SECTION A: STATISTICS

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

1. A company is introducing a job evaluation scheme. Points (x) will be awarded to each job based on the qualifications and skills needed and the level of responsibility. Pay (£y) will then be allocated to each job according to the number of points awarded.

    Before the scheme is introduced, a random sample of 8 employees was taken and the linear regression equation of pay on points was $y = 4.5x - 47$

    (a) Describe the correlation between points and pay.

    (1)

    (b) Give an interpretation of the gradient of this regression line.

    (1)

    (c) Explain why this model might not be appropriate for all jobs in the company.

    (1)

1a) positive correlation

b) With every extra point awarded, pay increases by £4.50.

c) If the points awarded is less than 11 then the pay is negative.

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

2. A sixth form college has 84 students in Year 12 and 56 students in Year 13

   The head teacher selects a stratified sample of 40 students, stratified by year group.

   (a) Describe how this sample could be taken.

   (3)

   The head teacher is investigating the relationship between the amount of sleep, s hours, that each student had the night before they took an aptitude test and their performance in the test, p marks.
   For the sample of 40 students, he finds the equation of the regression line of p on s to be

   $$p = 26.1 + 5.60s$$

   (b) With reference to this equation, describe the effect that an extra 0.5 hours of sleep may have, on average, on a student's performance in the aptitude test.

   (1)

   (c) Describe one limitation of this regression model.

   (1)

a Number year 12 students 1-84 and year 13 students 1-56, then use a random number source e.g. calculator to choose a simple random sample!

Yr 12 : Yr 13 = 84:56 = 3:2

40 split into 3:2 ratio : $\frac{40}{5}$ = 8 , 24:16

Choose 24 year 12 students and 16 year 13.

b. $\Delta p = 5.6 \times 0.5 = 2.8 \Rightarrow$ 2.8 mark increase

c. The model has no defined range $\Rightarrow$ students who have infinite sleep (don't wake up) will perform best.

3. Sara was studying the relationship between rainfall, r mm, and humidity, h%, in the UK. She takes a random sample of 11 days from May 1987 for Leuchars from the large data set.

She obtained the following results.

h	93	86	95	97	86	94	97	97	87	97	86
r	1.1	0.3	3.7	20.6	0	0	2.4	1.1	0.1	0.9	0.1

Sara examined the rainfall figures and found

$$Q_1 = 0.1 \qquad Q_2 = 0.9 \qquad Q_3 = 2.4$$

A value that is more than 1.5 times the interquartile range (IQR) above Q_3 is called an outlier.

(a) Show that $r = 20.6$ is an outlier.
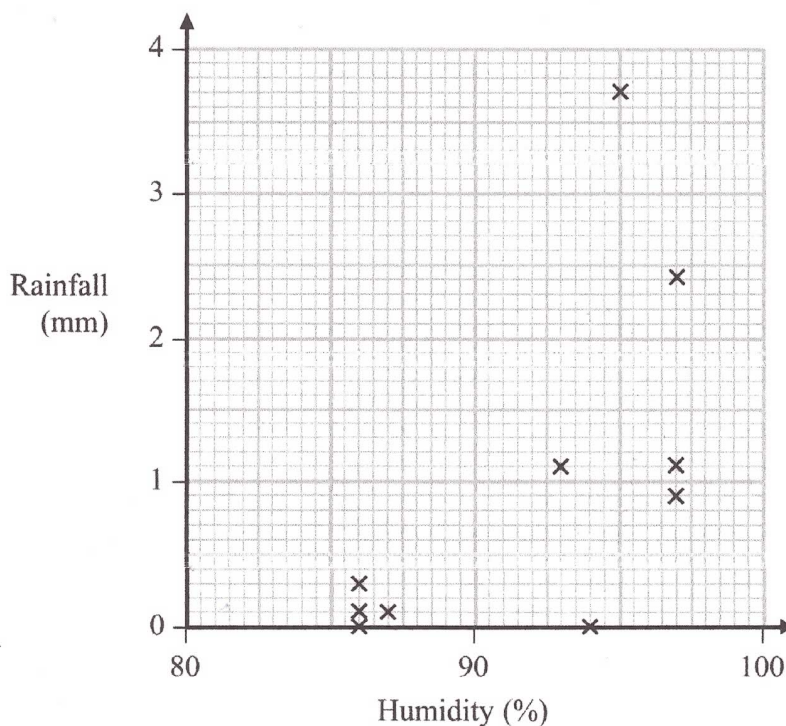
(1)

(b) Give a reason why Sara might:

(i) include

(ii) exclude

this day's reading.

(2)

Sara decided to exclude this day's reading and drew the following scatter diagram for the remaining 10 days' values of r and h.

(c) Give an interpretation of the correlation between rainfall and humidity.

(1)

Question continued

The equation of the regression line of r on h for these 10 days is $r = -12.8 + 0.15h$

(d) Give an interpretation of the gradient of this regression line.

(1)

(e) (i) Comment on the suitability of Sara's sampling method for this study.

(ii) Suggest how Sara could make better use of the large data set for her study.

(2)

(a) $IQR = Q_3 - Q = 2.4 - 0.1 = 2.3$

$2.4 + 1.5 \times 2.3 = 5.85$

$r = 20.6 > 5.85$, so $r = 20.6$ is an outlier

(b) (i) She should include it because it is a piece of data and all data should be considered.

