Write your name here

Surname	Other names

Pearson Edexcel GCSE

Centre Number □□□□□ Candidate Number □□□□

Statistics

Paper 1H

Higher Tier

Thursday 22 June 2017 – Morning
Time: 2 hours

Paper Reference
5ST1H/01

You must have:
Ruler graduated in centimetres and millimetres, protractor, pen HB pencil, eraser, electronic calculator.

Total Marks

Instructions

- Use **black** ink or ball-point pen.
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions.
- Answer the questions in the spaces provided – *there may be more space than you need.*

Information

- The total mark for this paper is 100.
- The marks for **each** question are shown in brackets – *use this as a guide as to how much time to spend on each question.*
- Questions labelled with an **asterisk** (*) are ones where the quality of your written communication will be assessed – *you should take particular care on these questions with your spelling, punctuation and grammar, as well as the clarity of expression.*

Advice

- Read each question carefully before you start to answer it.
- Keep an eye on the time.
- Try to answer every question.
- Check your answers if you have time at the end.

Turn over ▶

Pearson

Higher Tier Formulae

You must not write on this page.
Anything you write on this page will gain NO credit.

Mean of a frequency distribution $= \dfrac{\sum fx}{\sum f}$

Mean of a grouped frequency distribution $= \dfrac{\sum fx}{\sum f}$, where x is the mid-interval value.

Variance $= \dfrac{\sum (x-\bar{x})^2}{n}$

Standard deviation (set of numbers) $\sqrt{\left[\dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2\right]}$

or $\sqrt{\left[\dfrac{\sum (x-\bar{x})^2}{n}\right]}$

where $\bar{x}$ is the mean set of values.

Standard deviation (discrete frequency distribution) $\sqrt{\left[\dfrac{\sum fx^2}{\sum f} - \left(\dfrac{\sum fx}{\sum f}\right)^2\right]}$

