Write your name here

Surname	Other names

Pearson Edexcel Level 3 GCE

Centre Number	Candidate Number

Statistics

Advanced Subsidiary
Paper 1

Tuesday 19 June 2018 – Afternoon
Time: 1 hour 30 minutes

Paper Reference
8ST0/01

You must have:
Statistical Formulae and Tables booklet
Calculator

Total Marks

Candidates may use any calculator allowed by the regulations of the Joint Council for Qualifications. Calculators must not have retrievable mathematical formulae stored in them.

Instructions

- Use **black** ink or ball-point pen.
- If pencil is used for diagrams/sketches/graphs it must be dark (HB or B). Coloured pencils and highlighter pens must not be used.
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions and ensure that your answers to parts of questions are clearly labelled.
- Answer the questions in the spaces provided – *there may be more space than you need.*
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.
- Unless otherwise stated, statistical tests should be carried out at the 5% significance level.
- When a calculator is used, the answer should be given to three significant figures unless otherwise stated.

Information

- A booklet 'Statistical Formulae and Tables' is provided.
- There are 6 questions in this question paper. The total mark for this paper is 60.
- The marks for **each** question are shown in brackets – *use this as a guide as to how much time to spend on each question.*

Advice

- Read each question carefully before you start to answer it.
- Try to answer every question.
- Check your answers if you have time at the end.
- If you change your mind about an answer, cross it out and put your new answer and any working underneath.

Turn over ▶

P58389A

1/1/1/1/1/1/

P 5 8 3 8 9 A 0 1 2 4

Pearson

Answer ALL questions. Write your answers in the spaces provided.

1 Below is a list of five statistical expressions.

$$\sum x_i^3 \qquad \sigma^2 \qquad \overline{x} \qquad X \qquad \sum (x_i - \overline{x})^2$$

Write down which of these expressions are

(i) **statistics**,

(ii) **parameters**.

(4)

(Total for Question 1 is 4 marks)

DO NOT WRITE IN THIS AREA

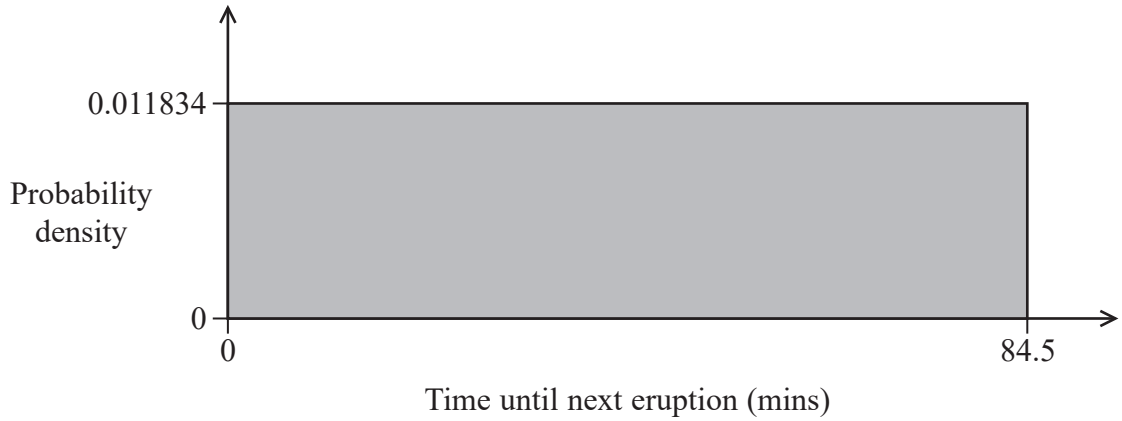
DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

2 A geyser is a water spring that regularly ejects hot water and steam up into the air in volcanic regions of the world.

Old Faithful is a famous geyser in Yellowstone National Park, which erupts regularly.

If I arrive at the geyser at a random time, the time until the next eruption can be modelled by a uniform distribution, as shown in **Figure 1**.



(Based on real data)

Figure 1: Probability density function of the time until the next eruption of Old Faithful

(a) Explain why the mean is equal to the median in this distribution.

(1)

(b) I arrive at a random time and stay for 20 minutes.

Find the probability that I will see Old Faithful erupting.

(2)

(c) I arrive at a random time every day for a seven-day week and stay for 20 minutes each day.

Find the probability that I will see Old Faithful erupting at least twice.

(3)

(Total for Question 2 is 6 marks)

3 Neville is trying to research fitness levels of people under 25 for his PhD. He is particularly interested in the exercise habits of young people with diabetes.

