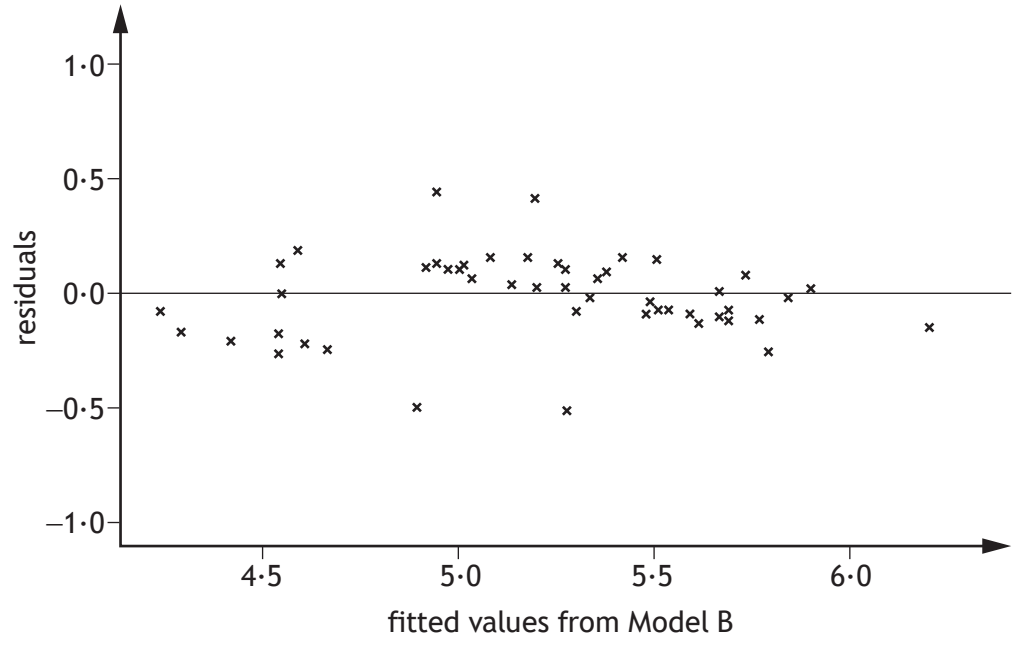


Figure 2B



(d) Use the residual plots to comment on the validity of each model. **3**

[Turn over

2. (continued)

Using Model A, the output from a 95% confidence interval for the transformed cost of a 15 metre long yacht is given below. One value has been deleted and replaced by *****.

Model A Confidence Interval Output

```
data: sqrt(cost) and length
sqrt(cost) = −204.693 + 55.437 length

variable    value
length      15

fit         SE(fit)     lower       upper
*****       95.24       592.0915    661.634
```

(e) Using this output for Model A, calculate the estimated cost and the 95% confidence interval for the actual cost of a 15 metre long yacht. **3**

Peter is considering spending £100 000 to purchase a yacht. He plans to substitute the value of 100 000 into each fitted model to obtain an estimate for the mean yacht length that he can buy at this price.

(f) (i) Give a reason why it would be inappropriate for Peter to substitute 100 000 for the cost into the equations of either model. **1**

(ii) Suggest a statistical process that Peter should do instead. **1**

[END OF QUESTION PAPER]

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE

National Qualifications

X803/77/12

Statistics Paper 2

Duration — 2 hours 45 minutes

Total marks — 83

SECTION 1 — 76 marks

Attempt ALL questions.

SECTION 2 — 7 marks

Attempt EITHER Part A OR Part B.

You may use a calculator.

To earn full marks you must show your working in your answers.

State the units for your answer where appropriate.

Write your answers clearly in the answer booklet provided. In the answer booklet you must clearly identify the question number you are attempting.

Use **blue** or **black** ink.

Before leaving the examination room you must give your answer booklet to the Invigilator; if you do not, you may lose all the marks for this paper.