or $\sqrt{\left[\dfrac{\sum f(x-\bar{x})^2}{\sum f}\right]}$

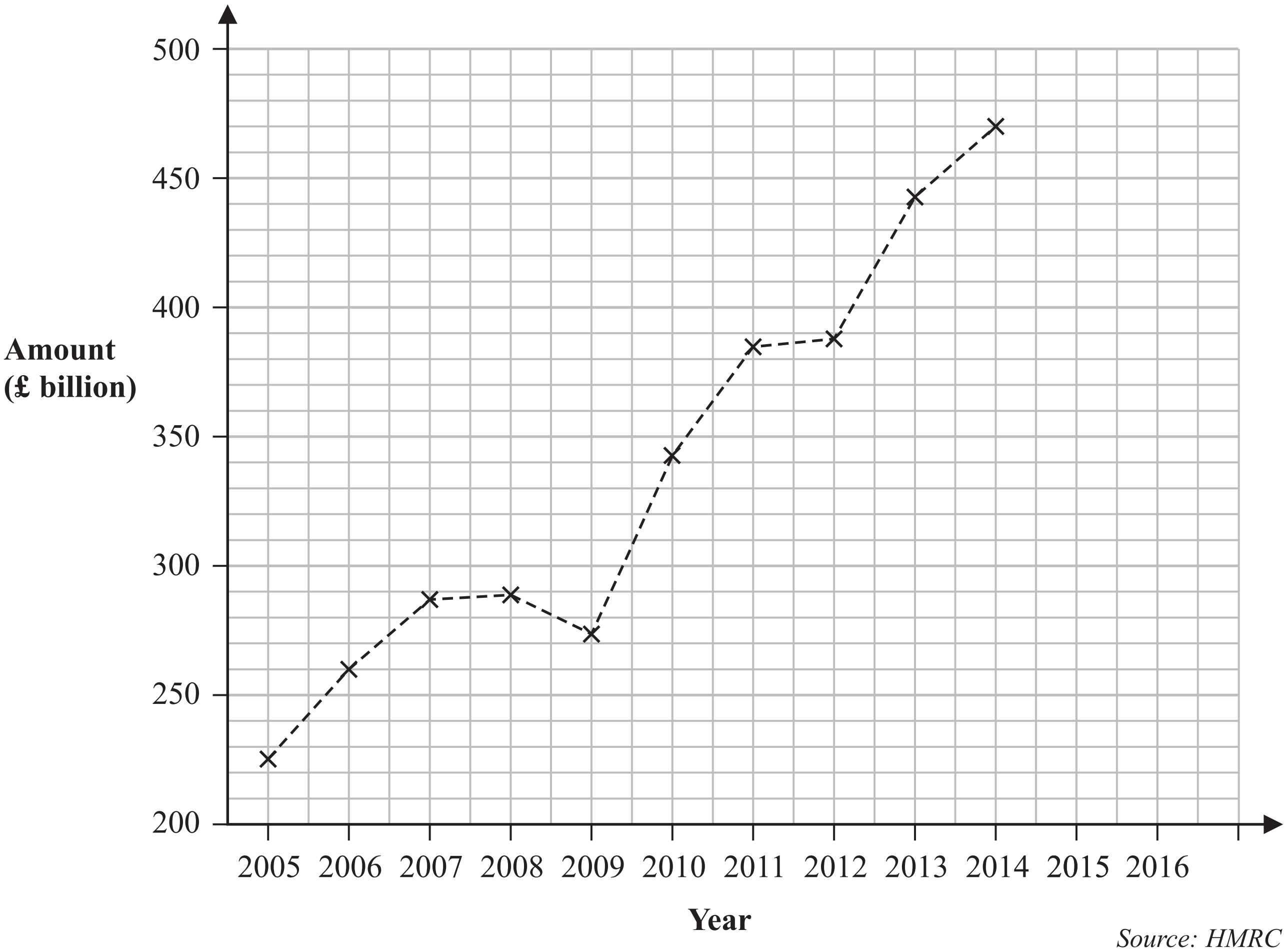
Spearman's Rank Correlation Coefficient $1 - \dfrac{6\sum d^2}{n(n^2-1)}$

Answer ALL the questions.

Write your answers in the spaces provided.

You must write down all the stages in your working.

1 The time series graph shows the total amount of money (£ billion) in Individual Savings Accounts in the United Kingdom for the years 2005–2014



Source: HMRC

(a) Write down the total amount of money in Individual Savings Accounts for the year 2014

£.. billion

(1)

(b) (i) Draw a trend line on the time series graph.

(ii) Describe the trend.

....

(2)

(c) Explain why using the trend line to predict the total amount of money in Individual Savings Accounts for the year 2016 may be unreliable.

....

....