He decides to collect data by writing a web questionnaire called 'Are you as fit as Usain Bolt?'

In the questionnaire, people answer several questions about their health, exercise and diet, and then they are presented with a score out of 100%, of the form

'You are 64% as fit as Usain Bolt!!'

People can then share this result on social media, with a link added for anyone else who might like to take the test.

(a) State the name of this sampling technique.

(1)

(b) State two possible advantages of this sampling technique.

(2)

One question on Neville's questionnaire reads:

> Do you complete at least 1 hour of vigorous exercise every week?
> [e.g. running, football, spin class, brisk swimming, aerobic dancing]

Neville filters his database to only include those people under 25 **with diabetes**, and runs a query to produce the table in **Figure 2**.

Yes	38
No	30

Figure 2: Responses to the above question for people under 25 with diabetes

The response to this question for **all** people under 25 was 61% yes.

(c) Making any necessary assumptions, use the information given to test whether the proportion of the population of people under 25 with diabetes who would answer 'yes' to this question is equal to 61%.

(6)

Question 3 continued

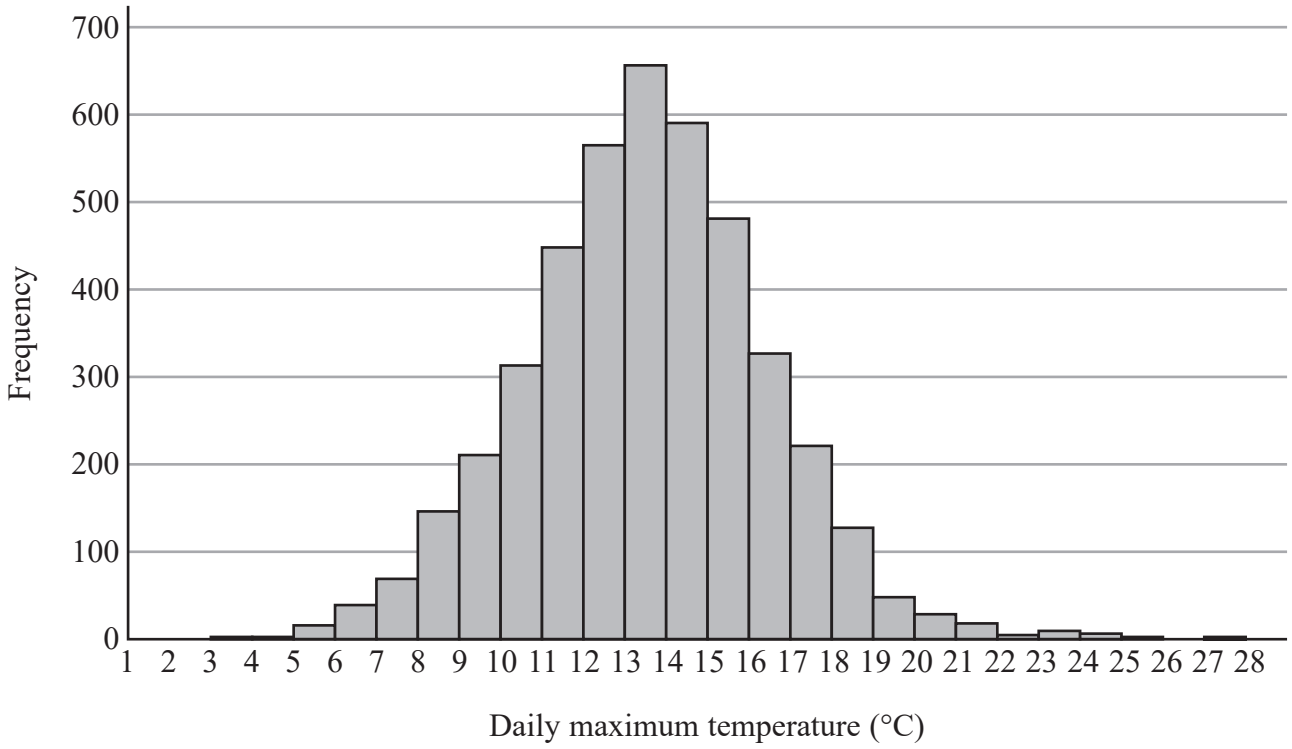
(d) Give two possible explanations for why Neville's data may not reflect the true proportion of people under 25 who complete at least one hour of vigorous exercise per week.