You may refer to the Statistics Advanced Higher Statistical Formulae and Tables.

X8037712

MARKS

SECTION 1 — 76 marks

Attempt ALL questions

1. A biased, five-sided spinner is numbered with scores 2, 4, 6, 8 and 10.

Let S be the score on a single spin, where $\mathsf{P}(S = s) = \frac{s}{k}$, for some constant k.

(a) Determine the value of k and hence tabulate the probability distribution of S. **2**

(b) Calculate $\mathsf{E}(S)$ and $\mathsf{V}(S)$. **2**

2. A primary school has the following staff.

	Teachers	Admin	Other
Female	18	7	5
Male	12	3	5

A member of staff is selected at random.

F is the event that the person selected is female, T is the event that the person selected is a teacher and A is the event that the person selected is admin staff.

(a) Find the probabilities of $\mathsf{P}(\mathrm{F} \cap \mathrm{T})$ and $\mathsf{P}(\mathrm{F} \cup \overline{\mathrm{A}})$. **2**

(b) Given that 80% of the teachers, 50% of the admin staff and 30% of the other staff drive to school, calculate the probability that

(i) a randomly selected member of staff drives to school **2**

(ii) a randomly selected member of staff is one of the admin staff, given that they did not drive to school. **3**

3. A parasitic mite is known to impact on the health of honey bee colonies. Previous research suggests that a median mite count of greater than 7 has a serious negative impact on colony health.

A beekeeper wishes to determine if the colony's health is at risk. The number of mites is recorded each day by placing a collecting tray below the colony and the mites fall out onto the tray and are then counted.

The beekeeper records the daily mite count for a colony over a representative 14 day period as follows:

6	8	11	13	6	14	11
9	6	7	11	8	6	14

Stating a necessary assumption, perform a Wilcoxon Signed-Rank test to determine whether the colony's health is at risk. **8**

4. A researcher is studying bank vole populations across 3 different habitats: woodland, farmland and moorland. The researcher uses non-destructive Ugglan traps to monitor bank vole numbers in a 24 hour period. Over the six week study period the mean number of bank voles captured in each 24 hour period can be modelled by the following Poisson distributions.

Woodland: $W \sim \text{Po}(5)$

Farmland: $F \sim \text{Po}(2{\cdot}3)$

Moorland: $M \sim \text{Po}(1{\cdot}2)$

In a given 24 hour period calculate the probability that the researcher

(a) (i) captures more than 10 bank voles at the woodland habitat **2**

(ii) gets exactly 2 captures across the non-woodland sites. **2**

(b) Using a suitable approximation calculate the probability that there are fewer than 340 captures across all three habitats in the 6 week period. **5**

5. A flooring installer wishes to know if there is a difference in warm up time (minutes) for two different brands of underfloor heating mats.

They obtain a random sample of 11 warm up times of brand A and another random sample of 15 warm up times for brand B. The summary statistics for the independent samples are provided below.

	Mean	**Standard deviation**
Brand A	54	5
Brand B	47	11

Carry out a t-test using the following hypotheses at the 5% level of significance.

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B$$

5

MARKS

6. Several studies have been conducted to investigate the influence of screen time on young people in the USA and the impact this has on their sleeping patterns.

In one study, students completed anonymous written questionnaires in rooms with trained researchers present. Students were chosen randomly, ensuring various ethnic groups were proportionally represented.

(a) State the sampling strategy used in this study. **1**

In a second study, students participated in an online survey having been recruited through an advertisement in a subscription-based online newsletter targeted at parents.

(b) State the sampling strategy used in the second study, giving a reason why results from this study might not be reliable. **2**

A statistically robust and broad-ranging survey of students' screen time found that the population mean daily screen time is 7 hours and 38 minutes ($\mu = 458$ minutes) with standard deviation, $\sigma = 130$ minutes.