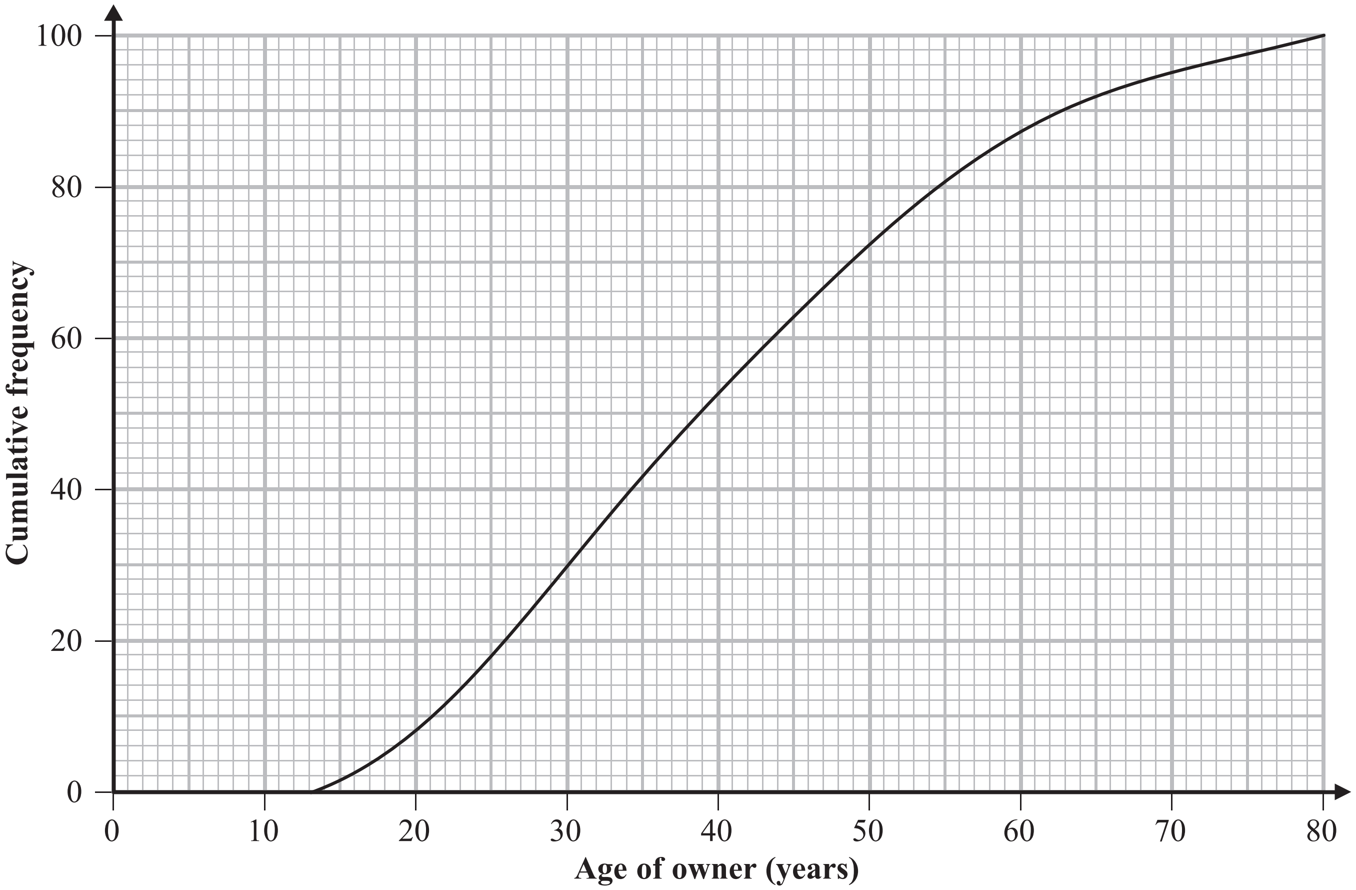
(1)

(Total for Question 1 is 4 marks)

2 Kunal was investigating the ages of owners of electronic tablets.

He used information from a survey carried out in the USA in 2012 to find the age distribution for a representative 100 people.

Kunal then drew this cumulative frequency graph for his information.



Source: adapted from comScore

(a) Find the number of these electronic tablet owners that are

(i) under 30 years old,

....

(ii) between 60 and 70 years old.

....

(3)

Kunal wants to use this survey to predict the percentage of electronic tablet owners in the **United Kingdom** that are under 30 years old.

(b) Explain whether or not it is sensible to use the results of this survey for his prediction.

....

....

....

....

....

(2)

This table gives information, from the same 2012 USA survey, about the ages of owners of smartphones.

Median	36 years
Interquartile range	23 years

*(c) Use the information from the table and from the graph to compare the ages of owners of electronic tablets with the ages of owners of smartphones in the USA.