(2)

(Total for Question 3 is 11 marks)

4 The Central England Temperature (CET) dataset is the longest running record of weather temperature in the world.

The dataset contains a record of the maximum air temperature in Central England for each day since 1878. The data for all October days in this period is presented in the histogram in **Figure 3**.



(Data source: HadCET dataset)

Figure 3: Daily maximum temperatures for October days in Central England, 1878–2016

Analysis of this data in a spreadsheet program produced the following summary statistics:

COUNT	4340
AVERAGE	13.47
VAR.S	8.27

(a) Using a suitable modelling distribution, estimate the probability that the maximum temperature will be below 11.5 °C on a randomly chosen October day in Central England.

You should fully describe the distribution that you use.

You should justify your choice of distribution.

(3)

The seeds of British bluebells have been observed to germinate in the autumn months when the maximum daily temperature **first** falls below 11.5 °C.

(b) Explain why the probability you estimated in (a) is not necessarily the same as the probability that bluebells will germinate on a randomly chosen October day in Central England.

(1)

Carolyn plans to walk out to the woods to study the bluebells on five occasions in October 2018. You may assume that these five days are a random sample of October days in Central England.

(c) Using your model, estimate the probability that the **mean** maximum temperature of these five days is below 11.5 °C.

(3)

(d) Give two negative criticisms of using your model to predict the maximum temperature on a particular day in Central England during October 2018.

In each case, suggest how you could improve your model to rectify this issue.

(4)

(Total for Question 4 is 11 marks)

BLANK PAGE

5 Deshandra is writing a report on the consumption of hot drinks around the world. She believes that countries where more coffee is consumed are likely to consume less tea.

She wants to test this belief by carrying out a hypothesis test.

She finds two tables of data on the internet titled

- 'The world's 50 biggest coffee drinking countries (in kg per capita per year)'
- 'Top 20 tea consuming nations (kg/capita/yr)'

(a) Explain why the data in these tables might be unsuitable for comparing consumption of tea and coffee by country.

(1)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

She does some further research and finds some new datasets. The **top 10 rows** of each dataset are presented in **Figures 4 and 5**.

Country	Tea consumption per capita (kg)
Turkey	3.16
Ireland	2.20
United Kingdom	1.95
Russia	1.39
Morocco	1.22
New Zealand	1.20
Egypt	1.01
Poland	1.00
Japan	0.97
Saudi Arabia	0.90
⋮	⋮

(Source: Adapted from real data)

Figure 4: Tea consumption by country in 2015

Country	Coffee consumption per capita (kg)
Finland	9.5
Norway	7.1
Netherlands	6.6
Slovenia	6.0
Austria	5.4
Serbia	5.3
Denmark	5.2
Germany	5.1
Belgium	4.8
Brazil	4.7
⋮	⋮

(Source: Adapted from real data)

Figure 5: Coffee consumption by country in 2015

(b) Explain how you would use spreadsheet software to organise the datasets into a format suitable for investigating correlation.

(3)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

There are some inconsistencies in the two datasets, as well as some missing data.
Deshandra decides to use a random sample taken from the countries that consistently feature in both datasets.

The data for this random sample is presented in **Figure 6**.

Country	Tea consumption (kg/capita/yr)	Coffee consumption (kg/capita/yr)
Netherlands	0.78	6.6
Japan	0.97	1.4
Estonia	0.43	4.1
Poland	1	3.0
United Kingdom	1.95	1.6
Tunisia	0.92	2.1
Malaysia	1.75	1.0
New Zealand	1.2	1.1
Georgia	0.75	1.3
Hong Kong	0.87	1.2

(Source: Adapted from real data)