A school was concerned about the levels of screen time experienced by its students and therefore offered to reward students if they reduced their screen time. Two months later, a random sample of 25 students was chosen and asked to record their screen time that day. The mean daily screen time for this group was 6 hours and 49 minutes (409 minutes).

(c) From this random sample, construct a 95% confidence interval for the population mean daily screen time, assuming that the standard deviation is still 130 minutes. **2**

A student representative calculated the same 95% confidence interval. They also calculated a 90% confidence interval which was (366·4, 451·6). When they met the school's headteacher to discuss the students' attempts to reduce screen time, they only presented the 90% confidence interval.

(d) Explain, with a reason, why the student representative might have presented the 90% confidence interval, rather than the 95% confidence interval to the headteacher. **2**

7. A continuous random variable X is uniformly distributed between 78 and 83.

$\overline{X}$ is the mean of 75 randomly observed values of X.

(a) Calculate $P(80{\cdot}45 < X < 80{\cdot}83)$. **1**

(b) Determine the distribution of $\overline{X}$ with justification. **3**

(c) Calculate $P(80{\cdot}45 < \overline{X} < 80{\cdot}83)$. **2**

MARKS

8. As part of a new company interview process, candidates are expected to solve a problem, on paper, where the time taken has been designed to be normally distributed with mean 15 minutes and standard deviation 2 minutes.

The times recorded for a random sample of 50 of these candidates has a mean of 16·1 minutes.

Perform a hypothesis test to assess whether or not there is any cause for concern about candidates taking too long at the task, stating a possible explanation for the result of your test. **7**

9. This question relates to insurance policies, from the viewpoint of an insurance company. Customers may choose from two different policies, A and B, which would pay out in the event of a claim that has a constant probability of happening. Both policies have the same monthly premium to be paid of £10.

Define the random variables A and B to be the monthly profit, in pounds, for the insurance company from each £10 paid in monthly premiums for the respective individual policies.

It is given that

$$\text{E}(A) = 2{\cdot}50 \qquad \text{SD}(A) = 4{\cdot}00$$
$$\text{E}(B) = 1{\cdot}00 \qquad \text{SD}(B) = 5{\cdot}00$$

The random variable C is defined as $A - B$.

(a) Calculate the expectation and variance of C. **2**

(b) Explain what the random variable C represents in this context. **1**

After a short advertising campaign, the insurance company sells 33 of policy A and 26 of policy B.

(c) Calculate the standard deviation of the total monthly profit that the insurance company can expect to make from these new policies. **3**

10. In a random sample of 50 UK shops selling fresh produce it was discovered that 13 of them had been selling produce past its sell-by date.

(a) Calculate a 95% confidence interval for the proportion of all UK shops selling produce past its sell-by date. **3**

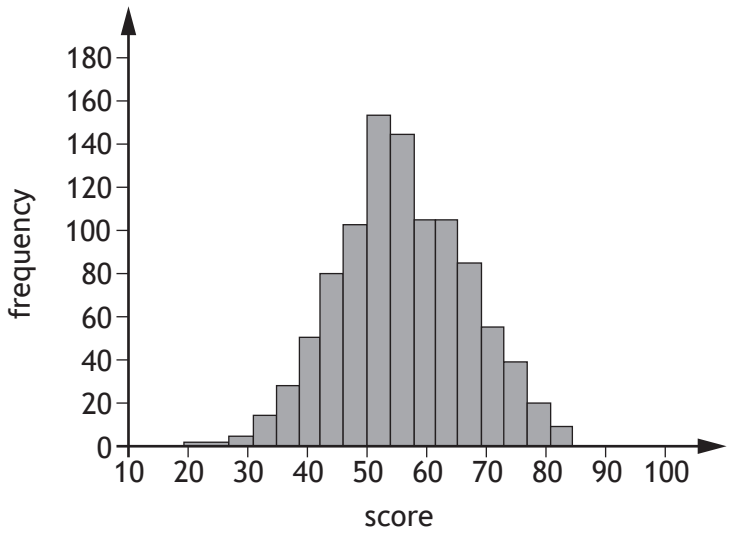
As part of future investigations, it is desired to have a more precise estimate of the proportion of shops by constructing a 95% confidence interval of width 0·04.