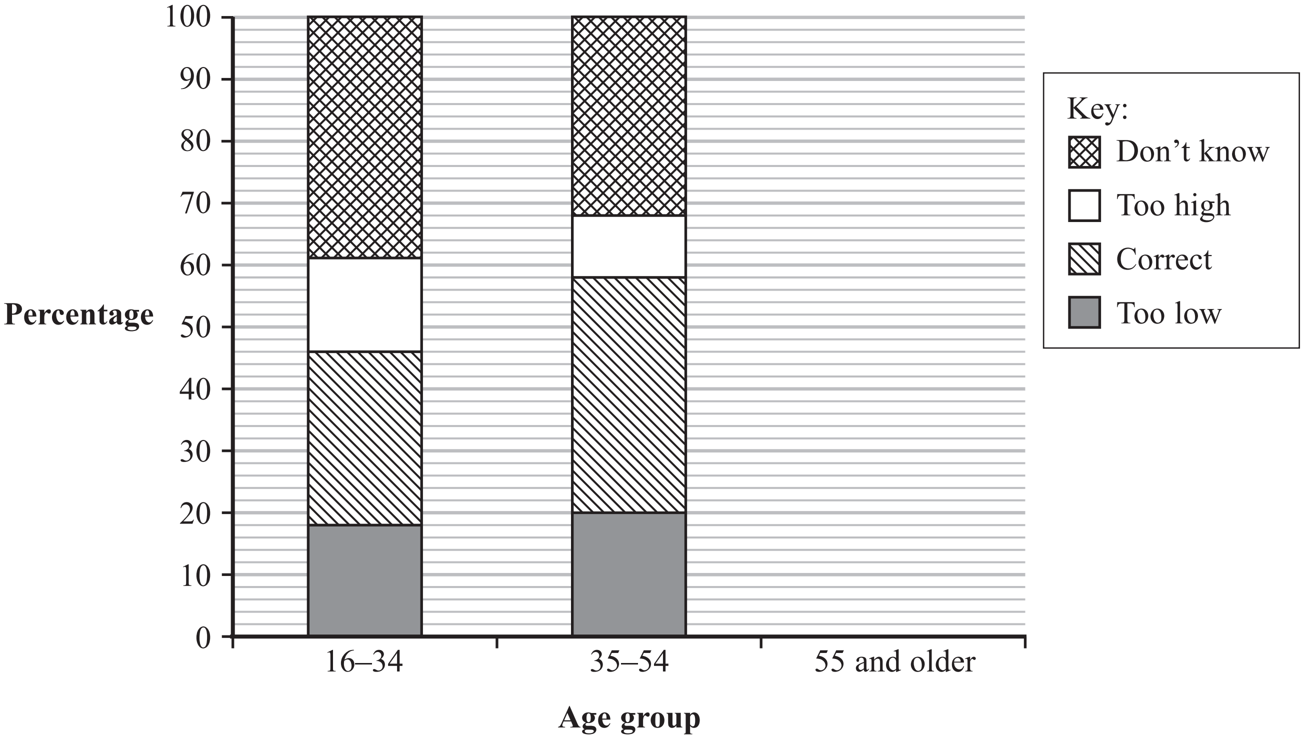
(4)

(Total for Question 2 is 9 marks)

3 In a survey people were asked this question.

"How much does a first class stamp cost for a standard letter?"

The composite bar chart shows information about the answers given by people in two different age groups.



Source: Ofcom Residential Postal Tracker

(a) Use the composite bar chart to find the percentage of the 35–54 age group that gave an answer that is

(i) too low,

....%

(ii) correct.

....%

(3)

Here is information about the answers given by people in the 55 and older age group.

	Age group
Cost of first class stamp	**55 and older**
Too low	16%
Correct	38%
Too high	14%
Don't know	32%
Total	100%

(b) Use the information in the table to complete the composite bar chart for people in the 55 and older age group.

(3)

(c) Compare the answers given by people in the 16–34 age group with the answers given by people in the 35–54 age group.

....

....

....

....