Figure 6: Tea and coffee consumption for Deshandra's random sample of countries

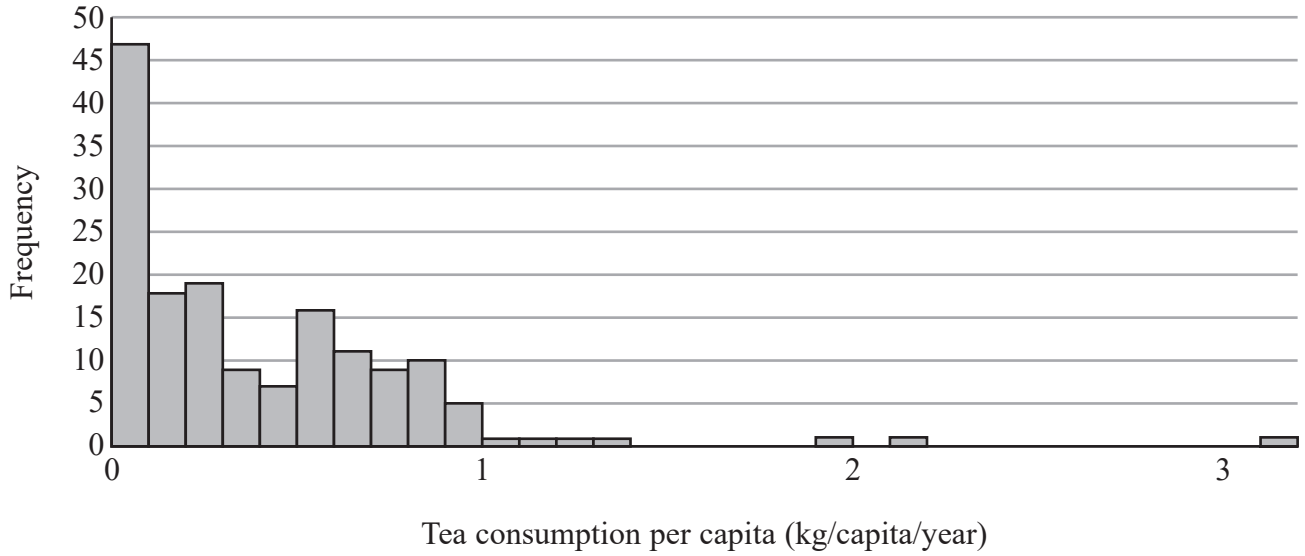
(c) Carry out a hypothesis test on Pearson's product-moment correlation coefficient to investigate Deshandra's belief that countries where more coffee is consumed are likely to consume less tea.

You should make clear the assumption that it is necessary for you to make in order for the test to be valid.

(6)

Question 5 continued

Deshandra then writes her report. She starts by constructing a histogram to show the distribution of tea consumption per capita, by country. This is presented in **Figure 7**.



(Source: Adapted from real data)

Figure 7: Distribution of tea consumption per capita, by country

(d) Explain, in context, why the assumption that you made in part (c) is unlikely to be valid.

(2)

(e) Suggest an alternative statistical measure that Deshandra could use for investigating correlation between coffee and tea consumption.

(1)

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

Deshandra finds the following information on a website that she includes in her report:

Mean consumption per person (whole world)

Tea: 0.675 kg/person/year

Coffee: 1.311 kg/person/year

(Source: Adapted from real data)

Deshandra then states in her report:

'In the world, twice as many cups of coffee are consumed than cups of tea.'

(f) Explain why this statement may not be reliable.

(2)

(Total for Question 5 is 15 marks)

6 The table in **Figure 8** is part of the 'Future home owners' section of the English Housing Survey 2015–16. This data is based on fieldwork carried out between April 2015 and March 2016 on a large sample of households.

Annex Table 1.11: Expectation to buy by ethnicity and tenure, 2015–16
all renters, excluding those who also own

		expect to buy	don't expect to buy	total	sample size
			thousands of households		
local authority	White	253	983	**1236**	*1210*
	all other ethnic groups	180	167	**348**	*301*
	total	433	1150	**1584**	*1511*
housing association	White	448	1516	**1964**	*1865*
	all other ethnic groups	155	160	**315**	*225*
	total	603	1677	**2279**	*2090*
all social renters	White	701	2499	**3200**	*3075*
	all other ethnic groups	335	328	**663**	*526*
	total	1036	2827	**3863**	*3601*
private renters	White	2077	1446	**3524**	*1638*
	all other ethnic groups	497	309	**806**	*332*
	total	2574	1755	**4329**	*1970*
total	**White**	**2778**	**3945**	**6724**	*4713*
	all other ethnic groups	**832**	**636**	**1468**	*858*
	total	**3610**	**4582**	**8192**	*5571*

(Source: Gov.uk)

Figure 8

(a) Carry out a hypothesis test to investigate whether there is an association between ethnic group ('White' or 'all other ethnic groups') and expectation to buy ('expect to buy' or 'don't expect to buy') for **private renters** in 2015–16.

(11)

Question 6 continued

Question 6 continued

(b) For each row in **Figure 8**, the value in the 'total' column is greater than the value in the 'sample size' column. Explain, in context, why this is the case.

(2)

(Total for Question 6 is 13 marks)

TOTAL FOR PAPER IS 60 MARKS

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

BLANK PAGE