(ii) She could exclude it since it is ane extreme value and affect the investigation.

(c) As humidity increases, rainfall increases.

(d) The gradient (0.15) represents that there's a 0.15mm increase in rainfall per percentage of humidity.

(e) (i) Sara's sampling method isn't very good since she only uses 11 days out of the whole month, and only uses one specific location.

(ii) Sara could use data from more than one UK location and also use a wider range of months with more days per month.

(Total for Question is 7 marks)

4. Jerry is studying visibility for Camborne using the large data set June 1987.

The table below contains two extracts from the large data set.

It shows the daily maximum relative humidity and the daily mean visibility.

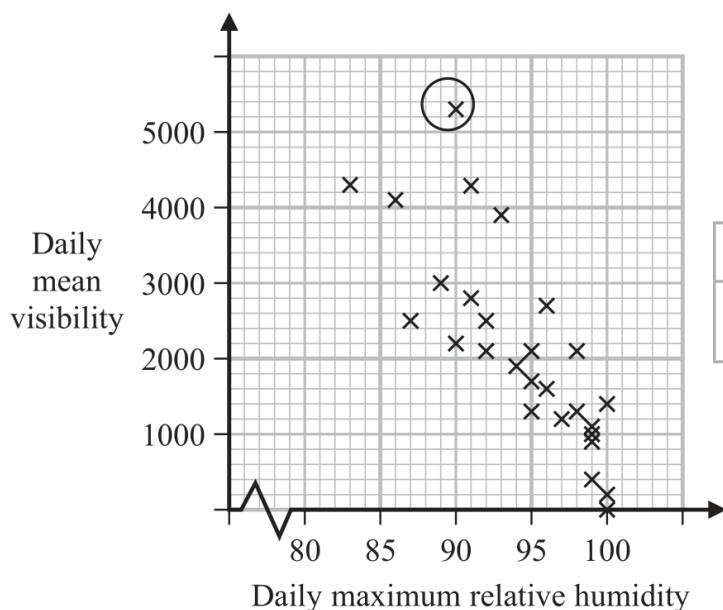| Date	Daily Maximum Relative Humidity	Daily Mean Visibility
Units	%	
10/06/1987	90	5300
28/06/1987	100	0

(The units for Daily Mean Visibility are deliberately omitted.)

Given that daily mean visibility is given to the nearest 100,

(a) write down the range of distances in metres that corresponds to the recorded value 0 for the daily mean visibility.

(1)

Jerry drew the following scatter diagram, Figure 2, and calculated some statistics using the June 1987 data for Camborne from the large data set.

	Q_1	IQR
Daily mean visibility	1100	1600
Daily maximum relative humidity (%)	92	8

Figure 2

Jerry defines an outlier as a value that is more than 1.5 times the interquartile range above Q_3 or more than 1.5 times the interquartile range below Q_1.

(b) Show that the point circled on the scatter diagram is an outlier for visibility.

(2)

(c) Interpret the correlation between the daily mean visibility and the daily maximum relative humidity.

(1)

Jerry drew the following scatter diagram, Figure 3, using the June 1987 data for Camborne from the large data set, but forgot to label the *x*–axis.
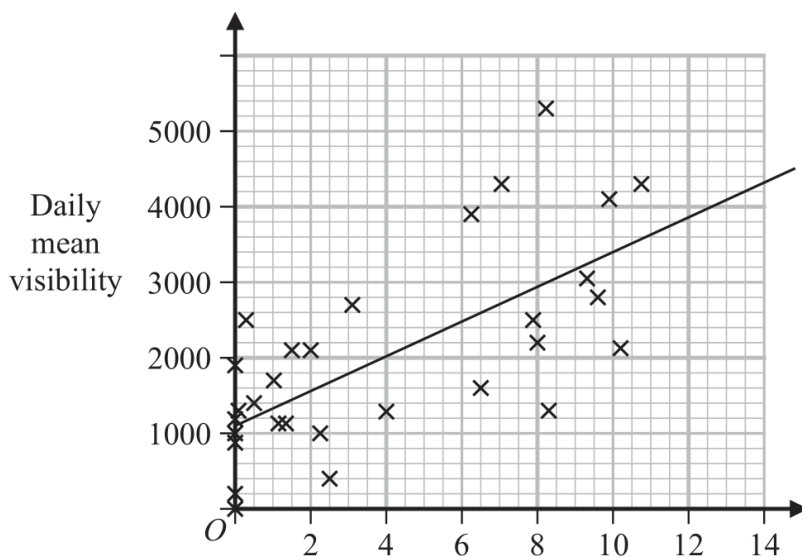

Figure 3

(d) Using your knowledge of the large data set, suggest which variable the *x*-axis on this scatter diagram represents.

familiarise yourself with how the quantities in the set are measured **(1)**

a) rounded to nearest hundred <u>decametres</u>

any value 0-50dm rounds down to 0 ⟹ 0 to 500m

b) data point: (90, 5300)

value at which point becomes an outlier:

$Q_3 = Q_1 + IQR = 1100 + 1600 = 2700$ m

$Q_3 + 1.5 IQR = 2700 + 1.5 \times 1600 = 5100$ m

5300 > 5100 ⟹ outlier

c) as the humidity increases, the mean visibility decreases

d) <u>we need something that increases visibility</u>

x = hours of sunshine

Turn over ▶