(2)

(Total for Question 3 is 8 marks)

4 A research company wants to find out how likely people in Great Britain are to vote in the next General Election.

The population is defined as all the voters in Great Britain.

(a) Suggest a suitable sampling frame.

....

(1)

The company is going to carry out its research by telephone.

(b) Give an advantage of collecting this information by telephone rather than using a postal questionnaire.

....

....

(1)

(c) Give one possible source of bias with a telephone survey.

....

....

(1)

The company asked 1000 people the following question.

On a scale from 1 to 10, with 1 being certain not to vote and 10 being certain to vote, how likely are you to vote in the next General Election?

1	2	3	4	5	6	7	8	9	10
Certain not to vote									Certain to vote

The table shows the percentage of people giving each response.

Response	1	2	3	4	5	6	7	8	9	10
Percentage	6	2	3	1	8	3	4	7	5	61

Source: Ipsos MORI

(d) Work out the number of people giving a response of 6 or higher.

....

(2)

(Total for Question 4 is 5 marks)

5 Ruchi is investigating the relationship between an athlete's height and the time the athlete takes to run 100 metres.

(a) Suggest a hypothesis she could use to investigate this.

....

....

(1)

Ruchi collects the data for the following variables.

The age of the athlete
The gender of the athlete
The hair colour of the athlete

*(b) Which one of these variables is best represented on a stem and leaf diagram?
Give a reason for your answer.

....

....

(2)

Ruchi also collects the data for the following variables.

The shoe size of the athlete
The number of steps taken to run 100 metres
The time taken to run 100 metres

*(c) Which one of these variables is best represented on a histogram?
Give a reason for your answer.

....

....

....

(2)

Ruchi wants to find out whether or not the distribution of the heights of the athletes is skewed. She finds the lower quartile and the upper quartile of the distribution.

(d) Write down the name of another statistic for the distribution that she will need to find.

....

(1)

(Total for Question 5 is 6 marks)

6 Pablo takes a reading test, a writing test and a speaking test.

Each test has the same total score.

The table shows Pablo’s score in each test.

Test	**Score**
Reading	55
Writing	41
Speaking	57

(a) Calculate Pablo’s mean test score.

....

(1)

Each test score is now given a weighting.

Test	**Score**	**Weighting**
Reading	55	30
Writing	41	45
Speaking	57	25

(b) (i) Describe in words why Pablo’s **weighted mean** test score will be less than his mean test score.

....

....

(ii) Calculate Pablo’s **weighted mean** test score.

....

(3)

The frequency table shows information about the reading test score, x, for 50 students.

Score	Frequency	
$30 < x \leqslant 40$	1	
$40 < x \leqslant 50$	3	
$50 < x \leqslant 60$	12	
$60 < x \leqslant 70$	18	
$70 < x \leqslant 80$	12	
$80 < x \leqslant 90$	3	
$90 < x \leqslant 100$	1	

Pablo scored 55 on this reading test.

(c) Calculate an estimate for the number of students who scored higher than Pablo on the reading test.

....

(2)

(d) Calculate the standard deviation of the 50 reading test scores.

You may use $\Sigma fx = 3250$ and $\Sigma fx^2 = 217\,850$

....

(2)

(e) Give one reason why it may be better to summarise the spread of these test scores using the standard deviation rather than the interquartile range.

....

....

(1)

(Total for Question 6 is 9 marks)

7 The table shows the chain base index numbers for the price of an annual season rail ticket from Gloucester to Birmingham for each of the years 2011 to 2015

Year	2010	2011	2012	2013	2014	2015
Chain base index number	▨	109	106	104	103	102

Source: Transport Focus

(a) Describe what the chain base index numbers show about the price of an annual season rail ticket for the years 2010 to 2015

....

....

....

....

(2)

The cost of an annual season rail ticket in 2010 was £3032

(b) Work out the cost of the ticket in 2011

....

