National
Qualifications
2021 ASSESSMENT RESOURCE

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.

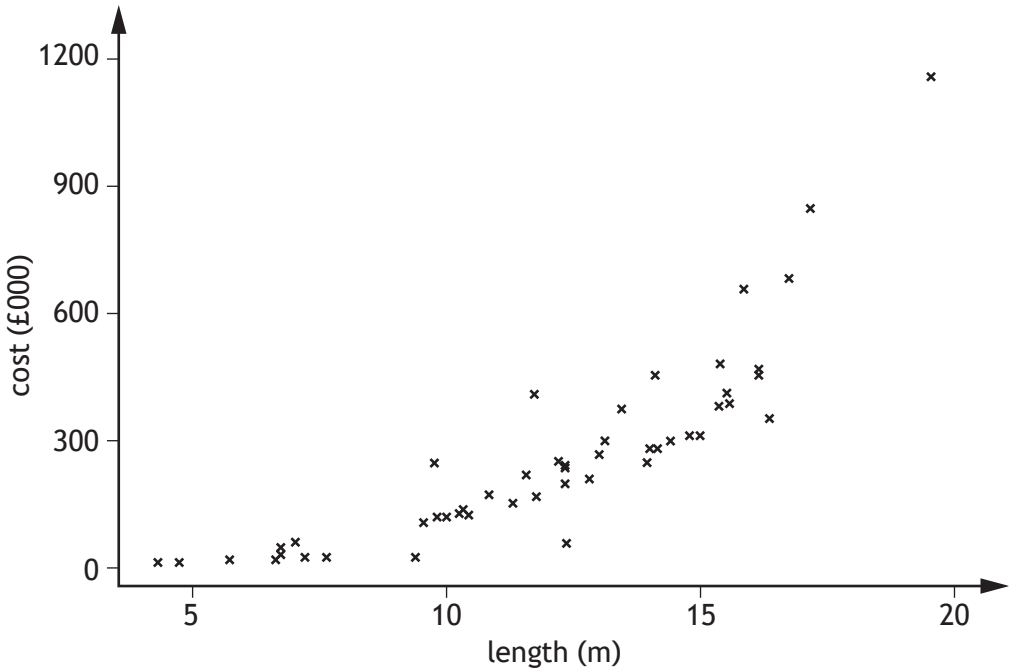
A reviewer of the report considers that the samples provide sufficient evidence that their parent distributions are roughly symmetrical and that it could be assumed that the distributions of times taken to draw cats and dogs are each normally distributed. In light of this they suggest that a t-test for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

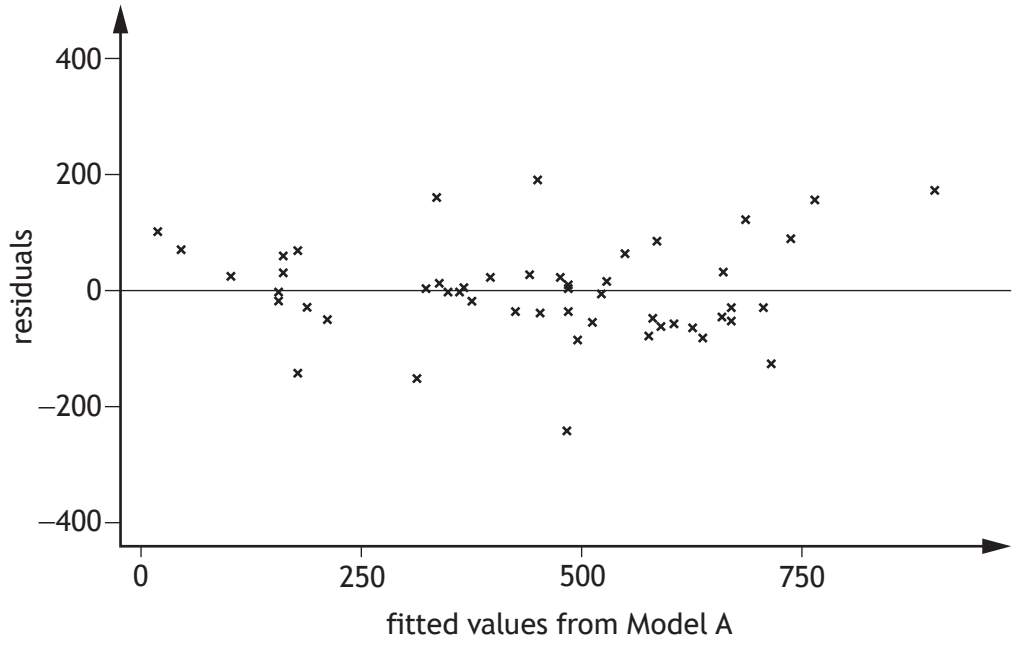
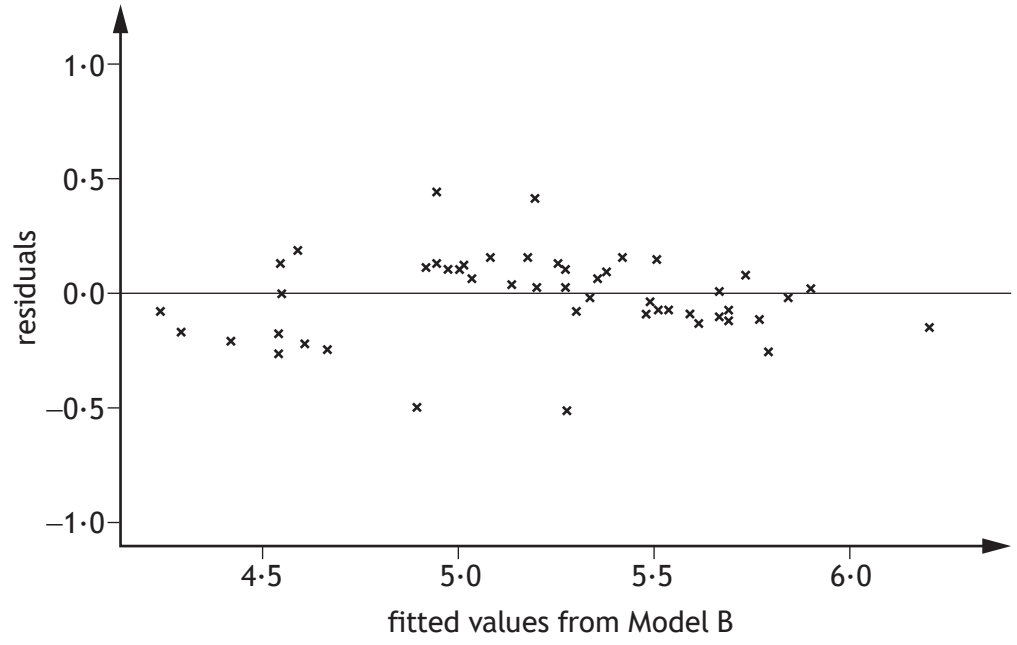


Figure 2B



(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