(b) Determine the minimum sample size required to achieve this, using the estimated proportion from the original random sample. **3**

[Turn over

11. Shift work is the term used to describe the working patterns of people whose working hours can vary from one week to another and who work at times outwith the normal working day. It is thought that shift work may have negative physical health consequences and a recent research study investigated the effects of shift work on mental capacities. **MARKS**

Participants in the study were asked to take a series of tests on different aspects of brain function and from these tests, an overall score was obtained for each participant. A low score may indicate lower mental capacity. A previous study on brain function had produced the results illustrated in the graph below.



(a) Comment on the distribution of scores that would inform any further statistical analysis. **1**

Researchers designed the study to attempt to discover if there were any differences in brain function between three different groups of workers. Those who had never undertaken shift work, those who had undertaken shift work up to and including 10 years and those who had undertaken shift work for more than 10 years.

Random samples of the three different groups of workers were taken and the scores are summarised in the table below, where $\overline{x}$ = sample mean, s = sample standard deviation and n = sample size.

Group	Shift work duration	$\overline{x}$	s	n
A	never	56·0	10·71	120
B	$\leq$ 10 years	55·4	10·08	70
C	$>$ 10 years	51·8	10·49	60

(b) Stating a necessary assumption, different from that mentioned in part (a), perform an appropriate statistical test to determine if the mean score for group B is different to that for group C. **7**

(c) Determine the maximum mean score for group B that would indicate that there is evidence, at the 10% level of significance, that the mean score for group A is greater than that for group B. **3**

[END OF SECTION 1]

MARKS

SECTION 2 — 7 marks

Attempt EITHER Part A OR Part B

Part A

12. Cereal bags are filled by a machine with their weights following a normal distribution with mean 500 grams and standard deviation 5·73 grams. Every 4 hours a random sample of 5 bags is taken and weighed. An x-bar chart is constructed to monitor the mean cereal bag weight.

(a) Calculate the 1σ limits for the x-bar chart. **2**

(b) (i) Calculate the probability that at least two of three consecutive sample means fall beyond the same 1σ limits. **4**

(ii) Explain why this probability would not be a good indication that the process is out of control. **1**

[Turn over

13. At the end of their first term at university, the same Statistics exam was taken by both Psychology and Biology students, who had been taught by different lecturers.

The frequencies of grades awarded to a random sample of students are shown below.

	Grade					
	A	**B**	**C**	**D**	**E**	**Total**
Psychology	13	11	6	5	1	36
Biology	4	4	6	4	5	23
Total	17	15	12	9	6	59

As part of a research project, an educational research student conducted a chi-squared test on the above data using the hypotheses given below.

H_0: the statistics grades are independent of the course subjects
H_1: the statistics grades are not independent of the course subjects

They calculated the chi-squared statistic to be 8·35 and concluded at the 10% significance level that they had evidence supporting the claim that the Statistics exam grades that had been awarded were not independent of the subject that had been studied.

(a) Give a possible explanation for why this claim might be true, with reference to the context of the research project. **1**

As part of their calculations, the research student had correctly generated the expected frequencies to be those shown below.

	Grade				
	A	**B**	**C**	**D**	**E**
Psychology	10·4	9·2	7·3	5·5	3·7
Biology	6·6	5·8	4·7	3·5	2·3

It is suspected by their supervisor that the hypothesis test had not been performed correctly by the research student. The supervisor used the same hypotheses and level of significance as the research student and conducted an improved chi-squared test.

(b) Carry out this improved chi-squared test, stating your conclusion. **5**

(c) Determine the course subject and exam result that contributed the most to the chi-squared test statistic that was calculated in part (b). **1**

[END OF SECTION 2]

[END OF QUESTION PAPER]