(2)

(c) Show that the cost of an annual season rail ticket has increased by more than 25% over the period from 2010 to 2015

(2)

(Total for Question 7 is 6 marks)

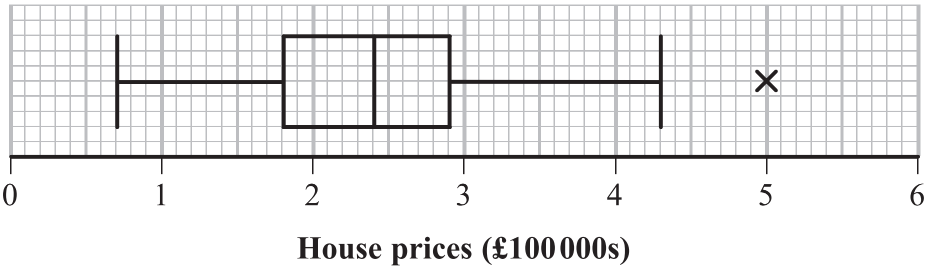
DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

8 Ashley collected information about house prices in her town last year.

She drew this box plot for this information.



Source: Land Registry

Ashley says that more than half the houses had a price greater than £200 000

(a) Explain how the box plot can be used to support what Ashley says.

....

....

(1)

The greatest price was £500 000

(b) Show, by calculation, that this price is an outlier.

(3)

(Total for Question 8 is 4 marks)

9 The owner of a bakery wants to open a new store in a large town.

He wants to find out where the residents of the town want the store to be located.

(a) The owner of the bakery may have to use primary data.
Give a reason why.

....

....

(1)

The owner is going to select a sample of the town’s residents.

He has to decide on an appropriate size for his sample.

(b) Give one possible problem if the sample size is too small.

....

....

....

(1)

The owner will choose between two sampling methods.

Method 1: Grouping the town into 8 geographical regions.
Then selecting all of the residents in two randomly selected regions.

Method 2: Grouping the town into 8 geographical regions.
Then randomly selecting residents from each region in proportion to the number of residents in that region.

(c) (i) Write down the name of sampling Method 1

....

(ii) Write down the name of sampling Method 2

....

(2)

*(d) Give **two** advantages of using Method 2 rather than Method 1

....

....

....

....

....

....

(2)

The owner is going to ask this question on a questionnaire.

"How far are you willing to travel to shop at the new store?"

(e) Design suitable response boxes for this question.

(2)

(Total for Question 9 is 8 marks)

Turn over ▶

10 The table shows the annual average ice cream consumption, in litres/person, and the Gross Domestic Product (GDP) per capita, in US Dollars, for 7 countries.

Country	**Ice cream consumption (litres/person)**	**GDP per capita (US Dollars)**	**Ice cream consumption ranks**	**GDP per capita ranks**	***d* (difference in ranks)**	
Norway	10.8	100 800	1			
Sweden	10.4	60 400	2			
Italy	6.3	35 900	3			
Croatia	6.1	13 600	4			
France	6.0	42 500	5			
Greece	5.5	21 900	6			
Bulgaria	2.3	7 500	7			

Source: Euroglaces and World Bank

(a) (i) Show that for these data $\sum d^2 = 10$

(ii) Calculate Spearman's rank correlation coefficient for these data.

....

(4)

(b) Describe the correlation found in part (a)(ii).

....

(1)

The data for the GDP is converted from US Dollars to Euros.

(c) Describe the effect this conversion will have on your answer to part (a)(ii).

....

(1)

Daniel thinks that greater ice cream consumption **causes** the GDP per capita of a country to increase.

(d) Discuss whether or not the Spearman’s rank correlation coefficient from part (a)(ii) can be used to support what Daniel thinks.

Give a reason for your answer.

....

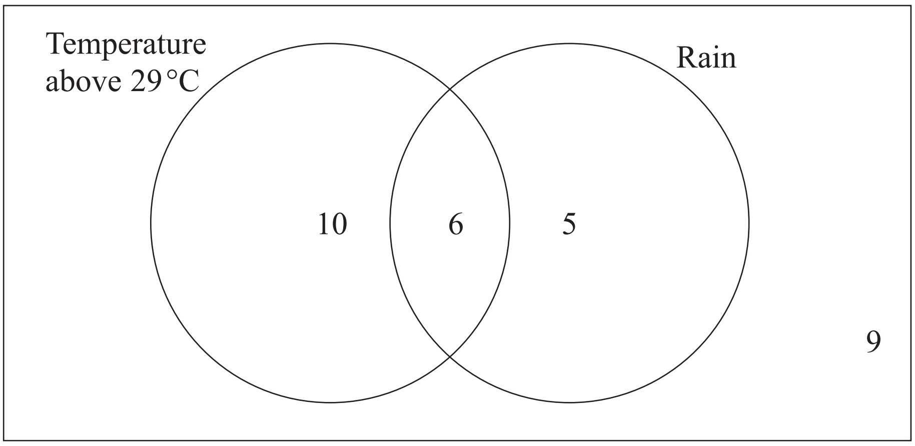
....

(1)

(Total for Question 10 is 7 marks)

11 The temperature and rainfall in Miami were recorded every day for 30 days.

The Venn diagram shows information about the number of days when the maximum temperature was above 29 °C and the number of days when it rained.



Source: www.weather.com

One of these days is selected at random.

(a) Find the probability that on this day

(i) the maximum temperature was above 29 °C,

....

(ii) it did **not** rain,

....

(iii) the maximum temperature was above 29 °C or it rained or both.

....

(4)

Two of the days are selected at random.

(b) Find the probability that on both of these days the maximum temperature was above 29 °C **and** it rained.

....

(2)

Greg thinks when it rains in Miami there is less chance that the maximum temperature will be above 29 °C than when it does not rain.

(c) By comparing two suitable probabilities, comment on what Greg thinks.

....

....

....

....

....

(2)

(Total for Question 11 is 8 marks)

12 A scientist wants to estimate the number of geese living around a lake.

The scientist captures a sample of 45 geese and puts a tag on each one. He then releases the geese.

The scientist waits one day and captures a sample of 18 geese.

He finds that 2 of these geese each have a tag.

(a) Estimate the total number of geese living around the lake.

....

(2)

(b) Give a statistical reason why the scientist waits one day before taking the second sample.

....

....

(1)

The scientist returns to the lake 1 year later and captures a sample of 18 geese.

He finds that 1 of these geese has a tag.

(c) Discuss the reliability of using this sample to estimate the total number of geese now living around the lake.

....

....

....

....

(2)

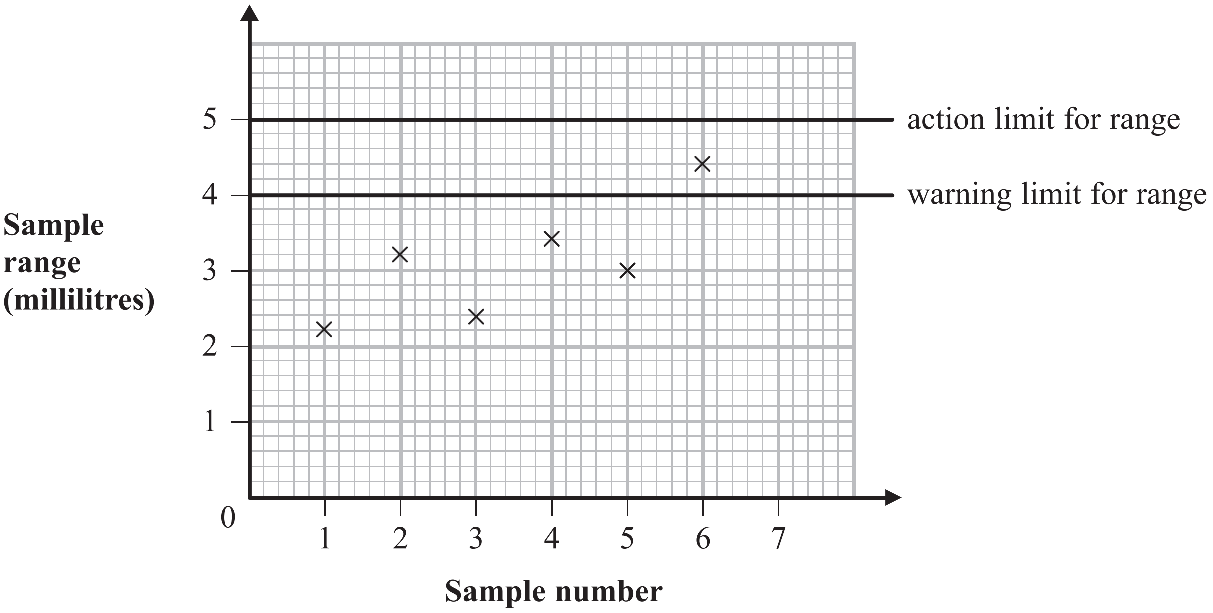
(Total for Question 12 is 5 marks)

13 A coffee machine is designed to produce 150 millilitres of coffee per serving.

For quality control, random samples of 3 servings are taken and the range of each sample is found.

A quality control chart is used to plot the sample ranges.

The first 6 sample ranges have been plotted.



(a) Describe what action should be taken after the 6th sample.

....

....

(1)

The amounts of coffee, in millilitres, in the 7th sample are

147.4 152.6 152.1

(b) (i) Find the value of the sample range for this sample.

.... millilitres

(ii) Plot this sample range on the quality control chart.

(iii) Describe what action should be taken after the 7th sample.

....

....

(3)

(Total for Question 13 is 4 marks)

14 There are 10 marbles in a bag.

7 of the marbles are red.

3 of the marbles are blue.

One marble is taken at random from the bag.

The colour of the marble is recorded and the marble is put back into the bag.

This process is repeated until the colour of each of 5 marbles has been recorded.

(a) (i) Write down the probability that the first marble taken from the bag is red.

....

(ii) Calculate the probability that the first 2 marbles taken from the bag are both red.

....

(3)

(b) Calculate the probability that more than half of the 5 marbles taken from the bag are red.

You may use $(p + q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$

....

(3)

(c) Show that the most likely number of red marbles taken from the bag is 4

You may use $(p + q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$

(2)

A new experiment takes place.

One marble is taken at random from the bag and its colour recorded.

In this experiment the marble is **not** put back into the bag.

This process is repeated until the colour of each of 5 marbles has been recorded.

(d) Discuss whether or not the binomial distribution can be used to model the number of red marbles taken from the bag in this experiment.

....

....

....

(2)

(Total for Question 14 is 10 marks)

15 In the 2015 Women's European Gymnastics Championships, two of the events were the vault and the balance beam.

Jeff thinks the scores may be modelled by normal distributions.

For the vault, the competitors' mean score was 14.5 and the standard deviation was 0.6

Source: European Union of Gymnastics

One competitor scored 14.1 in the vault.

(a) Calculate the standardised score for this competitor.
Give your answer correct to 1 decimal place.

....

(2)

The same competitor had a standardised score of 0.5 for the balance beam.

(b) Compare the performance of this competitor in the vault with her performance in the balance beam.

....

....

....

....

....

....

(2)

The highest score in the vault was 15.3

(c) Use standardised scores to discuss whether or not the scores in the vault may be modelled by a normal distribution.

(3)

(Total for Question 15 is 7 marks)

TOTAL FOR PAPER IS 100 MARKS

BLANK PAGE

BLANK PAGE

BLANK PAGE